Designing Long-term Group Fair Policies in Dynamical Systems

Miriam Rateike
Saarland University
Saarbrücken, Germany
rateike@cs.uni-saarland.de

Isabel Valera
Saarland University
Max Planck Institute for Software Systems
Saarbrücken, Germany
ivalera@cs.uni-saarland.de

Patrick Forré
AI4Science Lab, AMLab
Informatics Institute
University of Amsterdam
Amsterdam, The Netherlands
p.d.forre@uva.nl

ABSTRACT

Neglecting the effect that decisions have on individuals (and thus, on the underlying data distribution) when designing algorithmic decision-making policies may increase inequalities and unfairness in the long term—even if fairness considerations were taken into account in the policy design process. In this paper, we propose a novel framework for studying long-term group fairness in dynamical systems, in which current decisions may affect an individual's features in the next step, and thus, future decisions. Specifically, our framework allows us to identify a time-independent policy that converges, if deployed, to the *targeted* fair stationary state of the system in the long-term, independently of the initial data distribution. We model the system dynamics with a time-homogeneous Markov chain and optimize the policy leveraging the Markov Chain Convergence Theorem to ensure unique convergence. Our framework enables the utilization of historical temporal data to tackle challenges associated with delayed feedback when learning long-term fair policies in practice. Importantly, our framework shows that interventions on the data distribution (e.g., subsidies) can be used to achieve policy learning that is *both short- and long-term fair*. We provide examples of different targeted fair states of the system, encompassing a range of long-term goals for society and policymakers. In semi-synthetic simulations based on real-world datasets, we show how our approach facilitates identifying effective interventions for long-term fairness.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Social and professional topics**;

KEYWORDS

fairness, long-term, dynamical system, equilibrium, policy learning

ACM Reference Format:
Miriam Rateike, Isabel Valera, and Patrick Forré. 2024. Designing Long-term Group Fair Policies in Dynamical Systems. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 31 pages. https://doi.org/10.1145/3630106.3658538

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658538

1 INTRODUCTION

The majority of fairness notions that have been developed for trustworthy machine learning [19, 25] assume an unchanging data generation process, i.e., a static system. Consequently, existing work has explored techniques to integrate these fairness considerations into the design of algorithms in static systems [2, 19, 25, 89, 90]. However, in many practical settings [8, 22], we observe a dynamic interplay between algorithmic decisions and the individuals they affect, which alters the data generation process over time – effectively creating a dynamical system.

Consider a lending scenario, which we will be using as a running example throughout this paper[1], where a bank decides to give loans based on a individuals credit scores. This is a common setting in the literature on fair algorithmic decision-making [15, 17, 46, 83]. It becomes a dynamical system when decisions to grant loans result in changes in individuals' credit scores for subsequent lending applications. This phenomenon may occur for various reasons, such as bureaucratic processes influencing credit score changes in response to paid or unpaid loans after loan has been granted [46], social learning [27], recourse efforts [36], or strategic behavior of affected individuals [24]. In the presence of a feedback loop from decisions to the data generation process, decision-making can be viewed as an iterative process in many fairness scenarios. This results in a data distribution that evolves over time, shaping a dynamical system.

Prior work has shown that policies that do not take into account their impact on the underlying data distribution, may exacerbate inequalities and unfairness over time [17, 28, 29, 46, 54, 83, 91]. Recently, research has introduced optimization approaches aimed at learning decision-making policies that achieve long-term fairness. The majority of these efforts have demonstrated the effectiveness of reinforcement learning (RL) approaches in learning long-term fair policies when modeling system dynamics as Markov Decision Processes (MDPs) [10, 32, 64, 82, 85–87]. These approaches typically operate under the assumption of unknown dynamics, which determine how features change in response to decisions, and learn long-term fair policies through iterative online training using model-free RL. The goal then is to achieve a fair equilibrium in the long term. However, these approaches overlook certain characteristics inherent to common fairness applications. In addition, they usually compromise short-term fairness for long-term fair goals and thus neglect the short-term needs of individuals affected by algorithmic decisions.

Prior work suggests deploying a new policy at each time step, utilizing it to gather additional data, and iteratively refining the

[1] However, our results apply to a wide variety of other scenarios, such as university admissions [39, 43] and criminal recidivism [44].

policy until it reaches convergence. Policy updates occur after observing features and their changes in response to decision-making. In typical fairness scenarios, the time span between a decision and the observation of its impact on individuals' features can extend over years.[2] This would result in updating a policy every several years, which can significantly slow down the learning process in practice [79, 80]. Successful applications of online RL typically occur in settings where a simulator or game is accessible and provides instant feedback to the decisions [16, 59]. However, in the real world, meeting such requirements is often unfeasible.

Moreover, frequent policy updates may undermine trust in algorithmic decision-making [67]. On one hand, a policy could yield different decisions for the same individual due to random initializations. On the other hand, more importantly, individuals with the same features may receive different decisions before and after policy updates. This temporal inconsistency of decisions can lead to a perception of unfairness and a lack of trustworthiness in the algorithmic decision-making system.

Finally, compromising short-term fairness for long-term fair goals overlooks the critical need to address immediate fairness concerns. By solely prioritizing future fairness objectives, individuals may experience harm today. This can further erode trust in algorithmic decision-making processes.

Our Contributions. In this paper, we introduce a structured framework for thinking about long-term fairness. Inspired by [91], we connect the existing work on the Markov Chain Convergence Theorem with long-term fairness. Specifically, we employ Markov chains (MCs) as a framework to model system dynamics, where the sought-after long-term fair policy defines the Markov kernel. Similar to prior work [9, 82, 86, 87, 91], we assume time-homogeneous (fixed) dynamics. This allows us to: i) propose a new method for learning long-term group fair policies from historical temporal data, ii) demonstrate that interventions in the data generative system can ensure both short- and long-term fairness in the policy learning process. *To the best of our knowledge, our paper is the first to connect the Markov Chain Convergence Theorem to the goal of learning long-term fair policies.*

First, we propose a method for policy learning that can be seen as a form of model-based RL. We provide a structured framework that separates i) environment estimation, ii) problem modeling, and iii) optimization. This structured approach contributes to our understanding of long-term fairness and the design of policy and societal interventions. We assume that the environment (i.e., transition probabilities or the Markov kernel) can be explicitly estimated. Further, we assume that have access to a pre-existing temporal dataset that provides sufficient information for the reliable estimation of the environment. This dataset can be collected from other agents or humans in the past to learn the dynamical model. Subsequently, we impose the necessary convergence criteria from the Markov Chain Convergence Theorem on the policy-induced kernel. This ensures convergence to a unique equilibrium. Building upon this, we propose an optimization problem to find a policy that converges to a targeted fair equilibrium, provided such an equilibrium exists. In cases where the historical dataset is not representative or is unavailable, our framework can be extended to accommodate necessary policy updates.

Our policy, if found and deployed, ensures convergence to the fair long-term target without requiring further policy updates. This ensures consistent decision-making for individuals, where individuals across time steps experience the same probability of a decision (for given features). This can support public trust in algorithmic decision-making. Our policy is robust to covariate shifts, which are variations in feature distributions between training and test data under identical dynamical systems [65]. In our lending example, this means we can learn the policy from historical data collected from a different financial institution than deployed at, as long as the dynamic mechanism governing credit score changes remains consistent. Unlike previous approaches, our method enables evaluating a policy's temporal evolution and equilibrium before deployment, avoiding the societal risks associated with deploying a suboptimal policy on a population.

Second, we show that societal interventions (e.g., subsidies) on the distribution of the feature that defines the Markov Chain can ensure both short- and long-term fairness during policy learning. This eliminates the need for trading off long- and short-term fairness during policy learning as seen in prior works.

Our primary contributions are:

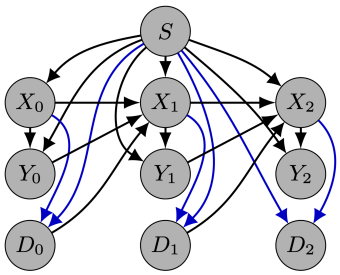
- We introduce a modeling approach for learning long-term fair policies, assuming access to a sufficient historical temporal dataset and a time-homogeneous environment.
- Based on this, we introduce an optimization problem for finding a policy that ensures convergence to a targeted long-term fair equilibrium. If the model assumptions are correct, this approach proves more efficient than purely data-driven methods.
- Given a policy that is guaranteed to converge to a (fair) equilibrium, we formally distinguish between interventions on the data distribution that affect the short-term convergence trajectory and those that affect the long-term equilibrium. These allows to design interventions to achieve fairness both short- and long-term fairness.
- We validate our method in simulated populations initialized with real-world data, and show i) that our method finds a decision-making policy that is robust to covariate shift and, compared to short-term policies, offers a significantly more stable learning process, achieving a better utility-fairness trade-off at convergence; ii) external interventions on the data distribution can ensure short-term fairness as the policy converges to the long-term fair equilibrium.

2 GUIDING EXAMPLE

We present as guiding example a credit lending scenario [15, 17, 46]. The data generative process is shown in Figure 1. For assumptions in this example, see § F.1. Our framework also applies to other generative processes; see an example in § F.2. While our results hold for continuous state and action spaces, for clarity, we predominantly use finite state and action spaces for formalization in this paper.

Data Generative Model. Let an individual with protected attribute S (e.g. gender) at time t be described by a non-sensitive feature X_t

[2] Common datasets report recidivism at intervals of 2 years [44], university graduation typically occurs within 3-6 years after admission, and credits can extend for up to 30 years.



$\gamma(s)$	$\mathbb{P}(S = s)$
$\mu_t(x \mid s)$	$\mathbb{P}(X_t = x \mid S = s)$
$\pi(d \mid x, s)$	$\mathbb{P}(D_t = d \mid X_t = x, S = s)$
$\ell(y \mid x, s)$	$\mathbb{P}(Y_t = y \mid X_t = x, S = s)$
$g(k \mid x, d, y, s)$	$\mathbb{P}(X_{t+1} = k \mid X_t = x, D_t = d, Y_t = y, S = s)$
P^s_π	$\mathbb{P}(X_{t+1} = k \mid X_t = x, S = s)$ $= \sum_{d,y} g(k \mid x, d, y, s)\pi(d \mid x, s)\ell(y \mid x, s)$

Figure 1: Data generative model over time steps (subscript) $t = \{0, 1, 2\}$. Non-sensitive feat. X_t, sensitive feat. S, ground truth Y_t, decision D_t. Time-varying feature distribution μ_t, fixed sensitive distribution γ, policy π (blue), ground truth distribution ℓ, dynamics g, time-homogeneous kernel P^s_π.

(e.g. credit score as a summary of monetary assets and credit history) and an outcome of interest Y_t (e.g. repayment ability). We assume the sensitive attribute to remain immutable over time and drop the attribute's time subscript. For simplicity, we assume binary sensitive attribute and outcome of interest $S, Y \in \{0, 1\}$ and a one-dimensional discrete non-sensitive feature $X \in \mathbb{Z}$. Let the population's sensitive attribute be distributed as $\gamma(s) := \mathbb{P}(S=s)$ and remain constant over time. We assume X to depend on S, such that the group-conditional feature distribution at time t is $\mu_t(x \mid s) := \mathbb{P}(X_t = x \mid S=s)$.

For example, different demographic groups may have different credit score distributions due to structural discrimination in society. The outcome of interest Y is assumed to depend on X and (potentially) on S resulting in the conditional distribution $\ell(y \mid x, s) := \mathbb{P}(Y_t = y \mid X_t = x, S=s)$. This distribution is assumed to remain unchanged over time. For example, payback probability may be tied to factors like income, which can be assumed to be encompassed within a credit score. We assume that there exists a decision maker that takes binary loan decisions based on X and (potentially) S and decides with probability $\pi(d \mid x, s) := \mathbb{P}(D_t = d \mid X_t = x, S=s)$.

Dynamical System. Consider dynamics where a decision D_t at time step t directly influences an individual's features X_{t+1} at the next step. We assume that the transition from the current feature state X_t to the next state X_{t+1} depends on the current features X_t, outcome Y_t, and (possibly) the sensitive attribute S. For example, after a positive lending decision, an individual's credit score may rise due to successful loan repayment. The magnitude of this increase could potentially be influenced by their sensitive attribute. We describe the probability of an individual with $S=s$ transitioning from a credit score of $X_t = x$ to $X_{t+1} = k$ in the next step, conditioned on the received decision $D=d$ and repayment $Y_t = y$, as the dynamics $g(k \mid x, d, y, s) := \mathbb{P}(X_{t+1} = k | X_t = x, D_t = d, Y_t = y, S=s)$. We assume dynamics to remain unchanged over time. Then the probability of a feature changing from $X_t = x$ to $X_{t+1} = k$ in the next step given $S=s$ is obtained by marginalizing out D_t and Y_t, resulting in transition probabilities

$$\mathbb{P}(X_{t+1} = k \mid X_t = x, S = s) = \sum_{d,y} g(k \mid x, d, y, s)\pi(d \mid x, s)\ell(y \mid x, s). \tag{1}$$

Importantly, the next step feature state depends only on the present feature state, and not on any past states. This dynamical system can then be seen as a time-homogeneous Markov chain with state space $\mathcal{X}$ and transition probabilities (1) that depend on a fixed policy π and a sensitive attribute S (given time-independent g and ℓ).

Time-homogeneous Markov Chain. We remind the reader of the formal definition of time-homogeneous Markov chains with discrete states space. For a formulation for general state spaces refer to § A or [51].

Definition 2.1 (Time-homogeneous Markov Chain [21]). *A time-homogeneous Markov chain on a discrete space $\mathcal{Z}$ with transition probability P is a sequence of random variables $(Z_t)_{t \in T}$ with joint distribution $\mathbb{P}$, such that for every $t \in T$ and $z, w \in \mathcal{Z}$ we have $\mathbb{P}(Z_{t+1} = w \mid Z_t = z) = P(z, w)$.*

In a Markov chain, each event's probability depends solely on the previous state. Recall that the transition probabilities must satisfy $P(z, w) \geq 0$ for all z, w, and $\sum_w P(z, w) = 1$ for all z. *Time-homogeneous* in this context pertains to the assumption that P remains constant and does not undergo any changes over time.

Having established that the guiding example can be modeled as a time-homogeneous Markov chain, we proceed to generalize the objective for long-term fair policies, encompassing previous approaches in the field.

3 DESIGNING LONG-TERM FAIR POLICIES

Consider a policymaker (e.g., bank, government), aiming to achieve a fair equilibrium that takes the form of a *fair distribution* in the long term. An example of such fair distribution could be equal credit score distributions across demographic groups [17]. We propose definitions of fair distributions (or targeted equilibria) further below (§ 6). To achieve a fair distribution in the long term, we need to describe the evolution of the group-conditional feature distribution $\mu_t(x \mid s)$ across time t. The behavior of the dynamical system is defined by the transition probabilities (1) together with the initial state distribution $\mu_0(x \mid s)$. At time t, the next step's feature distribution of group s can be computed for all $k \in \mathcal{X}$ as $\mu_{t+1}(k \mid s) = \sum_x \mu_t(x \mid s)\mathbb{P}(X_{t+1} = k \mid X_t = x, S = s)$.

Imagine a scenario where, at a specific time t, credit scores are already fair distributed. In such cases, the policymaker's objective is to uphold this fair distribution in the subsequent time step. This means that the policymaker asks the following to hold for each group s and all $x \in \mathcal{X}$:

$$\mu_{t+1}(x \mid s) = \mu_t(x \mid s) \tag{2}$$

To formally define this, we remind the reader of the definition of a stationary distribution, which is a state distribution that remains unchanged when multiplied by the transition probabilities:

Definition 3.1 (Stationary Distribution [21]). *A stationary distribution of a time-homogeneous Markov chain $(\mathcal{Z}, P)$ is a probability distribution μ, such that $\mu = \mu P$. More explicitly, for every $w \in \mathcal{Z}$ the following needs to hold: $\mu(w) = \sum_z \mu(z) \cdot P(z, w)$.*

Fundamental Objective for Long-term Fair Policies. We now provide a generalization of (2). Let a population's feature distribution over time be represented by a time-homogeneous Markov chain $(Z_t)_{t\in T}$ with a (general) state space $\mathcal{Z}$. The transition probabilities P^s_π depend on the sensitive attribute S and some policy π. Consider a scenario where our society is already in a fair state $(\mu^s)_{s\in\mathcal{S}}$. In this case, the policymaker would aim to find policy π that defines a transition probability P^s_π such that the next state remains fair. More formally, we would seek to satisfy the following equation:

$$\mu^s = \mu^s P^s_\pi \tag{3}$$

for all $s \in \mathcal{S}$. Therefore, the fair distribution $(\mu^s)_{s\in\mathcal{S}}$ should be the stationary distribution of the Markov chain defined by $(\mathcal{Z}, P^s_\pi)$. Any policy that aims for the fair stationary state $(\mu^s)_{s\in\mathcal{S}}$ will eventually need to find a policy that satisfies (3) to at least transition from a fair state to a fair state in the long term. *In this sense, (3) defines the fundamental problem of finding long-term fair policies.* Reaching a fair equilibrium is central to prior work (e.g., [86, 87, 91]), with differences primarily emerging from definitions of a fair equilibrium and the methods employed to achieve it.

Long-term Fair Policy Interventions. In our example, the objective of the policymaker is to identify a policy π that ensures the convergence of the credit score distribution to the intended fair distribution. We can formally describe this as follows: Suppose a policymaker aims to achieve a fair distribution $(\mu^s)_{s\in\mathcal{S}}$. The goal for the policymaker is then to find a policy π such that the induced transition probabilities P^s_π converge to the distribution $(\mu^s)_{s\in\mathcal{S}}$, and the distribution $(\mu^s)_{s\in\mathcal{S}}$ satisfies the defined fairness constraints. To identify a long-term fair policy, we propose a general optimization problem in § 5 that utilizes the Markov Chain Convergence Theorem, which we introduce next.

4 LONG-TERM AND SHORT-TERM INTERVENTIONS

In this section, we present the properties that our policy-induced transition probabilities must possess to fulfill the previously established objective of achieving long-term fair policies (§ 4.1). From this, we can derive theoretical insights on the type of interventions that yield short-term and long-term effects (§ 4.2).

4.1 Markov Chain Convergence Theorem

The Markov Chain Convergence Theorem establishes conditions for a time-homogeneous Markov chain to converge to a unique stationary distribution. In our model, the transition probabilities depend on the sensitive attribute, and we will apply in (4) the Markov Chain Convergence Theorem separately to each group's transition probabilities. We thus drop the superscript s. We first provide definitions for irreducibility and aperiodicity required for stating the Markov Chain Convergence Theorem thereafter.

Definition 4.1 (Irreducibility [21]). *A time-homogeneous Markov chain is irreducible if, for any two states $z, w \in \mathcal{Z}$, there exists a $t > 0$ such that $P^t(z, w) > 0$, where $P^t(z, w) = \mathbb{P}(Z_t = w \mid Z_0 = z)$ represents the probability of going from z to w in t steps.*

In other words, irreducibility ensures that there is a positive probability of reaching any state w from any state z after some finite number of steps. Note, for discrete state space $\mathcal{Z}$, every irreducible time-homogeneous Markov chain has a unique stationary distribution (Thm. 3.3 [21]).

Definition 4.2 (Aperiodicity [21]). *Consider an irreducible time-homogeneous Markov chain $(\mathcal{Z}, P)$. Let the set of return times from $z \in \mathcal{Z}$ be $R(z) = \{t \geq 1 : P^t(z, z) > 0\}$, where $P^t(z, z)$ represents the probability of returning to state z after t steps. The Markov chain is aperiodic if and only if the greatest common divisor (gcd) of $R(z)$ is equal to 1: $gcd(R(z)) = 1$ for all z in $\mathcal{Z}$.*

In words, aperiodicity refers to the absence of regular patterns in the sequence of return times to state z, i.e., the chain does not exhibit predictable cycles or periodic behavior.

Theorem 4.3 (Markov Chain Convergence Theorem [21]). *Let $(Z_t)_{t\in T}$ be an irreducible and aperiodic time-homogeneous Markov chain with discrete state space $\mathcal{Z}$ and transition probabilities P. Then the marginal distribution $\mathbb{P}(Z_t)$ converges to the unique stationary distribution μ as t approaches infinity (in total variation norm), regardless of the initial distribution $\mathbb{P}(Z_0)$.*

In other words, the Markov Chain Convergence Theorem states that, regardless of the initial distribution, the state distribution of an irreducible and aperiodic Markov chain eventually converges to the *unique* stationary distribution.

For general state spaces the Markov Chain Convergence Theorem can be proven under Harris recurrence, aperiodicity, and the existence of a stationary distribution [51] (see § A).

4.2 Characterizing Long- and Short-term Interventions

The Markov Chain Convergence Theorem establishes a formal foundation for differentiating interventions that influence the long-term dynamics of the system from those that do not. We summarize this in the following remark:

Remark 4.4 (Short- and Long-term Interventions). *In a time-homogeneous Markov chain $(Z_t)_{t\in T}$ with state space $\mathcal{Z}$ and transition probabilities P that fulfill the necessary properties of the Markov Chain Convergence Theorem, intervening on the state distribution μ_t at time t does not impact the stationary distribution μ (short-term intervention). Intervening on the transition probabilities P while preserving the necessary properties of the Markov Chain Convergence Theorem can alter μ (long-term intervention).*

This remark is important in the ongoing discourse regarding long-term fair policy design and societal interventions [27, 52, 76] and as well as for characterizing the trade-off between short-term and long-term fairness. For any given data generative model that can be described as a time-homogeneous Markov chain, we can determine whether an intervention has a long-term effect, before deploying it. In our lending example, altering the distribution of credit scores via one-time economic subsidies might yield short-time effects, but does not alter the long-term equilibrium. Instead,

the transition probabilities (1) are defined by the decision-making mechanism π, dynamics g, and the probability of the outcome of interest ℓ. An intervention in any of these probabilities, or a combination thereof, can influence the long-term equilibrium. A bank could change decision-making policy $\pi(d \mid x, s)$ that determines whether an individual receives a loan or not (as exemplified in this paper). Instead, given a fixed π, a recourse policymaker could also alter its recommendations for feature changes to individuals with declined credits, thereby changing the dynamics $g(k \mid x, d, y, s)$[3], or a government could regulate the cost of credit, leading to a change in the probabilities of repayment for the same features or risk score $\ell(y \mid x, s)$.

We can further characterize the trade-off between short-term and long-term interventions. We note that there is no inherent trade-off between short and long-term interventions in reaching the targeted fair equilibrium. However, a trade-off may emerge in the temporal and reSource costs associated with convergence. Short-term interventions, while addressing immediate population needs, could potentially extend (or reduce) the time required to reach the desired long-term target. We explore this empirically in § 7.3.

5 LONG-TERM FAIR POLICY OPTIMIZATION

We now reformulate objective (3) into a computationally solvable optimization problem for finding a time-independent policy. This policy, if deployed, leads the system to convergence to a fair stationary state in the long term, regardless of the initial data distribution.

Definition 5.1 (General Optimization Problem). *Assume a time-homogeneous Markov chain $(\mathcal{Z}, P^s_\pi)$ defined by a state space $\mathcal{Z}$ and a kernel P^s_π. To find policy π that ensures the Markov chain's convergence to a unique stationary distribution $(\mu^s)_{s\in\mathcal{S}}$, while minimizing a fair long-term objective J_{LT} and adhering to a set of fair long-term constraints C_{LT}, we propose the following optimization problem:*

$$\begin{aligned} &\min_{\pi} \quad J_{LT}((\mu^s)_{s\in\mathcal{S}}, \pi) \\ &\textit{subj. to} \quad C_{LT}((\mu^s)_{s\in\mathcal{S}}, \pi) \geq 0; \quad C_{conv}(P^s_\pi) \geq 0 \,\forall s \end{aligned} \tag{4}$$

where C_{conv} are convergence criteria according to the Markov Chain Convergence Theorem.

In words, we aim to find a policy π that minimizes a long-term objective J_{LT} subject to long-term constraints C_{LT} and convergence constraints C_{conv}. The objective J_{LT} and constraints C_{LT} are dependent on the policy-induced stationary distribution $(\mu^s)_{s\in\mathcal{S}}$, which represents the long-term equilibrium state of the data distribution and may also depend directly on the policy π. In § 6, we provide various instantiations of long-term objectives and constraints to illustrate different ways of parameterizing them. Convergence constraints C_{conv} are placed on the kernel P^s_π and guarantee convergence of the chain to a *unique stationary distribution for any starting distribution* according to the Markov Chain Convergence Theorem (Def. 4.3). The specific form of C_{conv} depends on the properties of the Markov chain, such as whether the state space is finite or continuous.

[3]Traditionally, algorithmic recourse is considered at an individual level within static scenarios often in causal settings [34]. This proposal here would introduce a dynamic perspective and consider a recourse policy on a population level.

Solving the Optimization Problem. In our example, the Markov chain is defined over a categorical feature X (credit score), resulting in a finite state space. In this case, the optimization problem becomes a linear constrained optimization problem and we can employ any efficient black-box optimization methods for this class of problems (e.g., [41]). We detail this for our example: The convergence constraints C_{conv} are determined by the aperiodicity and irreducibility properties of the corresponding Markov kernel (see § 4). It has been shown that an $n \times n$ transition matrix P constitutes an irreducible and aperiodic Markov chain if and only if all entries of $(P)^n$ are strictly positive [7]. We can thus impose as convergence constraint C_{conv} for finite states $\sum_{i=1}^{n} \left(T^s_\pi\right)^n > \mathbf{0}\,\forall s$, where n is the number of states ($n = |X|$), and $\mathbf{0}$ denotes the matrix with all entries equal to zero. The group-dependent stationary distribution μ^s_π based on T^s_π can be computed via eigendecomposition [81]. In the next section we introduce objective functions J_{LT} and constraints C_{LT} that capture notions of profit and predictive fairness. Importantly, as we exemplify, for finite state spaces, these objectives and constraints are linear. We acknowledge, however, the challenges associated with solving linear problems for high-dimensional discrete states. While our general optimization problem remains applicable in the context of a continuous state space, solving it becomes more challenging with the potential introduction of non-linearities and non-convexities. We defer solving these challenges to future work.

6 EXAMPLES OF DEFINING TARGETED FAIR STATES

Our framework enables users to define their preferred long-term group fairness criteria. Prior work on long-term fairness has suggested different types of long-term fairness notions (e.g., return parity [10, 82], equal acceptance rate [64]). Indeed, a majority of these previously established group fairness criteria can be expressed as functions of the stationary distribution, allowing for their seamless integration into our framework. Here, we illustrate how long-term fair targets are quantified by defining a long-term objective J_{LT} and constraints C_{LT} in Def. (4). We provide these examples assuming discrete X and binary D, Y, S as in our guiding example (§ 2). We refer to the notation $\mu_\pi(x \mid s)$ when we are interested in $(\mu^s)_{s\in\mathcal{S}}$ at certain values x and s. For more details see § C.

6.1 Profit

Assume that when a granted loan is repaid, the bank gains a profit of $(1-c)$; when a granted loan is not repaid, the bank faces a loss of c; and when no credit is granted, neither profit nor loss occurs. We quantify this profit as utility [13, 38], considering the cost $c \in [0, 1]$ of a positive decision, as: $\mathcal{U}(\pi; c) = \sum_{x,s} \pi(D{=}1 \mid x, s) \cdot (\ell(Y = 1 \mid x, s) - c)\, \mu_\pi(x \mid s)\gamma(s)$, where $\pi(D{=}1 \mid x, s)$ is the probability of a positive policy decision, $\ell(y \mid x, s)$ the positive ground truth distribution, $\mu_\pi(x \mid s)$ the stationary group-dependent feature distribution, and $\gamma(s)$ the distribution of the sensitive feature. A bank's objective may be to maximize utility (minimize financial loss, i.e., $J_{\text{LT}} := -\mathcal{U}(\pi, c)$). In contrast, a non-profit organization may aim to constrain its policy by maintaining a minimum profit level $\epsilon \geq 0$ over the long term to ensure program sustainability ($C_{\text{LT}} := \mathcal{U}(\pi; c) - \epsilon$).

6.2 Predictive Fairness

Ensuring long-term predictive fairness can help a policymaker meet regulatory requirements and maintain public trust. A common example of such group unfairness notion is *equal opportunity* [25]. This notion measures the disparity in the loan approval rate for eligible applicants based on their demographics, here formalized as: $\texttt{EOPUnf}(\pi) = \mid \mathbb{P}_\pi(D=1 \mid Y=1, S=0) - \mathbb{P}_\pi(D=1 \mid Y=1, S=1) \mid$, where $\mathbb{P}_\pi(D=1 \mid Y=1, S=s) = \frac{\sum_x \pi(D=1|x,s)\ell(Y=1|x,s)\mu_\pi(x|s)}{\sum_x \ell(Y=1|x,s)\mu_\pi(x|s)}$.

A policymaker may now define a maximum tolerable unfairness threshold as $\epsilon \geq 0$ to be held at equilibrium $C_{\text{LT}} := \epsilon - \texttt{EOPUnf}$. Alternatively, they may aim to minimize predictive unfairness EOPUnf in the long term by imposing $J_{\text{LT}} := \texttt{EOPUnf}(\pi)$. Our framework also allows for other group fairness criteria, e.g., demographic parity [19] or sufficiency [12].

In this section, we presented different long-term goals as illustrative examples for lending policies. Refer to § C for examples of fairness objectives or constraints on the feature distribution (e.g., equal qualification [86, 91]) or constraints placed on the type of policy that can be deployed (e.g., monotonicity). This section serves as a starting point for discussions on long-term fairness objectives. We strongly encourage exploring different targets by consulting research from social sciences and economics and involving affected communities in defining these objectives. In the following section, we validate our approach empirically and showcase the interplay between short- and long-term interventions.

7 EXPERIMENTAL RESULTS

We validate our proposed optimization problem in semi-synthetic simulations, where we initialize distributions using real-world data and assume dynamics similar to prior work [10, 82, 86, 87, 91]. We first demonstrate that the policy solution, if found, converges to the targeted stationary state under known dynamics and is robust to covariance shift (§ 7.1). We then assume dynamics are unknown (§ 7.2) and estimate them from a historical temporal dataset. Finally, we illustrate how our approach facilitates distinguishing between short-term and long-term interventions (§ 7.3). For dataset and setup details, see § D; for additional results, including exploration of different dynamics and long-term targets, see § E. While our experiments in finite state and action spaces serve as a proof of concept, we acknowledge that addressing complex dynamics, larger or continuous state and action spaces, and exploring more sophisticated optimization methods are crucial directions for future research. Our code is available at github.com/mrateike/designing-long-term-fair-policies.

Data and Dynamics. We conduct experiments on two datasets utilizing the graphical model introduced in § 2. The FICO loan repayment dataset [4, 68] includes a one-dimensional credit score X, which we discretize into four categories, along with binary attribute race S and repayment behavior Y. From the COMPAS recidivism dataset [44], we choose two-dimensional features X (age category and priors_count), and use binary attributes race S and 2-year-recidivism Y. We get static probabilities ℓ and γ and the starting distribution μ_0 from the probabilities provided (FICO) or estimate them from samples (COMPAS). Since the datasets are static and thus lacking information on feature changes in response to decisions, we assume dynamics g.[4] If not stated otherwise, we assume one-sided dynamics that are characterized by a particular (usually positive) decision leading to changes in a feature distribution, while other decisions do not incur any feature changes. Following prior work [17, 46], we assume that if an applicant defaults on their loan, their credit score remains the same; if the applicant repays the loan, their credit score is likely to increase (the higher the better). Similar to previous research, when dynamics are artificially assumed, our findings may lack generalizability.

Optimization Problem. We now exemplify a long-term target. Consider a bank that aims to maximize its profit ($\mathcal{U}$) while guaranteeing equal opportunity (EOPUnf) for loan approval. Given cost of a positive decision c and a small tolerated unfairness level ϵ, we seek for a policy:

$$\begin{aligned} &\pi^\star_{\texttt{EOP}} := \arg_\pi \max \mathcal{U}(\pi; c) \\ &\text{subj. to} \quad \texttt{EOPUnf}(\pi) \leq \epsilon;\ C_{\text{conv}}(T_\pi), \end{aligned} \tag{5}$$

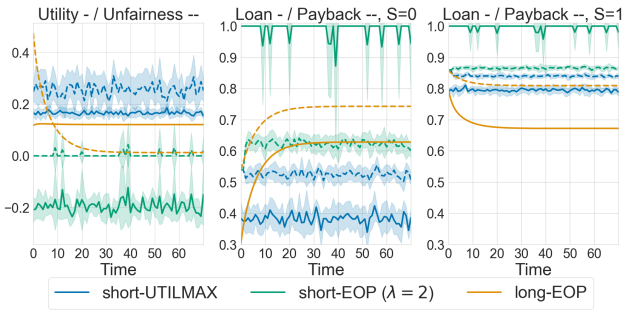
This target has been proposed for fair algorithmic decision-making in static systems [25]. Short-term policies aiming to fulfill this target at each time step have been examined in dynamical systems [15, 17, 91] and it has been imposed as long-term target [82]. We redefine this concept as a long-term goal for the stationary distribution to satisfy. We first apply the general principle (4) to formulate an optimization problem via long-term objectives J_{LT} and long-term constraints C_{LT} and convergence constraints C_{conv}. Next, we solve the optimization problem. Using the found policy $\pi^\star$ and the resulting Markov kernel $T_{\pi^\star}$, we generate the feature distribution across 200 steps. We solve the problem using the Sequential Least Squares Programming method from scikit-learn [61], initializing it (warm start) with a uniform policy where all decisions are random ($\pi(D=1 \mid x, s) = 0.5\ \forall s, x$). Hereafter, we refer to a policy as fair if it meets the ϵ-fairness criteria.

7.1 Convergence to Fair Target and Temporal Stability

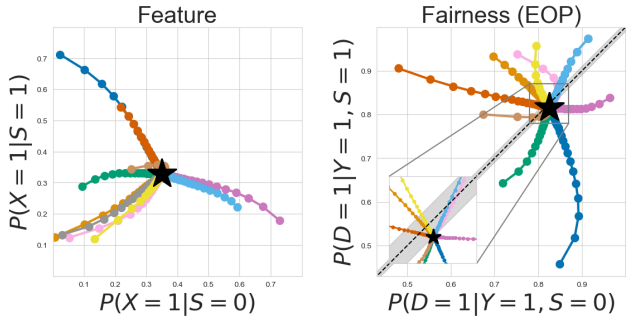
We use the FICO dataset to demonstrate that the policy, solution to the optimization problem specified above, converges to and maintains the targeted distribution without requiring updates, establishing a consistent decision-making framework. We then demonstrate that the identified policy converges to the same equilibrium across populations with different feature distributions but shared dynamics. This can be seen as a form of domain adaptation. This is particularly valuable when dynamics are unknown, as discussed in the following subsection. Meanwhile, this section serves as a proof of concept, assuming known dynamics. Additional results are in § E.1 and § E.2.

Convergence to Fair Target. We first validate that our policy is converging to the targeted fair state and compare it to both fair and unfair short-term policies. Figure 2a displays utility $\mathcal{U}$ and unfairness EOPUnf. Using the initial distribution $\mu_0(x \mid s)$ from FICO, we solve the optimization problem (5) for tolerated unfairness $\epsilon = 0.01$. The short-term policies consist of Logistic Regression models for

[4] While in practical situations expert knowledge can be employed to make assumptions about dynamics, caution is needed to prevent confirmation bias [57].



(a) Comparison of our long-term-EOP policy ($\epsilon = 0.01$) with unfair short-term-UTILMAX policy and fair short-term-EOP policy ($\lambda = 2$). Short-term policies over 10 seeds. On x-axis time steps, on y-axis left: utility $\mathcal{U}$ (solid, ↑), EOP-Unfairness `EOPUnf` (dashed, ↓); middle-right: loan probability $P(D=1 \mid S=s)$ (solid) and payback probability $P(Y=1 \mid S=s)$ (dashed) per sensitive s.



(b) Convergence of $\pi^{\star}_{\text{EOP}}$ to unique stationary distribution $\star$. Left: feature distribution (example of $X = 1$), right: EOP-fairness dashed, ϵ-EOP-fairness gray ($\epsilon = 0.01$). 200 time steps. Colors: random initial feature distributions. $c = 0.8$.

Figure 2: Validation of our policy: (a) Convergence to fair target and temporal stability, (b) Robustness to covariance shift.

10 random seeds, which are retrained at each time step; fairness is enforced using a Lagrangian approach ($\lambda = 2$). Our policy demonstrates high stability in both utility and fairness compared to short-term policies, which exhibit high variance across time. Note since our policy does not require training, we do not report standard deviation over different seeds. Furthermore, while our policy converges to the same fairness level as the short-term fair policy, it experiences only a marginal reduction in utility compared to the (unfair) utility-maximizing short-term policy. Thus, it does not suffer from a fairness-utility trade-off to the extent observed in the short-term policies. Figure 2a (middle, right) displays loan $\mathbb{P}(D=1 \mid S=s)$ and payback probabilities $\mathbb{P}(Y=1 \mid S=s)$ for non-privileged ($S = 0$) and privileged ($S = 1$) groups. The short-term fair policy achieves fairness by granting loans to everyone. For the utility-maximizing short-term policy, unfairness arises as the gap between individuals' probability of paying back and the probability of receiving a loan is much smaller for the privileged group. For our long-term policy, we observe that loan provision probabilities converge closely for both groups over time, while the gap between payback probability and loan granting probability remains similar between groups. Similar to prior research [82, 87], we observe that our policy achieves long-term objectives, but the convergence phase may pose short-term fairness challenges. In practice, it is essential to assess the potential impact of this on public trust.

Robustness to Covariate Shift. We learn the policy based on the initial FICO feature distribution ($\epsilon = 0.01, c = 0.8$) and subsequently conduct simulations using this policy with various randomly sampled initial feature distributions $\mu_0(x \mid s)$. Figure 2b displays the trajectories of the distributions. On the left side, we show the feature distribution for risk score $X = 1$ for both sensitive groups (S), on the right side, we show the group-dependent acceptance probability ($D=1$) among qualified individuals ($Y=1$). Recall that an EOP-fair policy requires these conditional acceptance rates to be equal (dashed diagonal line). The region satisfying the relaxed fairness constraint ϵ-EOP-fairness is shaded in gray. We observe that while the initial starting point affects the convergence process and time, our policy consistently converges to a single feature distribution that is an ϵ-EOP-fair stationary distribution (star within the gray area). In our example, this implies that a policy learned from historical data in one financial institution can be effectively applied to a population in another institution with different credit score distributions, provided there are shared mechanisms governing dynamics and repayment behavior.

7.2 Policy Learning Under Unknown Dynamics

The previous section demonstrated a proof of concept assuming known probability estimates and dynamics. Here, we expand our methodology to address unknown probabilities and dynamics. We presume access to a temporal dataset collected with an (unknown) suboptimal policy. As most datasets for fair decision-making are static, we generate a temporal dataset from the static COMPAS dataset [44] with 5278 samples through simulations akin to prior work [91]. In the offline approach, we derive the policy from probabilities estimated from the historical temporal dataset. For comparison, we include results obtained by adapting our framework to online learning. In the cold start scenario, we initialize the learning process with an uninformed guess and iteratively collect the historical dataset by sampling 500 data points at each step. In the warm start, we use probability estimates from the offline approach as an informed initial guess and iteratively update them by collecting new data and deploying the learned policy to the entire population for 10 time steps. These online approaches, however, introduce challenges related to delayed feedback, periodic policy updates, and exploration costs, as encountered in prior work. The learned policies are deployed in simulated environments under true dynamics for 200 time steps. Refer to § E.6 for details.

Results. Table 1 reports utility and EOP-unfairness at equilibrium as well as cumulatively over 200 time steps. Equilibrium measures indicate the effectiveness of the approach in achieving the target, while cumulative measures reflect the convergence cost, including

Table 1: Comparing learning $\pi^{\star}_{\text{EOP}}$ under known (`true`) and unknown dynamics: online with a cold start, offline from historical data collected with suboptimal policy, online with a warm start from historical data. COMPAS dataset, $\epsilon = 0.01$, $c = 0.6$. Top: at equil. ($\star$); bottom: cumulative ($\sum_{t=1}^{200}$). Mean ± std over 5 seeds.

		Utility (↑)	Unfairness (↓)
At $\star$	Known	0.0489	0.0089
	Offline	0.0486 ± 0.0006	0.0697 ± 0.0071
	Online cold	0.0425 ± 0.006	0.0797 ± 0.0162
	Online warm	0.0486 ± 0.0003	0.0103 ± 0.0067
Cumul.	Known	9.8095	3.3142
	Offline	9.7645 ± 0.1033	13.0371 ± 1.3748
	Online cold	8.3493 ± 1.0915	15.4073 ± 3.0694
	Online warm	9.3488 ± 0.0672	5.0753 ± 1.48

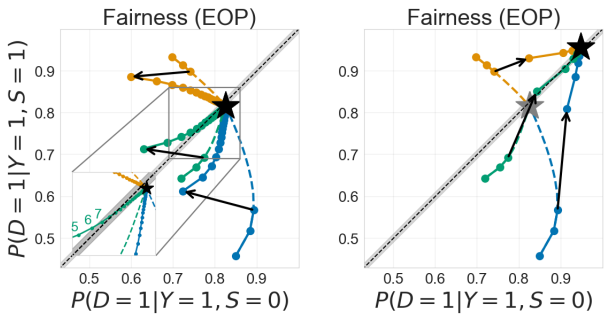


Figure 3: Convergence of $\pi^{\star}_{\text{EOP}}$ to equilibrium $\star$ under interventions ($\rightarrow$). Colors: initial feature distributions. Dashed colored: convergence w/o intervention to $\star$. Left: Short-term interv.; right: Long-term interv. $\epsilon = 0.01$, $c = 0.8$, FICO. Gray diagonal: ϵ-fairness. Green numbers: time steps.

the learning phase in online approaches. Overall, our results indicate that our approach can uncover long-term fair policies (from historical data), even under partially observed labels. These policies, upon deployment, converge to a fair target without requiring additional policy updates. As in prior work, reaching the targeted unfairness value largely depends on the quantity and quality of available training data. At equilibrium, the offline approach exhibits higher utility and lower unfairness compared to the online approach with a cold start. Yet unfairness remains beyond the targeted ϵ due to the misestimation arising from the limited dataset size available. Conversely, the online warm start approach, aiming to enhance the offline information through real-time interaction with the environment, achieves significantly lower unfairness close to the fair target. Cumulative results reveal that online learning of the long-term fair policy with a cold start incurs, as anticipated, higher costs in terms of utility and unfairness, and also high variance. The online warm start demonstrates a relatively small decrease in accumulated utility compared to offline learning, suggesting low costs associated with learning. The findings suggest that exploring the combination of offline learning and periodic policy updates with new information could be a promising avenue for future research on long-term fair policies, effectively balancing delayed feedback effects and accurate environment estimation. Next, we demonstrate how interventions can complement the deployment of long-term fair policies.

7.3 Short- and Long-Term Interventions During Policy Deployment

We now empirically demonstrate the interplay between short- and long-term interventions during policy deployment following Remark 4.2. We use the same setup as in § 7.1 and deploy the learned policy $\pi^{\star}_{\text{EOP}}$.

In Figure 3 (left), we show the convergence of the policy under a covariance shift occurring at $t = 3$ during deployment (short-term intervention). Such shifts may result from economic shocks or governmental subsidy programs, altering financial assets and consequently changing risk scores. We assume that these interventions do not impact the underlying dynamic mechanisms. We observe that such a change in the feature distribution leads to a shift in the fairness measure in the short term. For the orange population, the shift increases unfairness, moving it further from a fair distribution (dashed black line) and prolonging the trajectory to the equilibrium compared to the path without intervention. Conversely, for the green and blue populations, the intervention leads to a trajectory closer to a fair distribution. For the green population, fairness is achieved at step 7, and the trajectory remains fair thereafter until convergence. In the long term, the policy converges to the same targeted stationary distribution as it would without the external distribution shift (dashed colored trajectories). Our long-term fair policy thus guarantees convergence to the targeted fair equilibrium under unexpected short-term interventions, requiring no policy updates. *This means that a short-term intervention can function as an effective mechanism to ensure that achieving long-term fairness does not compromise short-term fairness.*

In Figure 3 (right), we show the convergence of the fixed learned policy under a shift of dynamics $g(k \mid x, d, y, s)$ occurring at $t = 3$ during deployment (long-term intervention on transition probabilities). Specifically, we assume the deployment of a (fair) recourse policy [34, 76] advising individuals facing negative decisions on altering their features to increase the likelihood of acceptance in the next time step. We note a shift in the resulting equilibrium (from $\star$ to $\star$). While both equilibria are fair (within gray ϵ-fair area), this means that the implementation of a fair recourse policy results in the approval of more loans for individuals likely to repay - (from accepting $\sim$81% to $\sim$94%). *This implies that given a long-term fair policy, intervening in other mechanisms that define the kernel can result in a more favorable fair equilibrium.* Additional results in § E.7 show that the fair recourse policy leads to an increase in both utility and fairness accumulated over time.

Our results show that short- and long-term interventions can effectively guide a long-term fair policy toward more favorable convergence behavior and equilibria, respectively. Next, we discuss strengths and limitations of our approach.

8 DISCUSSION: STRENGTHS AND LIMITATIONS

In this section, we discuss key assumptions and limitations. Additional discussion can be found in § B.

Modeling Assumptions. In this paper, we take several modeling assumptions that need to be carefully validated in practice as they simplify complex societal systems. First, we use Markov chains (MCs) to model the behavior of a system. Markov Decision Processes (MDPs), are a specific type of MCs and are widely recognized for providing a flexible approach to long-term fair policy learning [10, 64, 82, 85–87]. They operate under the Markov assumption that the future state depends solely on the current state and action. While effective in simplifying sequential decision-making, this assumption may not always hold in practice. A borrower's creditworthiness can be influenced by past behaviors and credit history, not accounted for in the Markov assumption. Yet, if a credit score incorporates all relevant historical financial information, the Markov assumption may be considered valid. Second, we make several assumptions about sensitive attributes. While these align with prior work [15, 82, 86, 91], they simplify the complexity of social concepts. Group fairness traditionally requires categorizing individuals into non-overlapping groups, which assumes meaningful divisions, neglects intersections of identity [20], and fails to account for those who reject predefined labels altogether. In our data generative model, the sensitive attribute is a root node and remains static over time, which may fail to capture the nature of social concepts within sociocultural contexts [30, 77].

Assumptions on Dynamics. The proposed general optimization problem (4) assumes a time-homogeneous kernel and thus dynamics defining it. Although real-world data often change over time, we treat the dynamics as static for a shorter duration. This is plausible, if they rely on bureaucratic [46] or algorithmic recourse policies [34] and is a common assumption in prior work [17, 86, 91]. If, however, the transition probabilities become time-dependent, updating the policy would be necessary. Further, we conceptualize dynamics as shifts in the distribution of features across an entire population. In this, we build upon the modeling assumptions and dynamics established by prior work [17, 46, 91]. While our current approach provides valuable insights into the macro-level dynamics of feature distributions (see also § E.4 and § E.5), we believe it would be interesting to explore more complex dynamics inspired by game theory [56, 58, 63]. This could offer a deeper understanding of the intricate interplay between individuals and features within the population. For offline policy learning, we assume a sufficient historical temporal dataset to learn the Markov kernel. For online learning, the conditions of aperiodicity and irreducibility automatically ensure exploration during data collection.

The Case of Non-existence of a Long-Term Fair Policy. In situations where a fair policy exists, our optimization problem (3) is designed to effectively discover it. Consider, however, the case that a solution does not exist. Then, as argued in § 3, no policymaker with different strategies of finding policies over time would find a solution to the same problem, with the same assumed distributions, dynamics, and constraints. If a solution to our optimization problem does not exist, this insight may prompt practitioners to explore alternative approaches for long-term fairness, such as redefining the fair state or non-stationary objectives [91].

Interplay of Long-Term and Short-Term Goals. Our framework aims to fulfill fairness in the long term. Prioritizing long-term objectives offers the potential to transform historical disparities [78]. This can temporarily come at the cost of reduced utility and fairness in the short term [86]. While strict adherence to short-term fairness may result in inferior long-term results [17, 46, 91], focusing on long-term goals alone becomes a risk if it diverts attention from population needs that require immediate intervention. If these needs are unaddressed, it may prompt individuals to alter their behavior, resulting in time-variant dynamics that constrain the adaptability of any policy learning approach. Our results in § 7.3 demonstrate that deploying a long-term fair policy together with short-term interventions on the feature distribution, can lead to fulfilling short-term fairness without compromising convergence to the long-term target. In our example, short-term interventions could take the form of subsidies either from the government or from the bank itself—to comply with regulations or due to the benefit of higher utility and customer satisfaction. We defer an in-depth study of the trade-off between short- and long-term fairness and the efforts to quantify it to future work.

9 SUMMARY AND OUTLOOK

We have introduced a general framework for achieving long-term fairness in dynamical systems, where algorithmic decisions in one time step impact individuals' features in the next time step, which are consequently used to make decisions. We proposed a technical approach for identifying a time-independent policy that is guaranteed to converge to a targeted fair stationary state, regardless of the initial data distribution. We model the system dynamics with a time-homogeneous Markov chain and enforce the conditions of the Markov chain convergence theorem to the Markov kernel through policy optimization. The theoretical results from this paper hold for general state spaces. Our framework can be applied to different dynamics and long-term fair goals. We demonstrate this in a guiding example of credit lending assuming a finite state space. In semi-synthetic simulations, we show the effectiveness of policy solutions to converge to targeted stationary population states in a stable manner. Our work extends our understanding of long-term fairness and the short- and long-term effects of algorithmic and societal interventions. Future work lies in applying our framework to a wider range of problems with more complex dynamics larger and continuous feature spaces, multiple sensitive attributes, and using more sophisticated optimization methods. This includes exploring the use of temporal datasets from the fair recommending literature [10] and designing long-term interventions beyond decision-making, such as learning long-term fair recourse policies.

ACKNOWLEDGMENTS

Special thanks to Diego Baptista Theuerkauf for providing continuously insightful feedback and support throughout the evolution of this paper. We also thank Jonas Klesen, Rose Hoberman and all anonymous reviewers for reviews and valuable comments. Thank you also to Lucas Dixon for fruitful discussions.

Miriam Rateike was supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B, ELLIS unit Amsterdam, and the 2023 Google PhD Fellowship in Machine Learning.

This work has been partially funded by the European Union (ERC-2021-STG, SAML, 101040177). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

REFERENCES

[1] Robert Adragna, Elliot Creager, David Madras, and Richard Zemel. 2020. Fairness and robustness in invariant learning: A case study in toxicity classification. *arXiv preprint arXiv:2011.06485* (2020).

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. In *Proceedings of the International Conference on Machine Learning*, Vol. 35. 60–69.

[3] Søren Asmussen and Peter W. Glynn. 2010. Harris Recurrence and MCMC: A Simplified Approach. *Research Report* 6 (2010). Thiele Centre for Applied Mathematics in Natural Science.

[4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning - Limitations and Opportunities*. URL: http://www.fairmlbook.org [Online; Accessed 22.06.2023].

[5] Flavia Barsotti and Rüya Gökhan Koçer. 2022. MinMax fairness: from Rawlsian Theory of Justice to solution for algorithmic bias. *AI & SOCIETY* (2022), 1–14.

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

[7] Somenath Biswas. 2022. Various proofs of the Fundamental Theorem of Markov Chains. *arXiv preprint arXiv:2204.00784* (2022).

[8] Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. 2018. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of ACM Conference on Recommender Systems*, Vol. 12. 224–232.

[9] Yifang Chen, Alex Cuellar, Haipeng Luo, Jignesh Modi, Heramb Nemlekar, and Stefanos Nikolaidis. 2020. Fair contextual multi-armed bandits: Theory and experiments. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*. PMLR, 181–190.

[10] Jianfeng Chi, Jian Shen, Xinyi Dai, Weinan Zhang, Yuan Tian, and Han Zhao. 2022. Towards Return Parity in Markov Decision Processes. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*. PMLR, 1161–1178.

[11] Silvia Chiappa. 2019. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7801–7808.

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

[13] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Vol. 32. 797–806.

[14] Bruce A Craig and Peter P Sendi. 2002. Estimation of the transition matrix of a discrete-time Markov chain. *Health economics* 11, 1 (2002), 33–42.

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

[16] Mark Cutler, Thomas J Walsh, and Jonathan P How. 2015. Real-world reinforcement learning via multifidelity simulators. *IEEE Transactions on Robotics* 31, 3 (2015), 655–671.

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

[18] Darrell Duffie and Peter Glynn. 2004. Estimation of continuous-time Markov processes sampled at random time intervals. *Econometrica* 72, 6 (2004), 1773–1808.

[19] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the Innovations in Theoretical Computer Science Conference*, Vol. 3. 214–226.

[20] James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. An intersectional definition of fairness. In *IEEE International Conference on Data Engineering*, Vol. 36. 1918–1921.

[21] Ari Freedman. 2017. Convergence theorem for finite markov chains. *Proc. REU* (2017).

[22] Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. 2022. Predictably unequal? The effects of machine learning on credit markets. *The Journal of Finance* 77, 1 (2022), 5–47.

[23] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. *Comput. Surveys* 46, 4 (2014), 1–37.

[24] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. 2016. Strategic classification. In *Proceedings of the ACM Conference on Innovations in Theoretical Computer Science*. 111–122.

[25] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems* 29 (2016).

[26] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. In *Proceedings of the International Conference on Machine Learning*. PMLR, 1929–1938.

[27] Hoda Heidari, Vedant Nanda, and Krishna Gummadi. 2019. On the Long-term Impact of Algorithmic Decision Policies: Effort Unfairness and Feature Segregation through Social Learning. In *Proceedings of the International Conference on Machine Learning*. PMLR, 2692–2701.

[28] Lily Hu and Yiling Chen. 2018. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the Conference on World Wide Web*. 1389–1398.

[29] Lily Hu and Yiling Chen. 2018. Welfare and distributional impacts of fair classification. *FAT/ML Workshop at the 35th International Conference on Machine Learning* (2018).

[30] Lily Hu and Issa Kohler-Hausmann. 2020. What's sex got to do with machine learning?. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 513–513.

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

[32] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2017. Fairness in reinforcement learning. In *Proceedings of the International Conference on Machine Learning*. PMLR, 1617–1626.

[33] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. 2016. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*. 325–333.

[34] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2022. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *Comput. Surveys* 55, 5 (2022), 1–29.

[35] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 2021. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 353–362.

[36] Amir-Hossein Karimi, Julius Von Kügelgen, Bernhard Schölkopf, and Isabel Valera. 2020. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *Advances in Neural Information Processing Systems* 33 (2020), 265–277.

[37] Niki Kilbertus, Philip J Ball, Matt J Kusner, Adrian Weller, and Ricardo Silva. 2020. The sensitivity of counterfactual fairness to unmeasured confounding. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*. PMLR, 616–626.

[38] Niki Kilbertus, Manuel Gomez Rodriguez, Bernhard Schölkopf, Krikamol Muandet, and Isabel Valera. 2020. Fair decisions despite imperfect predictions. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*. PMLR, 277–287.

[39] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. Algorithmic fairness. In *American Economic Association Papers and Proceedings*, Vol. 108. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 22–27.

[40] Ronny Kohavi and Barry Becker. 2013. UCI machine learning repository. URL: http://archive.ics. uci.edu/ml/datasets/Adult [Online; 22.06.2023].

[41] Dieter Kraft. 1988. A software package for sequential quadratic programming. *Forschungsbericht- Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt* (1988).

[42] Thanard Kurutach, Ignasi Clavera, Yan Duan, Aviv Tamar, and Pieter Abbeel. 2018. Model-ensemble trust-region policy optimization. *Proceedings of the International Conference on Learning Representations* (2018).

[43] Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in Neural Information Processing Systems* 30 (2017), 4069–4079.

[44] Jeff Larson, Vaggelis Atlidakis, and Marjorie Roswell. 2016. ProPublica COMPAS analysis and dataset. https://github.com/propublica/compas-analysis [Online; Accessed 22.06.2023].

[45] Yingying Li, Aoxiao Zhong, Guannan Qu, and Na Li. 2019. Online markov decision processes with time-varying transition probabilities and rewards. In *ICML Workshop on Real-world Sequential Decision Making*, Vol. 3.

[46] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed Impact of Fair Machine Learning. In *Proceedings of the International Conference on Machine Learning*, Vol. 35.

[47] Jie Lu, Anjin Liu, Fan Dong, Feng Gu, Joao Gama, and Guangquan Zhang. 2018. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering* 31, 12 (2018), 2346–2363.

[48] Maggie Makar and Alexander D'Amour. 2022. Fairness and robustness in anti-causal prediction. In *ICML Workshop on Spurious Correlations, Invariance and Stability*.

[49] Natalia Martinez, Martin Bertran, and Guillermo Sapiro. 2020. Minimax pareto fairness: A multi objective perspective. In *Proceedings of the International Conference on Machine Learning*. PMLR, 6755–6764.

[50] Tatsuya Matsushima, Hiroki Furuta, Yutaka Matsuo, Ofir Nachum, and Shixiang Gu. 2021. Deployment-efficient reinforcement learning via model-based offline optimization. *Proceedings of the International Conference on Learning Representations*.

[51] Sean P Meyn and Richard L Tweedie. 2012. *Markov chains and stochastic stability*. Springer Science & Business Media.

[52] Vishwali Mhasawade and Rumi Chunara. 2021. Causal multi-level fairness. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 784–794.

[53] Microsoft. 2020. Tempeh Repository: TEst Machine learning PErformance exHaustively. https://github.com/microsoft/tempeh.

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

[55] Anusha Nagabandi, Gregory Kahn, Ronald S Fearing, and Sergey Levine. 2018. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. In *IEEE International Conference on Robotics and Automation*. IEEE, 7559–7566.

[56] Adhyyan Narang, Evan Faulkner, Dmitriy Drusvyatskiy, Maryam Fazel, and Lillian J Ratliff. 2023. Multiplayer performative prediction: Learning in decision-dependent games. *Journal of Machine Learning Research* 24, 202 (2023), 1–56.

[57] Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology* 2, 2 (1998), 175–220.

[58] Shayegan Omidshafiei, Christos Papadimitriou, Georgios Piliouras, Karl Tuyls, Mark Rowland, Jean-Baptiste Lespiau, Wojciech M Czarnecki, Marc Lanctot, Julien Perolat, and Remi Munos. 2019. α-rank: Multi-agent evaluation by evolution. *Scientific reports* 9, 1 (2019), 9937.

[59] Błażej Osiński, Adam Jakubowski, Paweł Zięcina, Piotr Miłoś, Christopher Galias, Silviu Homoceanu, and Henryk Michalewski. 2020. Simulation-based reinforcement learning for real-world autonomous driving. In *IEEE International Conference on Robotics and Automation*. IEEE, 6411–6418.

[60] Elisha A Pazner. 1975. *Pitfalls in the theory of fairness*. Technical Report. Discussion Paper.

[61] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

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

[63] Georgios Piliouras and Fang-Yi Yu. 2023. Multi-agent performative prediction: From global stability and optimality to chaos. In *Proceedings of the ACM Conference on Economics and Computation*, Vol. 24. 1047–1074.

[64] Bhagyashree Puranik, Upamanyu Madhow, and Ramtin Pedarsani. 2022. Dynamic positive reinforcement for long-term fairness. *Workshop on Socially Responsible Machine Learning at the 10th International Conference on Learning Representations* (2022).

[65] Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. 2008. *Dataset shift in machine learning*. MIT Press.

[66] Reilly Raab and Yang Liu. 2021. Unintended selection: Persistent qualification rate disparities and interventions. *Advances in Neural Information Processing Systems* 34 (2021), 26053–26065.

[67] Miriam Rateike, Ayan Majumdar, Olga Mineeva, Krishna P Gummadi, and Isabel Valera. 2022. Don't Throw it Away! The Utility of Unlabeled Data in Fair Decision Making. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 1421–1433.

[68] Reserve, U. F. 2007. Report to the congress on credit scoring and its effects on the availability and affordability of credit. In *Board of Govenors of the Federal Reserve System*.

[69] Gareth O. Roberts and Jeffrey S. Rosenthal. 2004. General state space Markov chains and MCMC algorithms. *Probability Surveys* 1 (2004), 20–71.

[70] Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. 2017. When worlds collide: integrating different counterfactual assumptions in fairness. *Advances in Neural Information Processing Systems* 30 (2017).

[71] Michael Scheutzow and Dominik Schindler. 2021. Convergence of Markov Chain transition probabilities. *Electronic Communications in Probability* 26 (2021), 1–13.

[72] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. 2012. On causal and anticausal learning. In *Proceedings of the International Conference on Machine Learning*. 1255–1262.

[73] Jessica Schrouff, Natalie Harris, Sanmi Koyejo, Ibrahim M Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alexander Brown, Subhrajit Roy, Diana Mincu, Christina Chen, et al. 2022. Diagnosing failures of fairness transfer across distribution shift in real-world medical settings. *Advances in Neural Information Processing Systems* 35 (2022), 19304–19318.

[74] Chris Sherlaw-Johnson, Steve Gallivan, and Jim Burridge. 1995. Estimating a Markov transition matrix from observational data. *Journal of the Operational Research Society* 46, 3 (1995), 405–410.

[75] Yifan Sun, Yaqi Duan, Hao Gong, and Mengdi Wang. 2019. Learning low-dimensional state embeddings and metastable clusters from time series data. *Advances in Neural Information Processing Systems* 32 (2019).

[76] Julius von Kügelgen, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller, and Bernhard Schölkopf. 2022. On the fairness of causal algorithmic recourse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 9584–9594.

[77] Sandra Wachter. 2022. The theory of artificial immutability: Protecting algorithmic groups under anti-discrimination law. *Tulane Law Review* 97 (2022), 149.

[78] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2020. Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law. *West Virginia Law Review* 123 (2020), 735.

[79] Thomas J Walsh, Ali Nouri, Lihong Li, and Michael L Littman. 2009. Learning and planning in environments with delayed feedback. *Autonomous Agents and Multi-Agent Systems* 18 (2009), 83–105.

[80] Aline Weber, Blossom Metevier, Yuriy Brun, Philip S Thomas, and Bruno Castro da Silva. 2022. Enforcing delayed-impact fairness guarantees. *arXiv preprint arXiv:2208.11744* (2022).

[81] Marcus Weber. 2017. Eigenvalues of non-reversible Markov chains–A case study. *ZIB Report 17-13* (2017).

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

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

[84] Hao Wu, Andreas Mardt, Luca Pasquali, and Frank Noe. 2018. Deep generative markov state models. *Advances in Neural Information Processing Systems* 31 (2018).

[85] Yuancheng Xu, Chenghao Deng, Yanchao Sun, Ruijie Zheng, Xiyao Wang, Jieyu Zhao, and Furong Huang. 2023. Equal Long-term Benefit Rate: Adapting Static Fairness Notions to Sequential Decision Making. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*.

[86] Tongxin Yin, Reilly Raab, Mingyan Liu, and Yang Liu. 2024. Long-term fairness with unknown dynamics. *Advances in Neural Information Processing Systems* 36 (2024).

[87] Eric Yu, Zhizhen Qin, Min Kyung Lee, and Sicun Gao. 2022. Policy Optimization with Advantage Regularization for Long-Term Fairness in Decision Systems. *Advances in Neural Information Processing Systems* 35 (2022), 8211–8213.

[88] Jia Yuan Yu and Shie Mannor. 2009. Online learning in Markov decision processes with arbitrarily changing rewards and transitions. In *International Conference on Game Theory for Networks*. IEEE, 314–322.

[89] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2019. Fairness Constraints: A Flexible Approach for Fair Classification. *Journal of Machine Learning Research* 20, 75 (2019), 1–42. http://jmlr.org/papers/v20/18-262.html

[90] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, Vol. 54. PMLR, 962–970.

[91] Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellstrom, Kun Zhang, and Cheng Zhang. 2020. How do fair decisions fare in long-term qualification? *Advances in Neural Information Processing Systems* 33 (2020), 18457–18469.

A MARKOV CHAIN CONVERGENCE THEOREM FOR GENERAL STATE SPACES

The theoretical results from this paper hold for general state spaces. This means that for a general state space, a policymaker that aims for fairness will need to fulfill the fundamental objective (3) and thus solve the general optimization problem (4). For general state spaces, this entails, on the one hand, formulating objectives and constraints tailored to such spaces. On the other hand, we need to impose the necessary convergence criteria for the transition kernel defined over the general state space. We then require appropriate methods to efficiently the optimization problem.

In this section, we present the Markov convergence theorem for general state spaces, as well as the conditions to satisfy the conditions of the theorem. These are the convergence criteria we need to impose in the optimization problem (4) for general state spaces. We mainly follow the references of [3, 51, 69, 71].

Notation A.1. *The following notations will be used.*

(1) $\mathcal{X}$ *denotes a standard measurable space (aka standard Borel space), like* $\mathcal{X} = \mathbb{R}^D$ *or* $\mathcal{X} = \mathbb{N}$*, etc.*
(2) *We use* $\mathcal{B}_\mathcal{X}$ *to denote the* σ*-algebra of (Borel subsets of)* $\mathcal{X}$*.*
(3) $T : \mathcal{X} \dashrightarrow \mathcal{X}$ *denotes a Markov kernel (aka transition probability) from* $\mathcal{X}$ *to* $\mathcal{X}$*, i.e. formally a measurable map* $T : \mathcal{X} \to \mathcal{P}(\mathcal{X})$ *from* $\mathcal{X}$ *to the space of probability measures over* $\mathcal{X}$*.*
(4) *For a point* $x \in \mathcal{X}$ *and measurable set* $A \in \mathcal{B}_\mathcal{X}$ *we write* T *similar to a conditonal probability distribution:*

$$T(A|x) := T_x(A) := \text{ probability of } T \text{ hitting } A \text{ when starting from point } x. \tag{6}$$

(5) *We define the Markov kernel* $T^0 : \mathcal{X} \dashrightarrow \mathcal{X}$ *via:* $T^0(A|x) := 1_A(x)$*.*
(6) *We inductively define the Markov kernels* $T^n : \mathcal{X} \dashrightarrow \mathcal{X}$ *for* $n \in \mathbb{N}_1$ *via:*

$$\begin{aligned} T^n(A|x) &:= \int_\mathcal{X} T(A|y)\, T^{n-1}(dy|x) \\ &= \overbrace{(T \circ T \circ \cdots \circ T \circ T)}^{n\text{-times}}(A|x). \end{aligned} \tag{7}$$

Note that: $T^1 = T$*.*
(7) *As the sample spaces we consider the product space:*

$$\Omega := \prod_{n \in \mathbb{N}_1} \mathcal{X}. \tag{8}$$

(8) *For* $n \in \mathbb{N}_1$ *we have the canonical projections:*

$$X_n : \Omega \to \mathcal{X}, \qquad \omega = (x_n)_{n \in \mathbb{N}_1} \mapsto x_n =: X_n(\omega). \tag{9}$$

(9) *We use* $P_x := T_x^{\otimes \mathbb{N}_1}$ *to denote the probability measure on* Ω *of the homogeneous Markov chain induced by* T *that starts at* $X_0 = x$*. Note that for* $n \in \mathbb{N}_1$ *the marginal distribution is given by:*

$$P_x(X_n \in A) = T^n(A|x). \tag{10}$$

(10) *We abbreviate the tuple:* $\mathbf{X} := (X_n)_{n \in \mathbb{N}_1}$*. Note that* $\mathbf{X}$ *is a (homogeneous) Markov chain that starts at* $X_0 = x$ *under the probability distribution* P_x*. We will thus also refer to* $\mathbf{X}$ *as the (homogeneous) Markov chain corresponding to* T*.*
(11) *We abbreviate the probability of the Markov chain of ever hitting* $A \in \mathcal{B}_\mathcal{X}$ *when starting from* $x \in \mathcal{X}$ *as:*

$$L(A|x) := P_x\left(\bigcup_{n \in \mathbb{N}_1} \{X_n \in A\}\right). \tag{11}$$

(12) *We abbreviate the probability of the Markov chain hitting* $A \in \mathcal{B}_\mathcal{X}$ *infinitely often when starting from* $x \in \mathcal{X}$ *as:*

$$Q(A|x) := P_x\left(\{X_n \in A \text{ for infinitely many } n \in \mathbb{N}_1\}\right). \tag{12}$$

(13) *We abbreviate the expected number of times the Markov chain hits* $A \in \mathcal{B}_\mathcal{X}$ *when starting from* $x \in \mathcal{X}$ *as:*

$$\begin{aligned} U(A|x) &:= \sum_{n \in \mathbb{N}_1} T^n(A|x) = \mathbb{E}_x[\eta_A], \\ \eta_A &:= \sum_{n \in \mathbb{N}_1} 1_A(X_n). \end{aligned} \tag{13}$$

Definition A.2 (Irreducibility). T *is called* irreducible *if there exists a non-trivial* σ*-finite measure* ϕ *on* $\mathcal{X}$ *such that for* $A \in \mathcal{B}_\mathcal{X}$ *we have the implication:*

$$\phi(A) > 0 \quad \Longrightarrow \quad \forall x \in \mathcal{X}. \quad L(A|x) > 0. \tag{14}$$

The statement from [51] Prp. 4.2.2 allows for the following remark.

Remark A.3 (Maximal irreducibility measure). *If* T *is irreducible then there always exists a non-trivial* σ*-finite measure* ψ *that is maximal (in the terms of absolute continuity) among all those* ϕ *with property 14. Such a* ψ *is unique up to equivalence (in terms of absolute continuity) and is called a* maximal irreducibility measure *of* T*. For such a* ψ *we introduce the notation:*

$$\mathcal{B}_\mathcal{X}^T := \{A \in \mathcal{B}_\mathcal{X} \mid \psi(A) > 0\}. \tag{15}$$

Note that $\mathcal{B}_\mathcal{X}^T$ *does not depend on the choice of a maximal irreducibility measure* ψ *due to their equivalence. With this notation we then have for irreducible* T*:*

$$A \in \mathcal{B}_\mathcal{X}^T \quad \Longrightarrow \quad \forall x \in \mathcal{X}. \quad L(A|x) > 0. \tag{16}$$

Definition A.4 (Harris recurrence). T *is called* Harris recurrent *if* T *is irreducible and we have the implication:*

$$A \in \mathcal{B}_\mathcal{X}^T \quad \Longrightarrow \quad \forall x \in \mathcal{X}. \quad L(A|x) = 1. \tag{17}$$

Definition A.5 (Invariant probability measures). *An* invariant probability measure (ipm) *of* T *is a probability measure* μ *on* $\mathcal{X}$ *such that:*

$$T \circ \mu = \mu. \tag{18}$$

On measurable sets this can equivalently be re-written as:

$$\forall A \in \mathcal{B}_\mathcal{X}. \qquad \int_\mathcal{X} T(A|x)\, \mu(dx) = \mu(A). \tag{19}$$

Remark A.6. *Note that a general Markov kernel* T *can have either no, exactly one or many invariant probability measures.*

For irreducible T we have the following results from [51] Prp. 10.1.1, Thm. 10.4.4, 18.2.2, concerning existence and uniqueness of invariant probability measures.

Theorem A.7 (Existence and uniqueness of invariant probability measures). *Let* T *be irreducible.*

(1) *Then T has at most one invariant probability measure μ; and:*
(2) *the following are equivalent:*
   (a) *T has an invariant probability measure μ;*
   (b) *the following implication holds for $A \in \mathcal{B}_{\mathcal{X}}$:*

$$A \in \mathcal{B}_{\mathcal{X}}^{T} \quad \Longrightarrow \quad \forall x \in \mathcal{X}. \quad \limsup_{n\to\infty} T^n(A|x) > 0. \tag{20}$$

We have the following properties of invariant probability measures for irreducible T. These are cited from [51] Thm. 9.1.5, Prp. 10.1.1, Thm. 10.4.4, 10.4.9, 10.4.10, and, [71] Prp. A.1, Lem. 3.2.

Theorem A.8 (Properties of irreducible Markov kernels with invariant probability measures). *Let T be irreducible with invariant probability measure μ. Then the following statements hold:*

(1) *μ is a maximal irreducibility measure for T.*
(2) *μ satisfies the following condition for every $A \in \mathcal{B}_{\mathcal{X}}^{T}$ and $B \in \mathcal{B}_{\mathcal{X}}$:*

$$\begin{aligned} \mu(B) &= \int_A \mathbb{E}_x\left[\sum_{n=1}^{\tau_A} 1[X_n \in B]\right] \mu(dx), \\ \tau_A &:= \inf\{n \in \mathbb{N}_1 \,|\, X_n \in A\}. \end{aligned} \tag{21}$$

(3) *There exists a measurable set $\mathcal{H} \in \mathcal{B}_{\mathcal{X}}^{T}$ with $\mu(\mathcal{H}) = 1$ such that:*

$$\forall x \in \mathcal{H}. \quad T(\mathcal{H}|x) = 1, \tag{22}$$

T restricted to $\mathcal{H}$, $T : \mathcal{H} \dashrightarrow \mathcal{H}$, is well-defined and Harris recurrent (with invariant probability measure μ).

Definition A.9 (Aperiodicity). *Let T be irreducible. Then T is called:*

(1) periodic *if there exists $d \geq 2$ pairwise disjoint sets $A_1, \ldots, A_d \in \mathcal{B}_{\mathcal{X}}^{T}$, such that for every $j = 1, \ldots, d$, we have:*

$$\forall x \in A_j. \quad T(A_{j+1 (\mathrm{mod}\, d)}|x) = 1; \tag{23}$$

(2) aperiodic *if T is not periodic.*

With these notation we have the following convergence theorems, see [51] Thm. 13.3.3, 17.0.1, and, [71] Thm. 2.16, 2.17, Assm. 2.12, Prp. 2.2.

Theorem A.10 (Strong Markov chain convergence theorem). *Let μ be a probability measure on $\mathcal{X}$. Then the following are equivalent:*

(1) *T is aperiodic and Harris recurrent and μ is an invariant probability measure for T.*
(2) *For every $x \in \mathcal{X}$ we have the convergence in total variation norm:*

$$\lim_{n\to\infty} \mathrm{TV}(T_x^n, \mu) = 0. \tag{24}$$

Furthermore, if this is the case, then for every $g \in L^1(\mu)$ and every starting point $x \in \mathcal{X}$ we have the convergences:

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} g(X_k) = \mathbb{E}_\mu[g] \quad P_x\text{-a.s.} \tag{25}$$

Theorem A.11 (Markov chain convergence theorem). *Let μ be a probability measure on $\mathcal{X}$. Then the following are equivalent:*

(1) *T is aperiodic and irreducible and μ is an invariant probability measure for T.*
(2) *For every $x \in \mathcal{X}$ we have:*

$$\lim_{n\to\infty} \mathrm{TV}(T_x^n, \mu) < 1, \tag{26}$$

and, for μ-almost-all $x \in \mathcal{X}$ we have the convergence in total variation norm:

$$\lim_{n\to\infty} \mathrm{TV}(T_x^n, \mu) = 0. \tag{27}$$

Furthermore, if this is the case, then for every $g \in L^1(\mu)$ and μ-almost-all starting points $x \in \mathcal{X}$ we have the convergences:

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} g(X_k) = \mathbb{E}_\mu[g] \quad P_x\text{-a.s.} \tag{28}$$

We now want to investigate under which conditions we can achieve irreduciblity, aperiodicity or Harris recurrence. We first cite the results of [3] Thm. 1 and Cor. 1.

Theorem A.12 (Harris recurrence via irreducibility and density). *Let T be irreducible with invariant probability measure μ. Further, assume that T has a density w.r.t. an irreducibility measure ϕ, i.e.:*

$$T(A|x) = \int_A t(y|x)\, \phi(dy), \tag{29}$$

with a jointly measurable $t : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$. Then ϕ is a maximal irreducibility measure for T, μ has a strictly positive density w.r.t. ϕ and T is Harrris recurrent.

Corollary A.13 (Harris recurrence via irreducibility and Metropolis-Hastings form). *Let T be irreducible with invariant probability measure μ. Further, assume that T is of Metropolis-Hastings form w.r.t. an irreducibility measure ϕ:*

$$T(A|x) = (1 - a(x)) \cdot 1_A(x) + \int_A a(y|x) \cdot q(y|x)\, \phi(dy), \tag{30}$$

with jointly measurable $a, q : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ and $a(x) > 0$ for every $x \in \mathcal{X}$. Note that: $a(x) = \int a(y|x) \cdot q(y|x)\, \phi(dy)$. Then ϕ is a maximal irreducibility measure for T, μ has a strictly positive density w.r.t. ϕ and T is Harrris recurrent.

We now have all ingredients to derive the following criteria for the strong Markov chain convergence theorem A.10 to apply:

Corollary A.14 (Criterion for convergence via positive density). *Let ϕ be a non-trivial σ-finite measure on $\mathcal{X}$ such that T has a strictly positive jointly measurable density $t : \mathcal{X}\times\mathcal{X} \to \mathbb{R}_{>0}$ w.r.t. ϕ:*

$$T(A|x) = \int_A t(y|x)\, \phi(dy), \tag{31}$$

then T is irreducible, aperiodic and ϕ is a maximal irreducibility measure for T.

If, furthermore, T has an invariant probability measure μ then μ has a strictly positive density w.r.t. ϕ, T is Harris recurrent and the strong Markov chain convergence theorem A.10 applies.

Corollary A.15 (Criterion for convergence via positive Metropolis-Hastings form). *Let μ be an invariant probability measure of T. Further, assume that T is of Metropolis-Hastings form w.r.t. a non-trivial σ-finite measure ϕ:*

$$T(A|x) = (1 - a(x)) \cdot 1_A(x) + \int_A a(y|x) \cdot q(y|x)\, \phi(dy), \tag{32}$$

with strictly positive jointly measurable $a, q : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{>0}$ such that for every $x \in \mathcal{X}$ we have that:

$$a(x) := \int a(y|x) \cdot q(y|x)\, \phi(dy) \overset{!}{\in} (0, 1). \tag{33}$$

Then ϕ is a maximal irreducibility measure for T, μ has a strictly positive density w.r.t. ϕ, T is aperiodic, Harrris recurrent and the strong Markov chain convergence theorem A.10 applies.

Corollary A.16 (Criterion for convergence on countable spaces). *Let $\mathcal{X}$ be a countable space, i.e. finite or countably infinite. Let T be irreducible with invariant probability measure μ such that for all $x \in \mathcal{X}$ with $\mu(\{x\}) > 0$ we also have $T(\{x\}|x) > 0$. Then T is aperiodic and Harris recurrent and the strong Markov chain convergence theorem A.10 applies.*

B ADDITIONAL CLARIFICATIONS AND DISCUSSION

In this section, we provide additional clarifications and discussion on our proposed framework.

B.1 Existence of a Fair Stationary Distribution

Our approach also serves to determine whether a stationary distribution exists in the first place. In situations where a fair policy does indeed exist, our optimization problem (OP) is designed to effectively discover it. If a solution to our optimization problem does not exist, it implies that alternative methods (including, e.g., reinforcement learning), would also not find a policy inducing and maintaining the targeted fair stationary distribution under the same modeling assumptions. This stems from the fact that if the current state is fair, any alternative approach would still need to address the stationary equation (3) to maintain that state. This discovery can offer valuable insights to practitioners, prompting them to explore different perspectives on long-term fairness. For instance, this might involve revising non-stationary long-term fairness objectives, such as addressing oscillating long-term behaviors [91]. Alternatively, practitioners could consider redefining the targeted fair state that allows for stationary. By shedding light on these possibilities, our approach contributes to a deeper understanding of the dynamics and long-term fairness considerations.

B.2 Strengths and Limitations of a Model-based Approach

Model-based reinforcement learning focuses on scenarios where dynamics are known or can be estimated before policy learning, often requiring fewer samples compared to model-free approaches [42, 50, 55]. This is particularly valuable in delayed feedback scenarios, such as lending [46, 68], recidivism [44], or university admissions [43], where observing feedback may take in the real world months or years. Estimating societal dynamics from historical temporal data, especially for continuous state spaces, remains a challenge [18]. The quality of dynamics estimation in model-based approaches depends on factors like data quantity, quality, coverage, environmental complexity, and estimation methods. In certain scenarios, dynamics may be easier to estimate, especially, if state action and spaces are finite [14, 74]. Moreover, the dynamics considered here, which are driven not by complex human behavior but by policies such as rules governing changes in credit score groups, may contribute to a further simplification of the estimation process. Our approach represents an initial step in model-based learning of long-term fair policies, with potential extensions benefiting from advancements in learning Markov dynamics [75, 84] to address more complex scenarios.

B.3 Opportunities and Limitations of Time-invariant Policies

Our framework yields a single fixed, i.e., time-invariant policy. When the dynamics are constant, and policy learning and estimation of the dynamics occur simultaneously (as in reinforcement learning), then the learned policy requires frequent updates as more data becomes available. Our paper takes a different approach by separating the estimation problem (of the Markov kernel i.e., the dynamics) from the policy learning process and therefore does not require updating the policy. We believe that this holds several advantages, particularly in terms of predictability and trustworthiness. A fixed policy provides a consistent decision-making framework that stakeholders can anticipate and understand contributing to trustworthiness. In addition, a fixed policy simplifies operational processes, such as implementation and maintenance efforts, potentially leading to more efficient and effective outcomes.

When the dynamics vary with time, we can no longer rely on a single time-invariant policy for an infinite time horizon. If, however, the changes are slow and the dynamics remain constant within certain time intervals, our approach remains effective within the time intervals. Whenever the dynamics change, our approach would require re-estimating the dynamics and solving the optimization problem again to obtain a new policy. In this way, our method adapts to changing conditions and maintains its effectiveness over time. However, when dynamics change rapidly, the adaptability of any method is limited.

B.4 Use of Different Datasets

Our experimental section focuses on a two simulation setups, specifically centered around loan repayment based on FICO [68], which has been widely used by previous work on long-term fairness [15, 17, 46, 82, 87]. In addition we also show results on the COMPAS [44] dataset. In § E.5 and § E.4 we provide results for FICO with varying dynamics and initial distributions, essentially simulating different datasets of the same generative model. Note also that we provide an example of how the framework can be applied to a different generative model in § F.

B.5 Additional Related Work

We present additional related work, positioning our study within the broader context of prior research on sequential decision-making and robustness, which pursue different objectives. Our work differs from research on fair sequential decision learning under feedback loops, where decisions made at one time step influence the training data observed at the subsequent step [6, 33, 38, 67]. In this scenario, decisions introduce a sampling bias but do not affect the underlying generative process, as in our case. In our case, decisions influence the underlying data-generating process and consequently shift the data distribution.

Our work also differs from research focused on developing robust machine learning models that can perform well under distribution shifts, where deployment environments may differ from the training data environment [65]. Unlike the line of research that considers various Sources of shift [1, 48, 73], our approach leverages policy-induced data shifts to guide the system towards a state that aligns with our defined long-term fairness objectives. Rather than viewing data shifts as obstacles to overcome, we utilize them as a means to achieve fairness goals in the long term.

C ON LONG-TERM TARGETS

In this section, we offer additional examples and discussions on targeted fair states, exemplified in § 6. We first discuss how our framework allows for the integration of long-term fairness definitions established by prior research (C.1). Following that, we provide more examples of how our framework accommodates various types of constraints. These include distributional fairness concerning individual features (C.2), non-egalitarian fairness objectives (C.3), and constraints on the policy itself that are independent of the long-term equilibrium (C.4).

C.1 Definition of Long-term Fairness

We provide an overview of how our work relates to previously established long-term fairness notions.

Our framework aims to attain a state of long-term fairness. This entails that fairness formulations should be met in the long term and, importantly, once achieved, be maintained. Our goal differs fundamentally from approaches that aim to fulfill fairness at each time step. In this regard, [17] compare agents optimizing for short-term goals - e.g., a profit-maximization agent to an equality of opportunity fair agent and measure the long-term (in)equality of the initial credit score distribution across groups - without imposing it on the agents.

Prior work on long-term fairness introduces parity of return [10], which requires equal (discounted) rewards accumulated by the decision maker over time, where the reward could be defined as the ratio between true positive and overall positive decisions. [82] define long-term demographic parity (equal opportunity) as asking the cumulative expected individual rewards to be on average equal for (qualified members of) demographic groups. [86] aim to maximize the accumulated reward subject to accumulated unfairness (utility) constraint in a finite time horizon. The reward combines true positive and true negative rates, while the authors consider different (un)fairness measures: demographic parity, equal opportunity, and equal qualification rate. [87] formulate a (short-term) fairness metric (e.g., equality of opportunity) as a function of the state and increase its enforcement over time. Our framework provides the capability to enforce these fairness and reward considerations, specifically, we allow for feature complex objective functions (see § 6.1) as well as imposing group fairness criteria in the long-term (see § 6.2) for infinite time-horizons. Note that the formulation of a fair state is not limited to the possible fairness objectives and constraints discussed in § 6. Rather, we exemplify in that section that our framework can capture fairness objectives well-established in prior work [19, 25, 46, 91].

Next we provide examples for distributional fairness within our guiding example, such as equal distribution of credit scores [17], or equal qualification [66, 91].

C.2 Distributional Fairness

Policymakers may be interested in specific characteristics of a population's features X or qualifications Y (ground truth) on a group level [66, 91]. We measure group qualification Q as the group-conditioned proportion of positive labels assigned to individuals as $Q^s(\pi \mid s) = \sum_x \ell(Y = 1 \mid x, s)\mu_\pi(x \mid s)$, where $\ell(Y = 1 \mid x, s)$ is the positive ground truth distribution, and $\mu_\pi(x \mid s)$ describes the stationary group-dependent feature distribution. We measure inequity (of qualifications) as $\mathcal{I} := \mid Q(\pi \mid S = 0) - Q(\pi \mid S = 1) \mid$.

To promote financial stability, a policymaker like the government may pursue two different objectives. Firstly, they may aim to minimize default rates using the objective $J_{\text{LT}} := -\sum_s Q(\pi \mid s)\gamma(s)$. Alternatively, if the policymaker intends to increase credit opportunities, they may seek to maximize the population's average credit score with the objective $J_{\text{LT}} := -\sum_s \frac{1}{|X|} \sum_x \mu_\pi(x \mid s)\gamma(s)$, where $|X|$ represents the state space size. To achieve more equitable credit score distributions, the policymaker could impose the constraint $C_{\text{LT}} := \epsilon - \mid \mu_\pi(x \mid S = 0) - \mu_\pi(x \mid S = 1) \mid \forall x$. However, depending on the generative model, this approach might not eliminate inequality in repayment probabilities. In such cases, the policymaker may aim to ensure that individuals have the same payback ability using the constraint $C_{\text{LT}} := \epsilon - \mathcal{I}$. Note that measuring differences in continuous or high-dimensional distributions requires more sophisticated distance measures. However,equal credit score distributions or repayment probabilities may not guarantee equal access to credit, which can be imposed using predictive group fairness measures as introduced in § 6.2.

It is a well-known result in economics that prioritizing egalitarian distributions may not always align with individual (and societal) preferences [5, 49]. In such cases, it is sometimes more desirable to minimize the maximum societal risk to prevent unnecessary harm. We elaborate on this concept next.

C.3 On Minimax Objectives

While egalitarian allocations can align with societal values, they are generally considered Pareto inefficient [60]. In certain scenarios, policymakers may be interested in minimizing the maximum risk within a society [5]. This approach aims to prevent unnecessary harm by reducing the risk for one group without increasing the risk for another [49]. For instance, in the context of hiring, instead of equalizing the group-dependent repayment rates $Q(\pi, s)$, a policymaker may be interested in minimizing the maximum default risk $1 - Q(\pi, s)$ across groups. In other words, their objective could be $J_{\text{LT}} := \min_s -(1 - Q(\pi, s))$, rather than aiming for equal default or repayment rates.

At times, a policymaker may be concerned not only with the final outcome of the policy, as emphasized in the objectives and constraints introduced in the main paper and in this section above, but also with the way the policy is constructed. This consideration may stem from technical reasons or aim to enhance the perception of fairness in society. We discuss this aspect next.

C.4 Policy constraints

Our framework also allows to incorporate constraints on the type of policy being searched for. These constraints can be imposed on the policy independent of the stationary distribution. We provide an example here. If the features exhibit a monotonic relationship, where higher values of X_t tend to result in a higher probability of a positive outcome of interest $\ell(Y = 1 \mid x, s)$, the policymaker may be interested in a monotonous policy. A monotonous policy assigns higher decision probabilities as X_t increases. This means that if the feature increases (or decreases), the policy's response or outcome should also consistently increase (or decrease). Establishing predictability in the decision-making process is crucial for enhancing perceived fairness. In a lending example, as the risk score improves, there is a predictable increase in the likelihood of obtaining a loan approval. This transparency contributes to a clearer understanding of the decision criteria. In such cases, we can impose the additional constraint $\pi(k, s) \geq \pi(x, s), \forall k \geq x, s$.

D SIMULATION DETAILS

In this section, we present the details of the experiments and simulations that we show in the main paper in § 7 as well as for the additional results reported in § E. Specifically, we provide details on:

- The COMPAS dataset, its pre-processing, and assumptions (D.1)
- Solving the optimization problem for linear state spaces (D.2)
- Assumed underlying dynamics (D.3)
- Offline learning under unknown dynamics (D.4)
- Online learning under unknown dynamics (D.5)
- Computational reSources and run time (D.6)
- Optimization problem for a different fair target (policy $\pi^{\star}_{\text{QUAL}}$) with results reported in the following section (D.7).

D.1 COMPAS Dataset

In the main paper, we provided a detailed overview of how our modeling assumptions for the data generative model over time manifest in the lending scenario. In this section, we present an overview of how we use the COMPAS dataset in the simulations and how the same assumptions are applied to the recidivism scenario.

D.1.1 Pre-processing of Dataset. For our experiments, we use the COMPAS dataset from ProPublica [44]. The target variable Y indicates whether an individual faced rearrest within two years (coded as 1) or remained without rearrest (coded as 0). Importantly, this does not indicate whether the individual re-offended but rather whether the individual re-offended and this was detected or caught by the police within two years. The sensitive attribute S is Race (African-American 0, Caucasian 1).

We preprocess the dataset provided by ProPublica similar to the `tempeh` package [53], this includes features X: The age of defendants (`age`) in years as well as a categorized age feature (`age_cat`) in years (< 25, $25 - 45$, > 45), alongside historical information about prior incidents: counts of prior juvenile misdemeanors (`juv_misd_count`), counts of prior juvenile felonies (`juv_fel_count`), counts of other juvenile incidents that are relevant (`juv_other_count`), count or number of prior offenses or incidents the individual has on their record (`priors_count`), binary indicator if the most recent charge prior to the COMPAS score calculation is a felony or misdemeanor ((`c_charge_degree_F`), `c_charge_degree_M`). We then analyze importance of the features. For this we first train a decision tree classifier to predict the target variable with a 20-80 test-training split. Subsequently, we rank the feature importance from the trained classifier. The results indicate that `age` and `priors_count` are most contributing to the classifiers predictions.

We categorize `priors_count` into four subgroups based on whether the number of priors is 0, 1, $2 - 3$, or, > 3. For the information on an individual's age, we depend on the pre-categorized groups provided by feature `age_cat` (3 age groups). This results in 12 feature combinations (subgroups) and thus values of X.

D.1.2 Assumptions on Data Generative Model. We assume a data generative model for a recidivism scenario assuming a data generative process as in Figure 1)=. The protected attribute S is Race, the non-sensitive feature X_t (`age` and `priors_count`) and an outcome of interest Y_t refers to arrest for re-offense within 2 years. As above, we assume the sensitive attribute to remain immutable over time and drop the attribute's time subscript. For simplicity, we assume binary sensitive attribute and outcome of interest (0 - recidivism, 1 - no-recidivism) and a two-dimensional discrete non-sensitive feature $X_{\texttt{age}} \in \{0, 1, 2\}$ and $X_{\texttt{priors\_count}} \in \{0, 1, 2, 3\}$. We assume the population's sensitive attribute be distributed as $\gamma(s) := \mathbb{P}(S{=}s)$ and remain constant over time. We assume $X_{\texttt{priors\_count}}$ to depend on S, such that the group-conditional feature distribution at time t is $\mu_t(x \mid s) := \mathbb{P}(X_t{=}x \mid S{=}s)$. For example, different demographic groups may have different distributions of arrests for prior offenses due to structural discrimination in society (e.g., predictive policing). However, we assume that age is independent of Race. This assumption is consistent with the generative model. In this model, the presence of an edge, such as one from S to X, signifies a potential causal relationship that might exist but is not guaranteed to manifest.

Outcome of Interest. The outcome of interest Y is assumed to depend on X and (potentially) on S resulting in the label distribution $\ell(y \mid x, s) := \mathbb{P}(Y_t{=}y \mid X_t{=}x, S{=}s)$. The association between the features, age and the historical count of criminal offenses, and the likelihood of rearrest for re-offense, is complex, and establishing direct causation is challenging due to the presence of multiple contributing factors. Regarding the age, younger individuals may face challenges such as not fully developed impulse control, being influenced by peers, and having limited life experience. This can affect decision-making. Further, the lack of exposure to alternative career paths can further increase the risk of repeating criminal behavior and being caught by police. Regarding the count of prior offenses, repeated criminal behavior can create lasting patterns, especially, if rehabilitation is inadequate. Further, social stigma, difficulties in reintegrating into society, and the psychological impact of criminal histories can push individuals back into criminal activities. While the prior count may not directly cause rearrest for re-offense, we assume in this work that it serves as a significant indicator of a complex interplay of factors that increase the probability of individuals returning to criminal behavior. Our empirical findings from the COMPAS dataset suggest that the likelihood of arrest for re-offense decreases with advancing age and increases with a higher count of prior offenses. We explicitly acknowledge

that the assumptions we make here represent a simplified model of the world, neglecting the intricate socio-economic contexts of populations. It is also a valid question, whether incorporating algorithmic recommendations in bail decisions is appropriate at all.

Decision-making Policy. Here, we assume that there exists a policy that takes binary decisions related to pretrial release (0 - jail, 1 - bail), i.e., whether an individual should be granted bail or be held in jail based on X and (potentially) S and decides with probability $\pi(d \mid x, s) := \mathbb{P}(D_t = d \mid X_t = x, S = s)$. In the context of recidivism, we assume that automated pre-trial bail decisions solely depend on an individual's age and prior criminal count. We acknowledge that other variables might contribute to a more comprehensive understanding of an individual's risk profile.

Dynamical System. We denote the probability of an individual with $S = s$ transitioning from features $X_t = x$ to $X_{t+1} = k$ in the next step as the dynamics $g(k \mid x, d, y, s) := \mathbb{P}(X_{t+1} = k | X_t = x, D_t = d, Y_t = y, S = s)$ Importantly, the next step feature state depends only on the present feature state, and not on any past states. We assume that a pre-trail bail decision (D_1) directly impacts an individuals' features age and prior criminal history (X_{t+1}) in the next time step. Further, the transition from the current feature state (X_t) to the next state (X_{t+1}) is influenced not only by the decision and the current features but also by the outcome of interest (Y_t) and potentially the sensitive attribute (S). After an individual is released ($D = 1$), their prior criminal count remains the same ($X_{t+1} = X_t$) if they do not get arrested for reoffence ($Y_t = 0$) and increases if they do ($Y = 1$). If an individual is not released ($D = 0$), the prior criminal count remains the same (one-sided feedback, $X_{t+1} = X_t$)). We study an open population, where a small percentage of individuals undergo a transition from one age category to another at each time step independently of the specific nature of decisions and recidivism behavior.

We assume time-independent dynamics $g(k \mid x, d, y, s)$, where feature changes in response to decisions and individual attributes remain constant over time (see also B for a general comment on this assumption). Thus, for the time horizon considered here, we assume that the influence of bail decisions on an individual's prior criminal history does not vary over time. This means we assume no potential changes arising from societal or legal policies, interventions, or rehabilitation programs. We also assume that the distribution of the outcome of interest conditioned on an individual's features $\ell(y \mid x, s)$ remains constant over time. In the context of our example, this means that the probability of an individual re-offending based on their age and prior criminal history is assumed to be consistent across different time periods and not influenced by temporal shifts or changes.

D.2 Solving the Optimization Problem

Our framework can be thought of as a three-step process. First, just as previous work on algorithmic fairness empowers users to choose fairness criteria, our framework allows users to define the characteristics of a fair distribution applicable in their decision-making context (see § 6). The second step involves transforming the definition of fair characteristics into an optimization problem (OP). The third step consists of solving the OP. Given the nature of our optimization problem, which is linear and constraint-based, we can employ any efficient black-box optimization methods for this class of problems. Note that the OP seeks to find a policy π that induces a stationary distribution μ, which adheres to the previously defined fairness targets. As detailed in § 7, in the search of π, we first compute group-dependent kernel T^s_π, which is a linear combination of assumed/estimated dynamics and distributions and π. We then compute the group-dependent stationary distribution μ^s_π via eigendecomposition.

Solving the Optimization Problem for Finite State Spaces. In our guiding example and the corresponding simulation, we consider a time-homogeneous Markov chain $(\mathcal{Z}, P)$ with a finite state space $\mathcal{Z}$ (e.g., credit score categories). Consequently, the convergence constraints C_{conv} are determined by the *irreducibility* and *aperiodicity* properties of the corresponding Markov kernel (see § 4). It has been shown that an $n \times n$ transition matrix P constitutes an irreducible and aperiodic Markov chain if and only if all entries of $(P)^n$ are strictly positive [7]. We can thus impose as convergence constraint C_{conv} for finite states $\sum_{i=1}^{n} \left(T^s_\pi\right)^n > \mathbf{0} \, \forall s$, where n is the number of states ($n = |X|$), and $\mathbf{0}$ denotes the matrix with all entries equal to zero. Following Theorem 4.3, a sufficient condition for convergence to the unique stationary distribution is the positivity of the transition matrix P, where all elements are greater than zero. In our experiments, we ensure that the transition matrix P is positive, as we assume that $g(k \mid x, d, y, s) > 0$ for all d, s, y, x, k, while FICO and COMPAS data already yields $\ell(y \mid x, s) > 0$ for all y, x, s.

We compute the *stationary distribution* μ using eigendecomposition. Recall from Definition 3.1 that a stationary distribution of a time-homogeneous Markov chain $(\mathcal{Z}, P)$ is a probability distribution μ such that $\mu = \mu P$. More explicitly, for every $w \in \mathcal{Z}$, the following needs to hold: $\mu(w) = \sum_z \mu(z) \cdot P(z, w)$. If the transition matrix P is positive, $\mu = \mu P$ implies that μ is the eigenvector of P corresponding to eigenvalue 1. We then solve for the stationary distribution μ using linear algebra.

SLSQP Algorithm. We solve optimization problems (5) and (D.7) using the Sequential Least Squares Programming (SLSQP) method [41]. SLSQP is a method used to minimize a scalar function of multiple variables while accommodating bounds, equality and inequality constraints and can be used for solving both linear and non-linear constraints. The algorithm iteratively refines the solution by approximating the objective function and constraints using quadratic model. Specifically, SLSQP is designed to minimize scalar functions of one or more variables. In our case we are maximizing utility ($\pi^\star_{\mathsf{EOP}}$) or qualifications ($\pi^\star_{\mathsf{QUAL}}$) and searching for $\mathbb{P}(D = 1 \mid X = x, S = s)$ for all x and s, which are with $|X| = 4$ and $|S| = 2$, a total of 8 variables. Further, SLSQP can handle optimization problems with variable bounds. In our case, we set a minimum bound of 0 and a maximum bound of 1 as we are seeking for probabilities $\mathbb{P}(D = 1 \mid X = x, S = s)$ for all x and s. SLSQP can also handle both linear and non-linear equality and inequality constraints. In our example, where the state space is finite (i.e., X is categorical), all constraints are linear inequality or equality constraints. Finally, SLSQP uses a sequential approach, which means it iteratively improves the solution by solving a sequence of subproblems.

This approach often converges efficiently, even for non-convex and non-linear optimization problems.

We use the SLSQP solver from scikit-learn[5] [61] with step size eps $\approx 1.49\times10^{-10}$ and a max. number of iterations 200 and initialize the solver (warm start) with a uniform policy where all decisions are random, i.e., $\pi(D=1\,|\,x,s) = 0.5\;\forall x,s$.

D.3 Assumed Dynamics

We now provide details about the assumed dynamics. Refer to D.3.1 for FICO dynamics details and D.3.2 for assumed COMPAS dynamics.

D.3.1 Dynamics for FICO Lending Example. In our guiding example, we assume binary $s, y, d \in \{0, 1\}$ and four credit categories, i.e., we have $n = |\mathcal{X}| = 4$ states. For simplicity, we assume the following notation: $T_{sdy} := g(k\,|\,x, d, y, s)$. T_{sdy} is a $n \times n$ transition matrix that describes the Markov chain, where the rows and columns are indexed by the states, and $T_{sdy}(x, k)$, i.e., the number in the x-th row and k-th column, gives the probability of going to state $X_{t+1} = k$ at time $t + 1$, given that it is at state $X_t = x$ at time t and given that $S = s$, $D_t = d$, $Y_t = y$.

One-sided Dynamics. For all one-sided dynamics assumed in the main paper in § 7 and in the supplementary E, we assume:

$$T_{000},T_{001},T_{100},T_{101} = \begin{bmatrix} 0.9 & 0.03333 & 0.03333 & 0.03333 \\ 0.03333 & 0.9 & 0.03333 & 0.03333 \\ 0.03333 & 0.03333 & 0.9 & 0.03333 \\ 0.03333 & 0.03333 & 0.03333 & 0.9 \end{bmatrix}$$
$$T_{110},T_{010} = \begin{bmatrix} 0.9 & 0.9 & 0.9 & 0.9 \\ 0.03333 & 0.03333 & 0.03333 & 0.03333 \\ 0.03333 & 0.03333 & 0.03333 & 0.03333 \\ 0.03333 & 0.03333 & 0.03333 & 0.03333 \end{bmatrix} \tag{34}$$

One-sided General. For the one-sided dynamics in § 7.1 we additionally assume dynamics T_{sdy} that depend on the sensitive attribute in addition to (34):

$$T_{111} = \begin{bmatrix} 0.53333 & 0.03333 & 0.03333 & 0.03333 \\ 0.4 & 0.53333 & 0.03333 & 0.03333 \\ 0.03333 & 0.4 & 0.53333 & 0.03333 \\ 0.03333 & 0.03333 & 0.4 & 0.9 \end{bmatrix}$$
$$T_{011} = \begin{bmatrix} 0.33333 & 0.03333 & 0.03333 & 0.03333 \\ 0.6 & 0.33333 & 0.03333 & 0.03333 \\ 0.03333 & 0.6 & 0.33333 & 0.03333 \\ 0.03333 & 0.03333 & 0.6 & 0.9 \end{bmatrix}$$

Two-sided Fair Recourse. For the two-sided dynamics assumed in the main paper in § 7.3 we assume dynamics T_{sdy}:

$$T_{000} = \begin{bmatrix} 0.2 & 0.73333 & 0.03333 & 0.03333 \\ 0.03333 & 0.2 & 0.73333 & 0.03333 \\ 0.03333 & 0.03333 & 0.2 & 0.73333 \\ 0.03333 & 0.03333 & 0.03333 & 0.9 \end{bmatrix}$$
$$T_{001} = \begin{bmatrix} 0.1 & 0.83333 & 0.03333 & 0.03333 \\ 0.03333 & 0.1 & 0.83333 & 0.03333 \\ 0.03333 & 0.03333 & 0.1 & 0.83333 \\ 0.03333 & 0.03333 & 0.03333 & 0.9 \end{bmatrix}$$
$$T_{100},T_{101} = \begin{bmatrix} 0.5 & 0.43333 & 0.03333 & 0.03333 \\ 0.03333 & 0.5 & 0.43333 & 0.03333 \\ 0.03333 & 0.03333 & 0.5 & 0.43333 \\ 0.03333 & 0.03333 & 0.03333 & 0.9 \end{bmatrix}$$
$$T_{010},T_{110} = \begin{bmatrix} 0.9 & 0.03333 & 0.03333 & 0.03333 \\ 0.9 & 0.03333 & 0.03333 & 0.03333 \\ 0.9 & 0.03333 & 0.03333 & 0.03333 \\ 0.9 & 0.03333 & 0.03333 & 0.03333 \end{bmatrix}$$
$$T_{011},T_{111} = \begin{bmatrix} 0.33333 & 0.6 & 0.03333 & 0.03333 \\ 0.03333 & 0.33333 & 0.6 & 0.03333 \\ 0.03333 & 0.03333 & 0.33333 & 0.6 \\ 0.03333 & 0.03333 & 0.03333 & 0.9 \end{bmatrix}$$

Two-sided Recourse Dynamics. For `recourse` dynamics in § E.4, we assume the following dynamics T_{sdy}:

$$T_{000},T_{001} = \begin{bmatrix} 0.7 & 0.03333 & 0.03333 & 0.03333 \\ 0.23333 & 0.7 & 0.03333 & 0.03333 \\ 0.03333 & 0.23333 & 0.7 & 0.03333 \\ 0.03333 & 0.03333 & 0.23333 & 0.9 \end{bmatrix}$$
$$T_{100},T_{101} = \begin{bmatrix} 0.5 & 0.03333 & 0.03333 & 0.03333 \\ 0.43333 & 0.5 & 0.03333 & 0.03333 \\ 0.03333 & 0.43333 & 0.5 & 0.03333 \\ 0.03333 & 0.03333 & 0.43333 & 0.9 \end{bmatrix}$$
$$T_{010},T_{011} = \begin{bmatrix} 0.9 & 0.9 & 0.9 & 0.9 \\ 0.03333 & 0.03333 & 0.03333 & 0.03333 \\ 0.03333 & 0.03333 & 0.03333 & 0.03333 \\ 0.03333 & 0.03333 & 0.03333 & 0.03333 \end{bmatrix}$$
$$T_{110},T_{111} = \begin{bmatrix} 0.33333 & 0.03333 & 0.03333 & 0.03333 \\ 0.6 & 0.33333 & 0.03333 & 0.03333 \\ 0.03333 & 0.6 & 0.33333 & 0.03333 \\ 0.03333 & 0.03333 & 0.6 & 0.9 \end{bmatrix}$$

[5] https://docs.scipy.org/doc/scipy/reference/optimize.minimize-slsqp.html

Two-sided Discouraged Dynamics. For `discouraged` dynamics in § E.4, we assume the following dynamics T_{sdy}:

$$T_{000}, T_{001} = \begin{bmatrix} 0.9 & 0.63333 & 0.13333 & 0.03333 \\ 0.03333 & 0.3 & 0.53333 & 0.23333 \\ 0.03333 & 0.03333 & 0.3 & 0.43333 \\ 0.03333 & 0.03333 & 0.03333 & 0.3 \end{bmatrix}$$

$$T_{100}, T_{101} = \begin{bmatrix} 0.9 & 0.43333 & 0.13333 & 0.03333 \\ 0.03333 & 0.5 & 0.33333 & 0.23333 \\ 0.03333 & 0.03333 & 0.5 & 0.23333 \\ 0.03333 & 0.03333 & 0.03333 & 0.5 \end{bmatrix}$$

$$T_{010}, T_{011} = \begin{bmatrix} 0.9 & 0.9 & 0.9 & 0.9 \\ 0.03333 & 0.03333 & 0.03333 & 0.03333 \\ 0.03333 & 0.03333 & 0.03333 & 0.03333 \\ 0.03333 & 0.03333 & 0.03333 & 0.03333 \end{bmatrix}$$

$$T_{110}, T_{111} = \begin{bmatrix} 0.33333 & 0.03333 & 0.03333 & 0.03333 \\ 0.6 & 0.33333 & 0.03333 & 0.03333 \\ 0.03333 & 0.6 & 0.33333 & 0.03333 \\ 0.03333 & 0.03333 & 0.6 & 0.9 \end{bmatrix}$$

One-sided Slow. For the one-sided `slow` dynamics with results presented in E.5, we assume the following group-independent dynamics T_{sdy} in addition to (34):

$$T_{011} = T_{111} = \begin{bmatrix} 0.53333 & 0.03333 & 0.03333 & 0.03333 \\ 0.4 & 0.53333 & 0.03333 & 0.03333 \\ 0.03333 & 0.4 & 0.53333 & 0.03333 \\ 0.03333 & 0.03333 & 0.4 & 0.9 \end{bmatrix}$$

One-sided Medium. For the one-sided `medium` dynamics in E.5, we assume the following group-independent dynamics T_{sdy} in addition to (34):

$$T_{011} = T_{111} = \begin{bmatrix} 0.33333 & 0.03333 & 0.03333 & 0.03333 \\ 0.6 & 0.33333 & 0.03333 & 0.03333 \\ 0.03333 & 0.6 & 0.33333 & 0.03333 \\ 0.03333 & 0.03333 & 0.6 & 0.9 \end{bmatrix}$$

One-sided Fast. For the one-sided `fast` dynamics with results presented in E.5, we assume the following group-independent dynamics T_{sdy} in addition to (34):

$$T_{011}, T_{111} = \begin{bmatrix} 0.13333 & 0.03333 & 0.03333 & 0.03333 \\ 0.8 & 0.13333 & 0.03333 & 0.03333 \\ 0.03333 & 0.8 & 0.13333 & 0.03333 \\ 0.033335 & 0.03333 & 0.8 & 0.9 \end{bmatrix}$$

D.3.2 Dynamics for COMPAS Recidivism Example. In our guiding example, we assume binary $s, y, d \in \{0, 1\}$ and 12 credit categories, i.e., we have $n = |\mathcal{X}| = 12$ states. As above, for simplicity we assume the following notation: $T_{sdy} := g(k \mid x, d, y, s)$. As described in D.1, we assume an open population, where individuals both enter and exit. Consequently, there are intrinsic dynamics within the system that impact the distribution of features, beyond the influence of the policy decisions and individual reactions. As our dynamics center on population state distributions instead of individuals, we consequently assume a non-zero probability for "moving down" in age category or prior count. This assumption is based on population entries and exits due to births and deaths.

One-sided Dynamics. For all one-sided dynamics presented with COMPAS (in § 7.2, E.6.2 and E.7), we assume the following dynamics T_{sdy}. Note, here we are reporting for simplicity values rounded to two digits after the comma.

$$T_{000}, T_{001}, T_{100}, T_{101}, T_{011}, T_{111} = \begin{bmatrix} 0.8 & 0.01 & 0.01 & 0.01 & 0.19 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.8 & 0.01 & 0.01 & 0.01 & 0.1 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.8 & 0.01 & 0.01 & 0.01 & 0.1 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.8 & 0.01 & 0.01 & 0.01 & 0.1 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.01 & 0.8 & 0.01 & 0.01 & 0.01 & 0.11 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.8 & 0.01 & 0.01 & 0.01 & 0.1 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.8 & 0.01 & 0.01 & 0.01 & 0.1 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.8 & 0.01 & 0.01 & 0.01 & 0.1 \\ 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.9 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.9 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.9 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.9 \end{bmatrix}$$

$$T_{110}, T_{010} = \begin{bmatrix} 0.01 & 0.8 & 0.01 & 0.01 & 0.01 & 0.1 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.8 & 0.01 & 0.01 & 0.01 & 0.1 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.8 & 0.01 & 0.01 & 0.01 & 0.1 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.8 & 0.01 & 0.01 & 0.01 & 0.1 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.8 & 0.01 & 0.01 & 0.01 & 0.1 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.8 & 0.01 & 0.01 & 0.01 & 0.1 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.8 & 0.01 & 0.01 & 0.01 & 0.1 \\ 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.8 & 0.01 & 0.01 & 0.01 & 0.1 \\ 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.9 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.9 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.9 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.9 \end{bmatrix}$$

D.4 Offline Learning Under Unknown Dynamics

In this section, we provide details of the offline learning approach (see COMPAS results in § 7.2). Firstly, use all samples from the dataset to estimate $\mu_0(x \mid s)$ and $\gamma(s)$. Since COMPAS is a static dataset, we simulate a temporal dataset by assuming a suboptimal historical data collection policy, represented by π_0, which could, for instance, be the result of decisions made by a human decision maker. We assume partially observed labels and observe Y_t, but only for individuals with $D = 1$, as we only observe commited crime upon release. Next, we estimate $\ell(y \mid x, s)$ by computing the label distribution within the subset of individuals with observed outcomes.

Subsequent we simulate the updated features X_{t+1} of the next time step assuming one-sided dynamics $g(k \mid x, d, y, s)$. For more details on the dynamics, see § D.3.2. Consequently, we estimate $g(k \mid x, d, y, s)$ from the obtained samples by estimating the conditional distribution. Here, we assume that the policy maker is aware that dynamics are one sided and consequently only need to estimate the dynamics for individuals who received $D = 1$. Given our results serves as a proof of concept, we defer deploying sophisticated estimation methods to future work and literature addressing selective labeling and corresponding estimation errors.

D.5 Online Learning Under Unknown Dynamics

We now provide details on the two different online learning approaches (see § 7.2): online learning with a cold start and online learning with a warm start.

To compare with the offline approach, we assume that the online cold start approach has access to the same base information. However, it collects data iteratively through policy learning and deployment. This is different from the offline approach where data is collected with a different unknown suboptimal policy in the past and provided at once. In our assumed cold start online learning, the learner samples 500 random datapoints (without replacement) at each time step until all data is collected (with less than 500 in the last step). This results in 11 policy updates. By the last time step, the learner shares the same base information as the offline learner, differing only in the collected decisions and the resulting implications for learning from partially observed labels.

In the online warm start approach, we assume the learner has access to the historical dataset and can interact with the environment. Initially, it learns probabilities using the available historical data and subsequently collects new data periodically by applying the policy to the entire population, observing the next time steps' data. Our results show the policy learned through these iterative updates after 10 steps.

D.6 Computational ReSources and Run Time

Computational ReSources. All experiments were conducted on a MacBook Pro (Apple M1 Max chip). Since we can efficiently solve the optimization problem, these experiments are executed on standard hardware, eliminating the necessity for using GPUs.

Run Time. The optimization problems to find long-term policies in all experiments within this paper were consistently solved in under 10 seconds. Regarding the training of short-term fair policies on 5000 samples, the run times were approximately 20-23 minutes: 1245.92 seconds for `short-EOP` ($\lambda = 1$), 1244.25 seconds for `short-EOP` ($\lambda = 2$), and 1380.50 seconds for `short-MAXUTIL`.

D.7 Maximum Qualification Policy under Two-sided Dynamics

Maximum Qualifications. In the following section (§ E), we present results not only for the policy introduced in the main paper, which seeks to maximize utility subject to EOP-fairness, but also introduce another example of a fair policy. This alternative policy aims to maximize qualifications while maintaining non-negative utility.

Maximum Qualifications. Inspired by [91], assume a non-profit organization offering loans. Their goal is to optimize the overall payback ability ($\boldsymbol{Q}$) of the population to promote societal well-being. Additionally, they aim to sustain their lending program by prevent non-negative profits ($\boldsymbol{\mathcal{U}}$) in the long-term. We thus seek for:

$$\pi^{\star}_{\texttt{QUAL}} := \arg_{\pi} \max \boldsymbol{Q}(\pi) \qquad \text{subj. to} \quad \boldsymbol{\mathcal{U}}(\pi) \geq 0;\ C_{\text{conv}}(T_{\pi})$$

Two-sided Dynamics. In addition to one-sided dynamics, where only positive decisions impact the future, we also consider two-sided dynamics [91], where both positive and negative decisions lead to feature changes. We investigate two types of two-sided dynamics. Under `recourse` dynamics, individuals receiving unfavorable lending decisions take actions to improve their credit scores, facilitated through recourse [35] or social learning [27]. In `discouraged` dynamics, unfavorable lending decisions demotivate individuals, causing a decline in their credit scores. This may happen when individuals cannot access loans for necessary education, limiting their financial opportunities.

E ADDITIONAL RESULTS

In this section, we provide additional results related to the results discussed in § 7. Our analysis centers around our guiding example, employing the data distributions Sourced from FICO [68] unless otherwise specified. The structure of this section is as follows:

- In § E.1 we provide additional results for different starting distributions.
- In § E.2 we provide additional results the comparison to short-term policies.
- In § E.3 we provide additional results for varying the fairness threshold ϵ for our policy.
- In § E.4 we provide additional results for the different dynamic types (one- and two-sided).
- In § E.5 we provide additional results for varying the speed at which feature changes occur (`slow`, `medium`, `fast`).
- In § E.6 we provide additional results for offline learning under unknown dynamics.
- In § E.6 we provide additional results for online learning under unknown dynamics.
- In § E.7 we provide additional results for short-term and long-term interventions during policy deployment.

E.1 Different Starting Distributions

We provide additional results for the results shown § 7.1. Here We run simulations on 10 randomly sampled initial feature distributions $\mu_0(x \mid s)$, setting $\epsilon = 0.01, c = 0.8$. In addition to the results shown in the main paper, we here display in Figure 4 the resulting trajectories of all feature distributions.

E.2 Comparison to Static Policies

We provide additional results comparing our long-term policy to short-term policies.

Static Policy Training. The short-term policies are logistic regression models implemented using PyTorch. The forward method computes the logistic sigmoid of a linear combination of the input features, while the prediction method applies a threshold of 0.5 to the output probability to make binary predictions. The training process is carried out via gradient descent, with the train function optimizing a specified loss function. The `short-MAXUTIL` policy is trained using a binary cross-entropy loss. The fairness is enforced using a Lagrangian approach ($\lambda = 2$). The `short-EOP` policy is trained using a binary cross-entropy loss and regularization terms measuring equal opportunity unfairness with λ as hyperparameters controlling the trade-off between predictive accuracy and fairness. Training is performed for 2000 epochs with a learning rate of 0.05. We display results over 10 random initializations. The experiments in

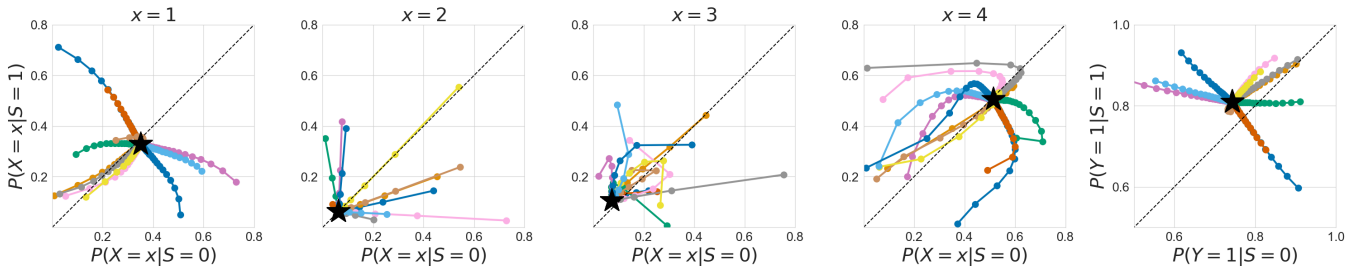


Figure 4: Convergence of feature distributions for $\pi^{\star}_{\texttt{EOP}}$ for different random starting distributions (colors) to unique stationary distributions $\mu = \star$. Trajectories over 200 time steps. $c = 0.8$, $\epsilon = 0.01$.

the main paper are shown for `short-EOP` with $\lambda = 2$. We show in the following results for different λ.

Feature and Outcome Trajectories. Figure 5 presents the trajectories of our long-term `long-EOP` ($\pi^{\star}_{\texttt{EOP}}$) and the static policies (unfair: `short-MAXUTIL`, fair: `short-EOP` ($\lambda = 2$)) over 200 time steps for a single short-term policy seed. We observe that our long-term policy converges to a stationary distribution and remains there once it has found it. In contrast, the trajectories of the short-term policies display non-stationarity, covering a wide range of distributions, as evidenced by the overlapping region. This indicates that the short-term policies exhibit a high variance and do not stabilize into a stationary distribution.

Utility, Fairness and Loan and Repayment Probabilities. Figure 6 (top left) displays $\mathcal{U}$ and `EOPUnf` over the first 100 time steps. We observe that short-term policies, which are updated at each time step, tend to exhibit greater variance compared to the long-term policy, which remains fixed at $t = 0$ - even as the underlying data distribution evolves in response to decision-making. Among the two short-term fair policies, the fairer one ($\lambda = 2$) approaches nearly zero unfairness, whereas the less fair one ($\lambda = 1$) displays a higher level of unfairness. Specifically, the more fair policy ($\lambda = 2$) reaches a low (negative) utility, while the less fair one ($\lambda = 1$) maintains a higher (though still negative) utility. The unfair short-term policy (`UTILMAX`) achieves positive utility but does so at the cost of a high level of unfairness. This highlights the trade-off between fairness and utility that short-term policies encounter. Conversely, our long-term fair policy maintains a level of unfairness close to zero while experiencing only a modest reduction in utility compared to the unfair short-term policy. This underscores our policy's capacity to attain long-term fairness while ensuring a higher level of utility, leveraging the long-term perspective to effectively shape the population distribution.

Figure 6 (top right, bottom left) presents the loan probability $\mathbb{P}(D = 1 \mid S = s)$ and payback probability $\mathbb{P}(Y = 1 \mid S = s)$ for non-privileged ($S = 0$) and privileged ($S = 1$) groups. In addition to the results presented in the main paper (Figure 2a), we observe a difference between the two short-term fair policies in our analysis in this appendix. The more equitable policy ($\lambda = 2$) achieves a low level of unfairness by granting loans with a probability of 1 to individuals across all social groups. The less equitable policy ($\lambda = 1$) provides loans to the underprivileged group with an average probability of approximately 0.85, while the privileged group receives loans at an average probability of around 0.9.

Crucially, the less equitable policy ($\lambda = 1$) exhibits a much higher variability in loan approval probabilities for the underprivileged group across different time steps compared to the privileged group. This highlights that unfairness does not solely manifest at the mean level but also in the variability across time. Both policies tend to grant loans at probabilities exceeding the actual repayment probabilities within the population. This suggests an "over-serving" phenomenon, implying that the policies on average extend loans to individuals who may not meet the necessary qualifications for borrowing.

In contrast, our policy maintains stability and converges to a low difference in loan approval probabilities between groups without significant temporal variance. Importantly, our loan approval probabilities remain below the loan repayment (as for the short-term unfair policy (`UTILMAX`)) probabilities, indicating that, on average, the policies are extending loans to individuals who are indeed eligible for them. In addition, for our policy, the gap between loan provision and repayment probabilities is similar across sensitive groups.

Effective Utility, Inequity and Unfairness. Figure 7 illustrates effective (accumulated) measures of utility, inequity, and (EOP) unfairness over time for the different policies, where results for static policies are reported over 10 random initializations. We observe that the short-term unfair policy (`short-UITLMAX` consistently accumulates the highest utility across all dynamics, while simultaneously maintaining a high level of effective unfairness and inequality. Conversely, the short-term fair policies (`short-EOP`($\lambda = 1$) and ($\lambda = 2$)) exhibit negative effective utility, but they do achieve lower levels of effective fairness and inequity.

For our long-term policy (`long-EOP`), we find that it accumulates positive utility over time. Although its utility remains below that of the short-term unfair policy, our policy exhibits very low levels of effective unfairness. Importantly, it also yields minimal accumulated inequity, even though it was not specifically optimized for this.

Analyzing the cumulative effects of policies is essential for evaluating the long-term impact of each policy choice. This analysis can, for instance, help determine whether investing in fairness pays off in the long-term and whether sacrificing short-term fairness in the initial stages ultimately benefits society in the long run.

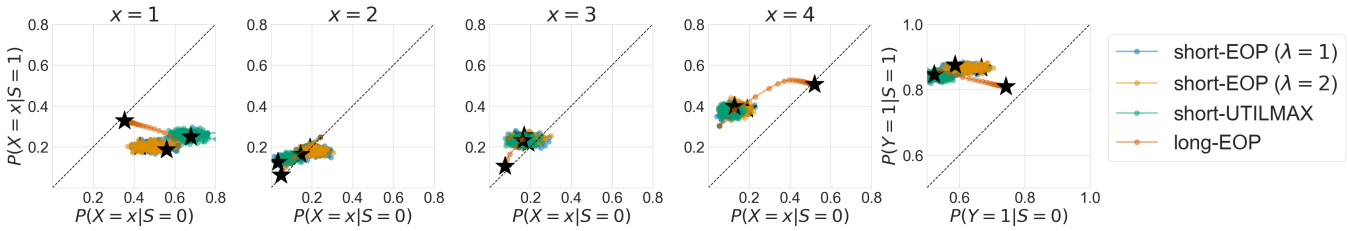


Figure 5: Convergence of feature distributions for our long-term `long-EOP` ($\pi^{\star}_{\text{EOP}}$) and the static policies (unfair: `short-MAXUTIL`, fair: `short-EOP` ($\lambda = 2$). Trajectories over 200 time steps. $c = 0.8$, $\epsilon = 0.026$. Last distribution values are marked with $\star$.

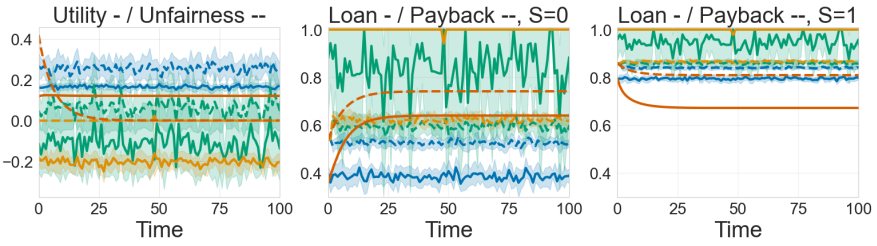


Figure 6: Results for our long-term `long-EOP` ($\pi^{\star}_{\text{EOP}}$) and the static policies (unfair: `short-MAXUTIL`, fair: `short-EOP` ($\lambda = 1$ and $\lambda = 2$). Top Left: Utility (solid, ↑) with $c = 0.8$ and EOP-Unfairness (dashed, ↓). Top right / Bottom left: Loan (solid) and payback probability (dashed) per policy and sensitive S. Colors indicate policies as in Figure 5.

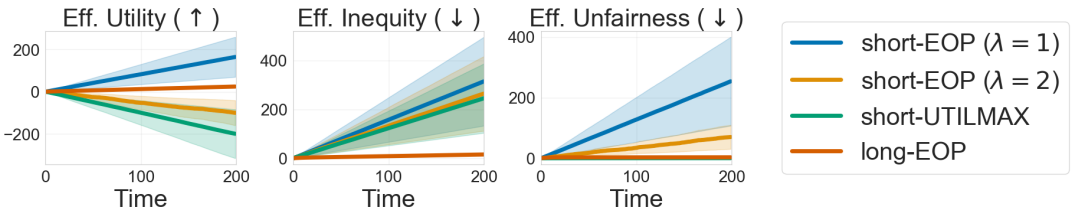


Figure 7: Results for our long-term `long-EOP` ($\pi^{\star}_{\text{EOP}}$) and the static policies (unfair: `short-MAXUTIL`, fair: `short-EOP` ($\lambda = 2$). Effective (cumulative) utility $\mathcal{U}$, inequity $\mathcal{I}$, and (EOP) unfairness `EOPUnf` for different policies.

E.3 Different Fairness Levels

We provide additional results, where we use the initial distribution $\mu_0(x \mid s)$ from FICO and solve the optimization problem provided in the main paper for four different fairness levels ϵ. This results in four policies $\pi^{\star}_{\text{EOP}}$.

Feature and Outcome Trajectories. Figure 8 presents the trajectories of $\pi^{\star}_{\text{EOP}}$ over 200 time steps for different fairness thresholds ϵ. We observe that although the convergence process, time, and final stationary distribution ($\star$) are very similar for different targeted fairness levels.

Utility and Loan and Repayment Probabilities. Figure 9 (top left) displays $\mathcal{U}$ and `EOPUnf` over the first 50 time steps (until convergence). We observe that all policies converge to a similar utility level while maintaining their respective ϵ level, confirming the effectiveness of our optimization problem. Figure 9 (top right, bottom left) presents the loan probability $\mathbb{P}(D = 1 \mid S = s)$ and payback probability $\mathbb{P}(Y = 1 \mid S = s)$ for non-privileged ($S = 0$) and privileged ($S = 1$) groups. While the probabilities across sensitive groups ultimately stabilize close together in the long term, the initial 20 steps exhibit a large difference in loan and payback probabilities. Optimizing for long-term goals may thus lead to unfairness in the short term, and it is important to carefully evaluate the potential impact of this on public trust in the policy.

E.4 Different Dynamic Types

Results in this subsection are for different dynamic types: `one-sided`, `recourse`, and `discouraged`. See D.3 for more details on these specific dynamics. We solve both optimization problems for each of the three dynamics, where solving the problem introduced in the main paper provides $\pi^{\star}_{\text{EOP}}$ and solving the optimization problem provided in § D.7 provides $\pi^{\star}_{\text{QUAL}}$.

Feature and Outcome Trajectories. Figure 10 presents the trajectories of $\pi^{\star}_{\text{EOP}}$ and $\pi^{\star}_{\text{QUAL}}$ over 200 time steps for different types of dynamics. We observe that although the initial distribution remains

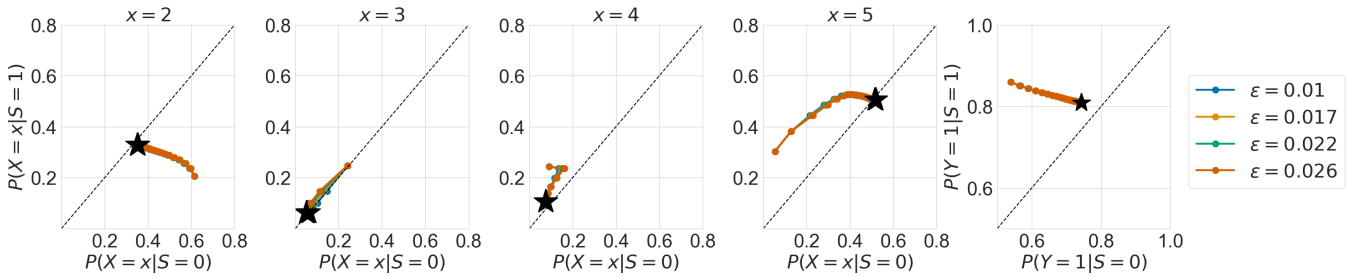


Figure 8: Convergence of feature distributions for $\pi^{\star}_{\text{EOP}}$ for different fairness thresholds ϵ to unique stationary distributions $\mu = \star$. Trajectories over 200 time steps. $c = 0.8$.

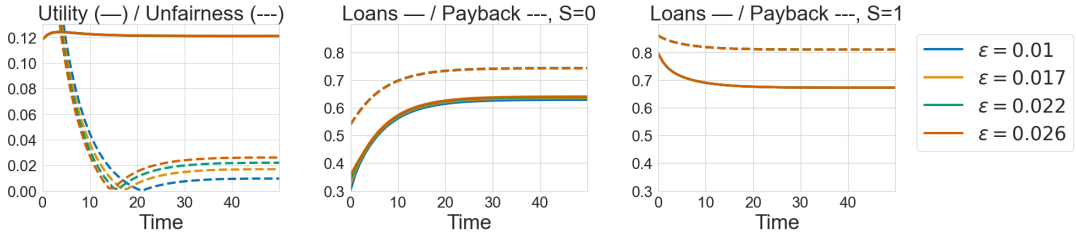


Figure 9: Results for different ϵ-EOP-fair $\pi^{\star}_{\text{EOP}}$. Top Left: Utility (solid, ↑) with $c = 0.8$ and EOP-Unfairness (dashed, ↓). Top right / Bottom left: Loan (solid) and payback probability (dashed) per policy and sensitive S.

unchanged, the convergence process, time, and final stationary distribution ($\star$) differ depending on the dynamics. Notably, the stationary distribution of $\pi^{\star}_{\text{QUAL}}$ appears to be similar for `one-sided` and `discouraged` dynamics. On the other hand, the results for all other dynamics and policies demonstrate distinct but relatively close outcomes.

Utility, Fairness and Loan and Repayment Probabilities. Figure 11 showcases the group-dependent probabilities of receiving a loan, $\mathbb{P}(D_t = 1 \mid S = s)$, and repayment, $\mathbb{P}(Y_t = 1 \mid S = s)$, for both the non-privileged ($S = 0$) and privileged ($S = 1$) groups. The probabilities are displayed for the convergence phase (first 50 time steps) for policies $\pi^{\star}_{\text{EOP}}$ and $\pi^{\star}_{\text{QUAL}}$ across dynamics types. When the payback probabilities are higher compared to the loan probabilities, it suggests an underserved community where fewer credits are granted than would be repaid. In the case of `one-sided` dynamics, we find that for $\pi^{\star}_{\text{EOP}}$, the loan and repayment probabilities are relatively close to each other at each time step. However, for $\pi^{\star}_{\text{QUAL}}$, the gap between repayment and loan probabilities widens as time progresses. At convergence, both sensitive groups exhibit a repayment rate of approximately 0.8, while the loan-granting probability is around 0.4. This suggests that, in the `one-sided` dynamics, for $\pi^{\star}_{\text{QUAL}}$ the repayment rate is higher compared to the loan granting rate, indicating that a significant number of individuals who would repay their loan are not being granted one. In the case of `one-sided` dynamics, similar to the `discouraged` dynamics, we observe different short-term and long-term effects. Specifically, for $\pi^{\star}_{\text{EOP}}$, the probability of receiving a loan initially differs between the sensitive groups within the first 20 time steps. However, as time progresses, these probabilities tend to become closer to each other. This suggests a potential reduction in the disparity of loan access between the sensitive groups over time under the influence of the $\pi^{\star}_{\text{EOP}}$ policy. In the case of `recourse` dynamics, we observe that the loan granting and repayment probabilities tend to stabilize closely together in the long term across sensitive groups and under both policies—except for $\pi^{\star}_{\text{QUAL}}$ when $S = 1$. In this particular case, the $\pi^{\star}_{\text{QUAL}}$ policy sets $\pi(D = 1 \mid X = x, S = 1) = 0$ for all values of x. This scenario serves as an example where optimizing for long-term distributional goals without enforcing predictive fairness constraints can lead to individuals with a high probability of repayment being consistently denied loans.

E.5 Different Dynamic Speeds

We begin by assuming one-sided dynamics and then introduce variation in the speed of transitioning between different credit classes. This variation encompasses three levels: `slow`, `medium`, and `fast`, each representing the rate at which borrowers' credit scores evolve in response to decisions. Additional information about these specific dynamics can be found in § D.3. For each of the three dynamics, we solve both introduced optimization problems to obtain $\pi^{\star}_{\text{EOP}}$ as presented in the main paper and $\pi^{\star}_{\text{QUAL}}$ introduced in § D.7.

Feature and Outcome Trajectories. Figure 12 depicts the trajectories over 200 time steps for $\pi^{\star}_{\text{EOP}}$ and $\pi^{\star}_{\text{QUAL}}$ under different speeds of one-sided dynamics. While the initial distribution remains the same for all runs, the convergence process, time, and final stationary distribution ($\star$) vary depending on the dynamics speed. Regarding the group-dependent distribution of Y, we observe that $\pi^{\star}_{\text{QUAL}}$ achieves a higher distribution (which in addition is closer to the equal outcome

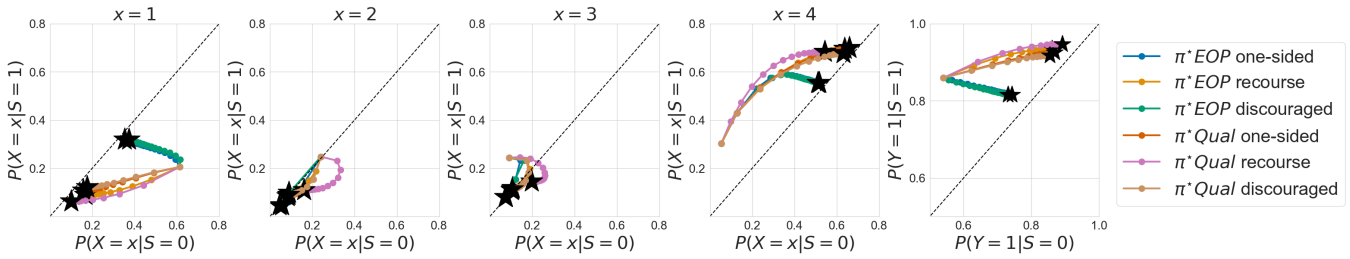


Figure 10: Convergence of $\pi^{\star}_{\text{EOP}}$ and $\pi^{\star}_{\text{QUAL}}$ for different type of dynamics towards different unique stationary distributions $\mu = \star$. Trajectories over 200 time steps. Top four plots: feature distribution μ_t. Bottom left: distribution of the outcome of interest. Equal feature/outcome distribution dashed. Initial distribution μ_0 =FICO, $c = 0.8$, $\epsilon = 0.01$.

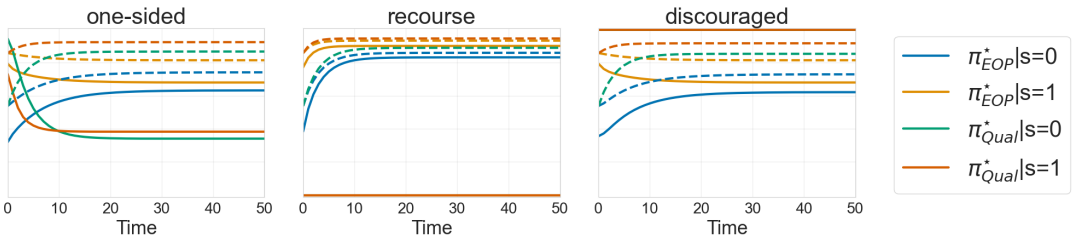


Figure 11: Loan probability $\mathbb{P}(D = 1 \mid S = s)$ (solid) and repayment probability $\mathbb{P}(Y = 1 \mid S = s)$ (dashed) for different type of dynamics (one-sided, recourse, discouraged) and policies $\pi^{\star}_{\text{EOP}}, \pi^{\star}_{\text{QUAL}}$ per sensitive attribute $s \in \{0, 1\}$. Initial distribution μ_0 = FICO, $c = 0.8$, $\epsilon = 0.01$.

distribution) compared to $\pi^{\star}_{\text{EOP}}$. This can be attributed to the fact that $\pi^{\star}_{\text{QUAL}}$ explicitly optimizes for maximizing the total distribution of Y. Additionally, we notice that for both policies slower dynamics result in lower stationary distributions of Y compared to faster dynamics.

Utility, Fairness and Loan and Repayment Probabilities. Figure 13 depicts the group-dependent probabilities of receiving a loan, $\mathbb{P}(D = 1 \mid S = s)$, and repayment, $\mathbb{P}(Y = 1 \mid S = s)$, for both non-privileged ($S = 0$) and privileged ($S = 1$) groups. The probabilities are shown for the convergence phase (initial 50 time steps) of policies $\pi^{\star}_{\text{EOP}}$ and $\pi^{\star}_{\text{QUAL}}$ across different speeds of one-sided dynamics. Higher payback probabilities compared to loan probabilities can indicate an underserved community where fewer credits are granted than would be repaid. Across all dynamics, we observe small differences in the repayment distributions for each policy. The repayment probabilities are consistently higher for the non-protected group compared to the protected group. Moreover, in general, $\pi^{\star}_{\text{QUAL}}$ yields higher repayment rates than $\pi^{\star}_{\text{EOP}}$. However, the loan probabilities—which indicate a group's access to credit—exhibit differences across dynamics and policies. As expected, the utility-maximizing $\pi^{\star}_{\text{EOP}}$ generally provides higher loan rates compared to $\pi^{\star}_{\text{QUAL}}$. While the loan rates remain similar across dynamics for $\pi^{\star}_{\text{EOP}}$, they vary for $\pi^{\star}_{\text{QUAL}}$. Under `slow` dynamics, $\pi^{\star}_{\text{QUAL}}$ yields low loan probabilities for the protected group, which then increases for `medium` and `fast` dynamics. Furthermore for $\pi^{\star}_{\text{QUAL}}$, the discrepancy between acceptance rates for sensitive groups is greatest at `slow` dynamics, and decreases significantly at `medium` dynamics - at the expense of the non-protected group. Finally, for `fast` dynamics, the acceptance rates for sensitive groups are approximately equal.

These observations emphasize the importance of conducting further investigations into the formulation of long-term goals, taking into account their dependence on dynamics and the short-term consequences. This includes not only considering the type of dynamics (one-sided or two-sided), but also the speed at which individuals' feature changes in response to a decision.

Effective Utility, Inequity and Unfairness. Figure 14 illustrates effective (accumulated) measures of utility, inequity, and (EOP) unfairness over time. For all dynamics, the policies align with their respective targets. $\pi^{\star}_{\text{EOP}}$ accumulates the highest utility across all dynamics while maintaining a low effective unfairness after an initial convergence period. On the other hand, $\pi^{\star}_{\text{QUAL}}$ exhibits a small negative effective utility due to the imposed zero-utility constraint, but achieves lower effective inequity by maximizing the total distribution of the outcome of interest. We observe that the speed of dynamics does not significantly affect effective utility for both policies and effective unfairness for the $\pi^{\star}_{\text{EOP}}$ policy. However the speed of dynamics does have an impact effective inequity, although its effect varies for each policy. Among the $\pi^{\star}_{\text{EOP}}$ policies, we find that the `medium` dynamics result in the lowest effective inequity, whereas among the $\pi^{\star}_{\text{QUAL}}$ policies, the `fast` dynamics exhibit the lowest effective inequity. While the effective utility is minimally affected by the speed of dynamics in the case of $\pi^{\star}_{\text{EOP}}$, we observe different results for effective inequity. Among the $\pi^{\star}_{\text{EOP}}$ policies, the `medium` dynamics result in the lowest effective inequity. Conversely, among

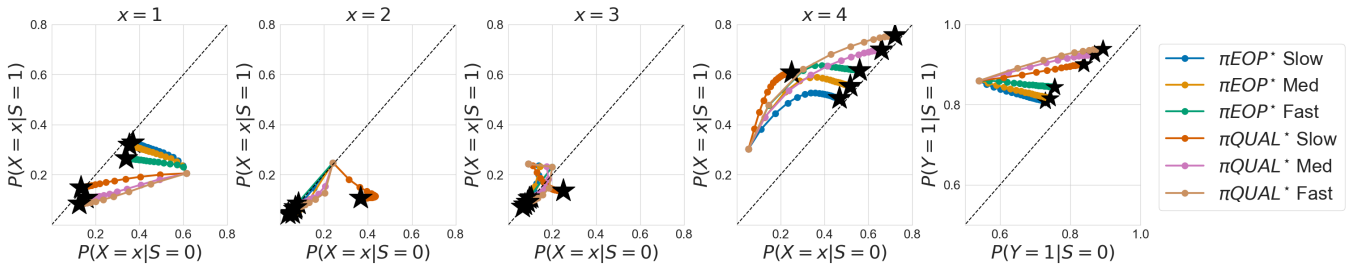


Figure 12: Convergence of $\pi^{\star}_{\text{EOP}}$ and $\pi^{\star}_{\text{QUAL}}$ for different speeds of dynamics towards different unique stationary distributions $\mu = \star$. Trajectories over 200 time steps. Left four plots: feature distribution μ_t. Right: distribution of the outcome of interest. Equal feature/outcome distribution dashed. Initial distribution μ_0 =FICO, $c = 0.8$, $\epsilon = 0.01$.

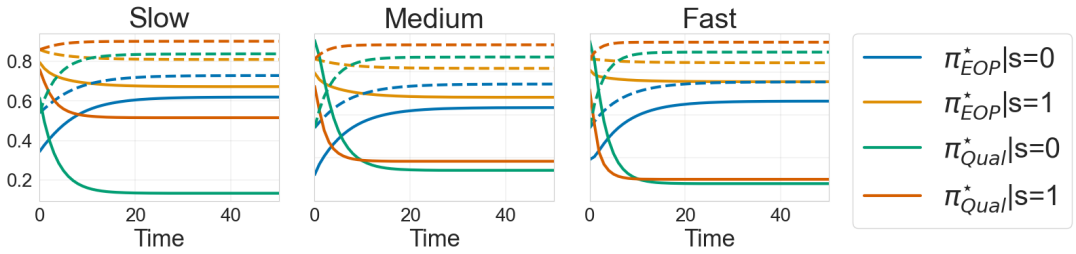


Figure 13: Loan probability $\mathbb{P}(D = 1 \mid S = s)$ (solid) and repayment probability $\mathbb{P}(Y = 1 \mid S = s)$ (dashed) for different speed of one-sided dynamics (`slow`, `medium`, `fast`) and policies $\pi^{\star}_{\text{EOP}}, \pi^{\star}_{\text{QUAL}}$ per sensitive attribute $s \in \{0, 1\}$. Initial distribution μ_0 = FICO, $c = 0.8$, $\epsilon = 0.01$.

the $\pi^{\star}_{\text{QUAL}}$ policies, the `fast` dynamics exhibit the lowest effective inequity. These observations highlight that the final outcomes of decision policies are not only influenced by the type of dynamics (one-sided and two-sided), but also by the speed of dynamics. It is thus crucial to also consider the rate at which individuals are able to change features within one time step. This consideration can for example be important in the context of recourse, where not all individuals may have the ability to implement the minimum recommended actions, potentially due to individual limitations. Consequently, only a fraction of individuals would be able to move up in their credit class in response to a negative decision.

E.6 Offline Learning Under Unknown Dynamics

We conduct additional experiments to investigate the impact of estimation errors in the underlying distributions on the quality of results on FICO (§ E.6.1) and COMPAS (§ E.6.2). Throughout we assume partially observed outcomes of interests (labels) Y.

E.6.1 FICO Lending Example. We investigate the sensitivity of our derived policy to the estimation of Y for different decision policies compared to access to the true distribution of Y. In a pracical loan example, label Y might be partially observed (i.e., observed only for individuals who received a positive loan decision). In this case, the estimate of Y may no longer be as accurate for one sensitive group as for another. Thereby, different policies reveal different amounts of labels for different subgroups. We first generate a temporal dataset comprising two time steps. These samples were drawn from the FICO base distribution, and we assumed the dynamics of One-sided General (as described in § D.3). The dataset is comprised of 50,000 samples aligning with the dataset scales employed in the fairness literature, such as the Adult dataset [40]. We deploy two different policies that influence the data observed at $t = 1$, `random` and `biased`, with the following formulations:

- `random` is defined by $\mathbb{P}(D = 1 \mid X, S) = 0.5$ for all X, S;
- `bias` is defined for all S by $\mathbb{P}(D = 1 \mid X, S) = 0.1$ if $X <= 2$ and for $S = 0$ as $\mathbb{P}(D = 1 \mid X, S) = 0.3$ if $X > 2$ and for $S = 1$ as $\mathbb{P}(D = 1 \mid X, S) = 0.9$.

The true distribution of features and label at $t = 0$ are shown in Figure 15a. The distributions of decisions and observed labels under the different policies are shown in Figures 15b - 15c.

We then estimate both $\ell(y \mid x, s)$ and $g(k \mid x, d, y, s)$ from the observed samples, with the latter being dependent on the former. Subsequently, we solve the optimization problem ($c = 0.9$, $\epsilon = 0.00005$) using these estimated distributions yielding two different policies (one per estimation). Consequently, we simulate the performance of the discovered policies under the true distributions and μ_0 =FICO. In the evaluation, we compare the results to the policy obtained under the true probability estimate $\ell(y \mid x, s)$ as supplied by FICO (`true`).

Feature and Outcome Trajectories. Figure 16 displays the trajectories of $\pi^{\star}_{\text{EOP}}$ for 200 time steps for the optimal policies obtained under both the true and estimated distributions and dynamics. Notably, the initial distribution remains the same, and the policies slightly vary in their convergence process to the stationary distribution ($\star$), while staying close to each other. It is important to emphasize that all policies successfully achieve a stationary distribution. This is due

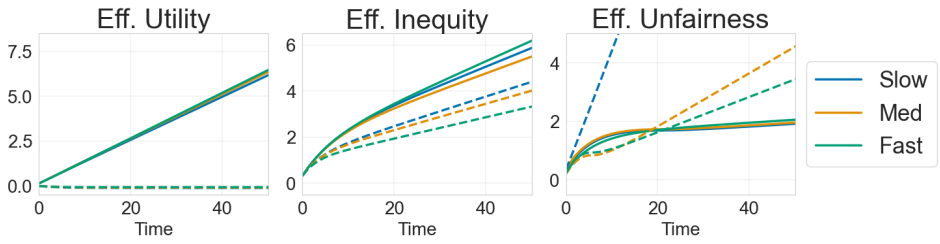
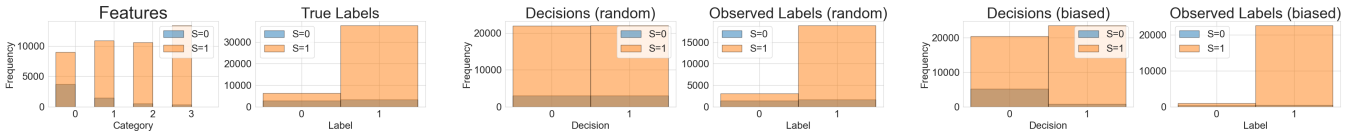


Figure 14: Effective (cumulative) utility $\mathcal{U}$, inequity $\mathcal{I}$, and (EOP) unfairness `EOPUnf` for different policies ($\pi^{\star}_{\text{EOP}}$ solid, $\pi^{\star}_{\text{QUAL}}$ dashed).



(a) True distributions of features and labels.

(b) Distribution of decisions and observed labels for `random`.

(c) Distribution of decisions and observed labels for `bias`.

Figure 15: Data distributions for different temporal datasets based on FICO used to estimate label distributions and dynamics.

to the fact that even though we employ estimated distributions as inputs for the optimization problem, we are still solving the optimization problem for a policy that induces a stationary distribution that satisfies the fairness criteria. We showcase this in the next results.

Utility, Fairness and Loan and Repayment Probabilities. Figure 17 (left) displays $\mathcal{U}$ and `EOPUnf` over the first 50 time steps (until convergence). We observe that the policies exhibit a different level of unfairness, while still achieving low unfairness. The policy derived from the true probabilities and dynamics achieves lowest unfairness, the policy derived from probabilities and dynamics collected under a random policy has slightly higher unfairness, and the policy derived from probabilities and dynamics collected under a biased policy has the highest unfairness. In terms of utility, where we aim for maximization without imposing a strict constraint, we observe that all policies exhibit a similar utility level. Figure 9 (middle, right) displays the loan probability $\mathbb{P}(D = 1 \mid S = s)$ and payback probability $\mathbb{P}(Y = 1 \mid S = s)$ for non-privileged ($S = 0$) and privileged ($S = 1$) groups. While there is no difference in loan and payback probabilities for the privileged group ($S = 1$) between the policies, we observe a small difference for the unprivileged group ($S = 0$). The policy derived from true probabilities and dynamics provides fewer loans to the unprivileged group compared to the policy derived from probabilities and dynamics collected under the random policy. Interestingly, the policy derived from probabilities and dynamics collected under a biased policy grants the most loans to the unprivileged group. Note, that our unfairness metric in the left plot is equal opportunity [25], not demographic parity [19]. Consequently, this observation may be explained by the policy obtained from biased estimation providing loans to a higher number of individuals from the unprivileged group who may not be able to repay them. Thus, while we do achieve a stationary distribution using estimated probabilities, it is important to note that convergence to the intended fair state is not guaranteed when estimation errors are present. However, if the estimations closely approximate the true distribution, the resulting stationary distribution achieves similar utility and fairness properties as the stationary distribution that would have been achieved had the policy found under the true probabilities.

E.6.2 COMPAS Recidivism Example. We first generate a temporal dataset comprising two time steps. For this, we use the samples provided by the COMPAS dataset, and assume the dynamics of One-sided General (as described in § D.3). The dataset is comprised of 5278 samples. We deploy two different policies that influence the data observed, `random`, and `bias`, with the following formulations:

- `random` is defined by $\mathbb{P}(D = 1 \mid X = x, S = s) = 0.5$ for all x, s;
- `bias` decides with the following probabilities $\mathbb{P}(D = 1 \mid X = x, S = s) :=$

(X1, X2)	(0,0)	(0,1)	(0,2)	(0,3)	(1,0)	(1,1)	(1,2)	(1,3)	(2,0)	(2,1)	(2,2)	(2,3)
S = 0	0.3	0.2	0.1	0.1	0.3	0.3	0.1	0.1	0.1	0.1	0.1	0.1
S = 1	0.5	0.4	0.3	0.2	0.4	0.4	0.2	0.2	0.2	0.2	0.2	0.2

The true distribution of features and label at $t = 0$ are shown in Figure 18a. The distributions of decisions and observed labels under the different policies are shown in Figures 18b and 18c.

We then estimate $\gamma(s)$, $\ell(y \mid x, s)$ and one-sided $g(k \mid x, d, y, s)$ (for $d = 1$) from the observed samples. Subsequently, we solve the optimization problem (5) with $c = 0.65$, $\epsilon = 0.001$, using these estimated distributions. This gives us two different policies (one per estimation), `random` and `bias`. For comparison, we also show the performance of a policy learned under optimal conditions, i.e., from known dynamics `true`. Consequently, we simulate the performance of the discovered policies under the true distributions and μ_0 estimated from COMPAS.

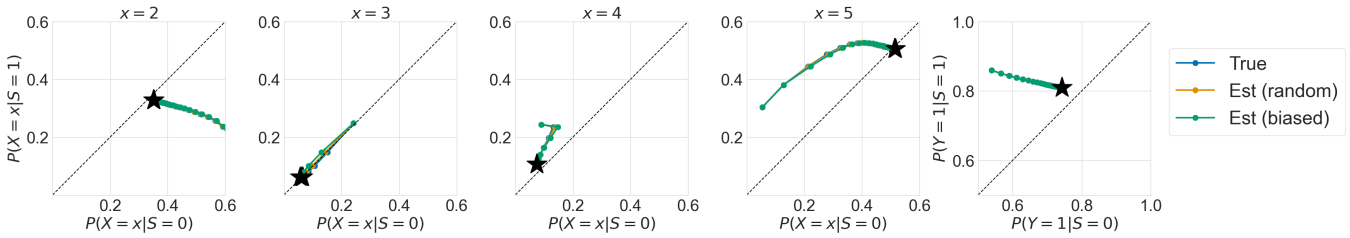


Figure 16: Convergence of $\pi^{\star}_{\texttt{EOP}}$ under true and estimations of $\ell(y \mid x, s)$ and $g(k \mid x, d, y, s)$ and under different type of initial policies (`random`, `threshold`, `bias`). 200 time steps, last time step marked $\star$. Top four plots: feature distribution μ_t. Bottom left: distribution of the outcome of interest. Equal feature/outcome distribution dashed.

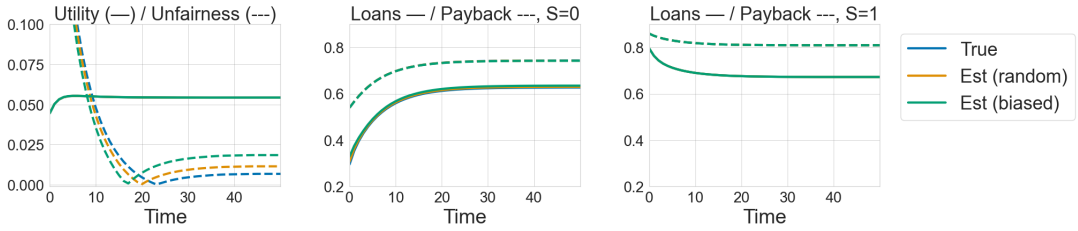
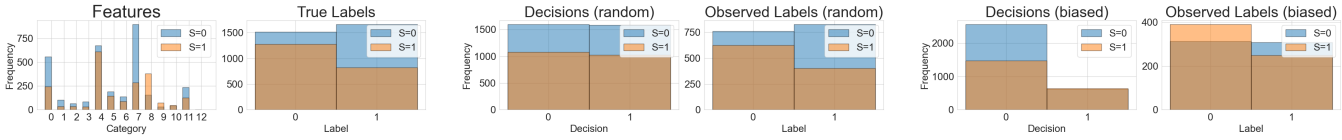


Figure 17: Results for our $\pi^{\star}_{\texttt{EOP}}$ under true and estimations of $\ell(y \mid x, s)$ under different type of initial policies (`random`, `threshold`, `bias`). Top Left: Utility (solid, ↑) and EOP-Unfairness (dashed, ↓) over first 50 time steps. Remaining: Loan (solid) and payback probability (dashed) per policy and sensitive S.



(a) True distributions of features and labels.

(b) Distribution of decisions and observed labels for `random`.

(c) Distribution of decisions and observed labels for `bias`.

Figure 18: Data distributions for different temporal datasets based on COMPAS.

Feature and Outcome Trajectories. Figure 19 displays the trajectories of $\pi^{\star}_{\texttt{EOP}}$ for 200 time steps for the optimal policies obtained under both the true and estimated distributions and dynamics. Note, the initial distribution remains the same. As above for the results from the FICO dataset, the policies slightly vary in their convergence process to a stationary distribution ($\star$), while staying close to each other. As above, we emphasize that all policies successfully achieve a stationary distribution. This is due to the fact that even though we employ estimated distributions as inputs for the optimization problem, we are still solving the optimization problem for a policy that induces a stationary distribution that satisfies the fairness criteria. We showcase this in the next results.

Long-term Utility and Fairness. Figure 20 illustrates the equilibrium values of $\mathcal{U}$ (left side of each subplot) and $\epsilon-\texttt{EOPUnf}$ (right side of each subplot). For $\epsilon - \texttt{EOPUnf}$, the dashed line represents $\epsilon = 0$, while the gray shaded area depicts the range where $\epsilon-\texttt{EOPUnf}$ is satisfied. Each subplot reports results for a different policy $\pi^{\star}_{\texttt{EOP}}$ learned under different unfairness relaxations ϵ and costs c. We present results for learning policies from unknown dynamics and estimate the dynamics using historical data collected a `random` policy and with a `bias` (which grants less bail for the underprivileged). These results are then compared to policies learned from known dynamics, `true`.

We observe that policies learned from unknown dynamics consistently yield a lower level of utility compared to `true` across the values of ϵ and c considered here. This difference can be attributed to the fact that `true` has learned to maximize utility while ensuring ϵ-`EOPUnf`. Policies under unknown dynamics, however, experience sub-optimal utility due to misestimation of the system's dynamics.

The impact of this misestimation on fairness varies. When $c = 0.6$, policies learned under unknown dynamics result in more fair equilibria (closer to the dashed line) compared to the equilibrium of the `true` policy, which is located on the outer edge of the ϵ-boundary. Conversely, for $c = 0.75$, the resulting equilibria for policies under unknown dynamics are slightly more unfair and located outside of the ϵ boundary. Additionally, these equilibria exhibit a

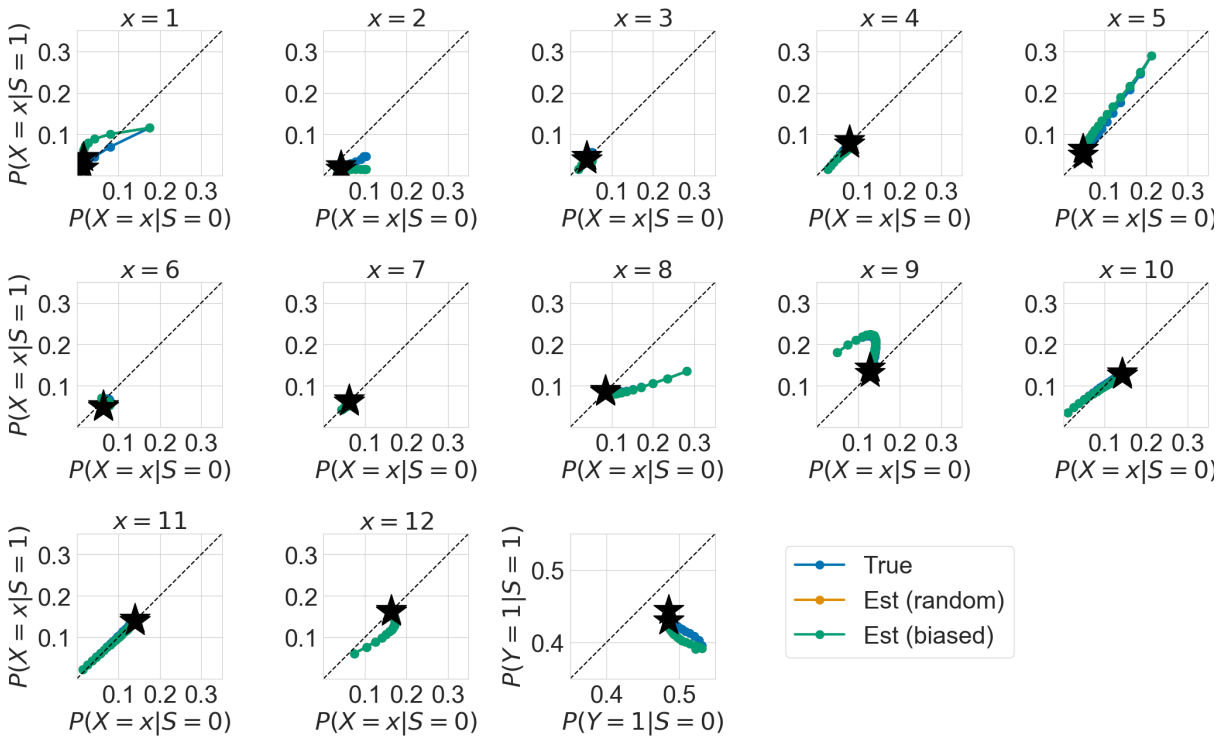


Figure 19: Convergence of $\pi^{\star}_{\text{EOP}}$ under offline estimations of $\ell(y \mid x, s)$ and $g(k \mid x, d, y, s)$ from historical datasets collected under different type of policies (`random`, `bias`) compared to known dynamics (`true`). 200 **time steps, stationary distribution $\star$. Plots show the feature distribution μ_t for $x = 1, \ldots, 12$, and distribution of the outcome of interest. Equal feature/outcome distribution dashed. Results are for $\pi^{\star}_{\text{EOP}}$ reported in the main paper with $c = 0.65$, $\epsilon = 0.01$**

lower total acceptance rate among qualified individuals compared to `true`. This suggests that policies learned from unknown dynamics may suffer from reduced utility at equilibrium. Their outcomes in terms of fairness and acceptance rates at equilibrium appear highly dependent on the specific scenario and parameters considered.

Between `bias` and `random`, we observe a small difference in utility, with the biased policy yielding slightly less utility than the random one. Fairness differences are marginal. Acknowledging that this observation might not extend to other setups, in our specific scenario, it implies that the policy responsible for dataset collection has a minimal impact on the estimation error. It rather seems that the crucial factor for dataset quality is the amount of data and the coverage of the feature space. We proceed to explore an online learning setup under unknown dynamics, where the quantity of data and coverage play a crucial role in the early steps.

E.7 Short- and Long-term Interventions During Policy Deployment

We briefly describe additional experiments concerning short- and long-term interventions. We run simulations on 3 randomly sampled feature distributions $\mu_0(x \mid s)$, setting $\epsilon = 0.01, c = 0.8$.

In Figure 21, the top two plots (21a) and (21b) reveal that a change in the feature distribution (four plots from the left) leads to a shift in the distribution of decisions, influencing the group fairness criteria (EOP-fairness, right plot) considered here.

In (21c) and (21d), we observe cumulative measures for different approaches. While short-term interventions do not seem to significantly alter the accumulated utility, their impact on fairness varies depending on the scenario. Unfairness sometimes decreases (for blue and green populations) and other times increases (for the orange population).

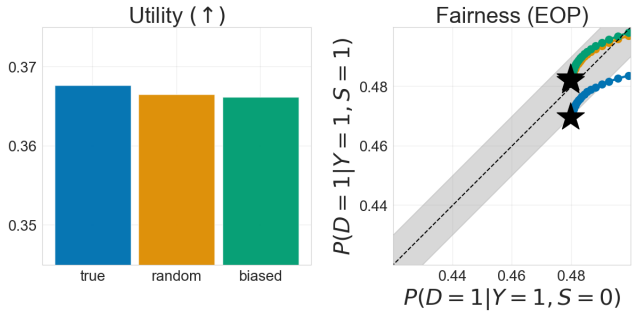
The long-term intervention exemplified in the paper suggested implementing a recourse policy. Interestingly, the deployment of such a policy increases cumulative utility and decreases cumulative EOP-unfairness for all setups.

F EXAMPLE SCENARIOS

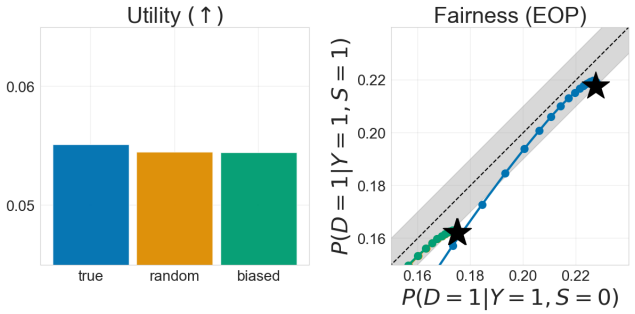
In this section, we discuss assumptions underlying the generative model presented in the main paper (which assumes a causal relationship $X \to Y$) in F.1. Then, we illustrate how our model can be applied to a different generative model (which assumes causal direction $Y \to X$) in F.2.

F.1 Assumptions of the Guiding Data Generative Model

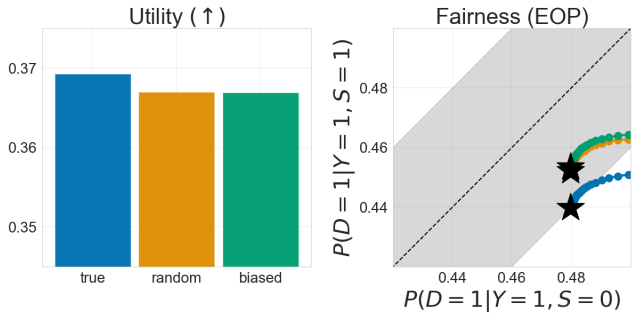
In this section, we discuss the assumptions taken in the data generative model introduced in § 2.



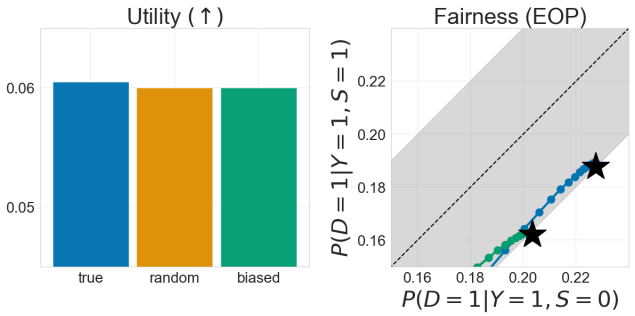
(a) $c = 0.6$, $\epsilon = 0.01$



(b) $c = 0.75$, $\epsilon = 0.01$



(c) $c = 0.6$, $\epsilon = 0.01$



(d) $c = 0.75$, $\epsilon = 0.01$

Figure 20: Long-term utility and fairness leaning under unknown dynamics from historical data collected under different policies (`random`, `bias`) compared to known dynamics (`true`). Results on COMPAS for different cost c and fairness relaxations ϵ at $t = 200$.

Assumptions F.1. *S is a root node and X_t, Y_t and D_t (potentially) depend on S.*

It is commonly assumed in the causality and fairness literature that sensitive features are root nodes in the graphical representation of the data generative model [11, 37, 43], although there is some debate on this topic [30, 52]. The assumption that the sensitive attribute S influences X_t is based on the observation that in practical scenarios, nearly every (human) characteristic is causally influenced by the sensitive attribute [11, 43]. In some cases, it is also assumed that S influences Y_t [11], while in other cases, this assumption is not made [46]. The extent to which the decision D_t is directly influenced by the sensitive attribute S depends on the decision policy being employed. Policies that strive for (statistical) fairness often require explicit consideration of the protected attribute in their decision-making process [13, 19, 25].

Assumptions F.2. *The outcome of interest Y_t depends on features X_t.*

The assumption that changes in X_t lead to changes in Y_t is prevalent in scenarios involving lending [15, 17, 31, 46]. This assumption is also implicit in problems where individuals seek recourse, e.g., via minimal consequential recommendations [34] or social learning [27].

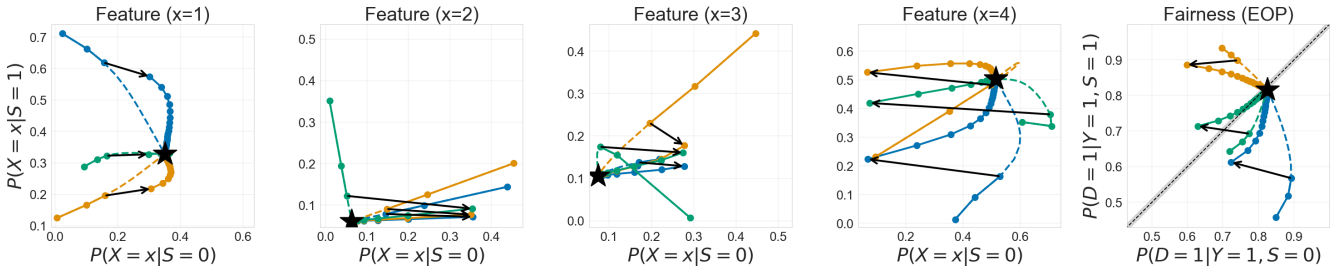
Assumptions F.3. *Decision D_t depends on features X_t.*

In algorithmic decision-making, the primary objective of a policy is typically to predict the unobserved label or outcome of interest, denoted as Y, based on the observable features, denoted as X [72]. We make the assumption that an individual's observed features at a particular point in time are sufficient to make a decision and conditioned on these features, the decision is independent of past features, labels, and decisions. This assumption aligns with prior work in the field [15, 34, 91].

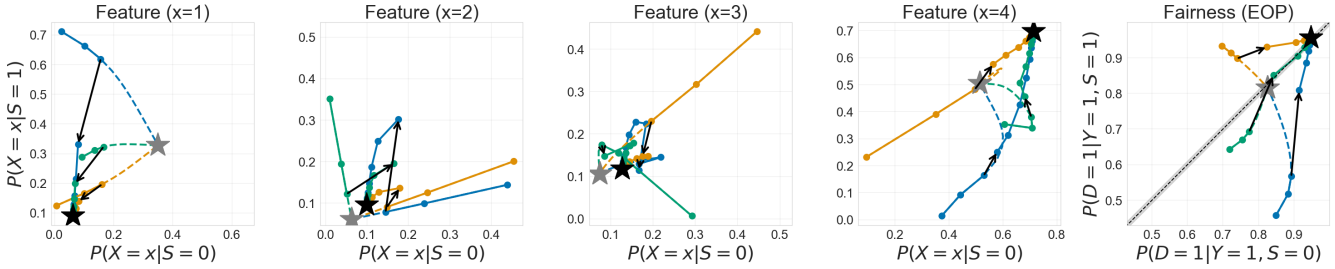
Assumptions F.4. *An individual's sensitive attribute S is immutable over time.*

For simplicity, we assume that individuals do not change their sensitive attribute. This assumption aligns with previous works that consider a closed population [15, 17, 46, 76]. A closed population refers to a group of individuals that remains constant throughout the study or analysis. It implies that there are no additions or removals from the population of interest. Other work considers that individuals join and leave the population over time, leading to a changing distribution of the sensitive attribute [26]. The assumption that individuals do not change their sensitive attribute is controversial because, on the one hand, social categories are often ontologically unstable [4, 30], and as such their boundaries are not clearly defined and dynamic. On the other hand, it ignores that individuals may be assigned identities at birth which they have the agency to correct at a given time. For example, an individual assigned one religion at birth may have a different religion at a later stage in life.

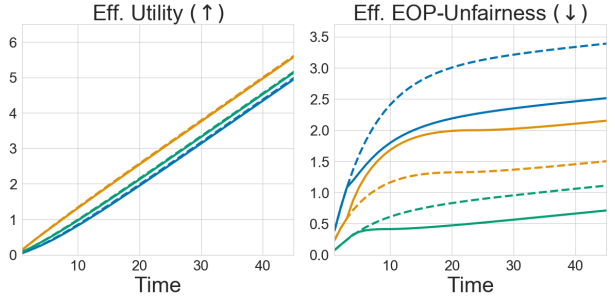
Assumptions F.5. *An individual's next step's features X_{t+1} depend on its current step's feature X_t, decision D_t, outcome of interest Y_t, and sensitive S.*



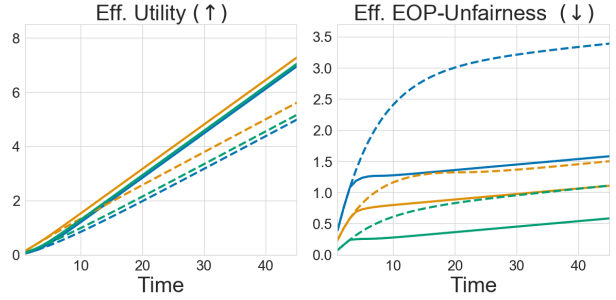
(a) Short-term interventions on feature distribution μ_t at $t = 3$.



(b) Long-term interventions on dynamics $g(k \mid x, d, y, s)$ (fair recourse policy).



(c) Cumulative utility/fairness. Short-term interventions.



(d) Cumulative utility/fairness. Long-term interventions.

Figure 21: Distribution shifts due to short-term (a) and long-term (b) interventions. Left four plots: Distributions of features X. Right plots: EOP-fairness dashed, ϵ-EOP-fair gray ($\epsilon = 0.01$). 200 time steps, $c = 0.8$. Colors: random initial feature distributions.

This assumption, as discussed in previous literature, can be attributed to either bureaucratic policies [46] or changes in individual behavior, in response to recommendations [35] or social learning [27]. In the lending context, it is commonly assumed that the higher the credit score the better. Then the assumption is: individuals approved for a loan ($D = 1$) experience a positive score change upon successful repayment ($Y = 1$) and a negative score change in case of default ($Y = 0$), while individuals rejected for a loan ($D = 0$) are assumed to have no score change [15, 17, 46]. In scenarios where individuals who are not granted a loan ($D = 0$) seek recourse, it would be assumed that a negative decision leads to an increase in credit score, to elicit a positive decision change in subsequent time steps [27, 35].

For the transition probabilities to be time-homogeneous, we take the following assumptions:

Assumptions F.6. *Dynamics $g(k \mid x, d, y, s)$ remain fixed over time.*

This is a common assumption in the literature [15, 17, 31, 91]. Although real-world data often exhibits temporal changes, we make the simplifying assumption of static dynamics. We can treat the dynamics as constant for specific durations. This is reasonable in situations where changes are based on policies involving bureaucratic adjustments [46] or algorithmic recourse recommendations [34], and where it is desirable for these policies to remain unchanged or not be retrained at every time step [62]. In practical applications, MDPs with time-varying transition probabilities present challenges, and the literature addresses this through online learning algorithms (e.g., [45, 88]).

Assumptions F.7. *Label distribution $\ell(y \mid x, s)$ remains fixed over time.*

This assumption is widely recognized in the literature [15, 17, 27, 31, 35, 91]. However, in real-world scenarios, the relationship between input data X_t and the target output Y_t may change over time, resulting in changes in the conditional distribution $\ell(y \mid x, s)$. This phenomenon is commonly referred to as *concept drift* [23, 47]. In the lending scenario, concept drift may arise from changes in individuals' repayment behavior or alterations in the process of

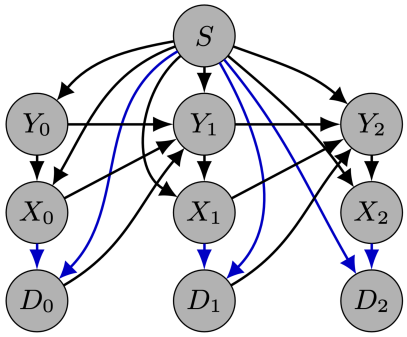


Figure 22: Data generative model: qualifications over time. Time steps (subscript) $t = \{0, 1, 2\}$**.**

generating credit scores based on underlying features like income, assets, etc.

F.2 Additional Example: Qualifications over Time

In this section, we provide an additional example, which could also be covered by our framework. The example was provided by [91] with their data generative model displayed in Figure 22. The primary distinction from the example presented in Section 2 lies in the assumption that $Y_t \to X_t$. [91] employ their model to replicate lending and recidivism scenarios over time in their experiments, using FICO and COMPAS data, respectively. However, most prior work has modeled the (FICO) lending examples as $X_t \to Y_t$ [15, 17, 46]. The same holds for recidivism (COMPAS) [70]. We, therefore, frame the example as a repeated admission example where Y_t denotes a (presumably hidden) qualification state at time t, following [43, 67].

Data Generative Model. Let an individual with protected attribute S (e.g., gender) at time t be described by a qualification Y_t and a non-sensitive feature X_t (e.g., grade or recommendations levels). We assume the sensitive attribute to remain fixed over time, and drop the attributes time subscript. For simplicity, we assume binary sensitive attribute and qualification, i.e., $S, Y_t \in \{0, 1\}$ and a one-dimensional discrete non-sensitive feature $X_t \in \mathbb{Z}$. Let the population's sensitive attributes be distributed as $\gamma(s) := \mathbb{P}(S = s)$ and assume them to remain constant over time. We assume Y_t to depend on S, such that the group-conditional qualification distribution at time t is $\mu_t(y|s) := \mathbb{P}(Y_t = y \mid S = s)$. For example, different demographic groups may have different qualification distributions due to structural discrimination in society. We assume that the non-sensitive features X_t are influenced by the qualification Y_t and, possibly (e.g., due to structural discrimination), the sensitive attribute S. This leads to the feature distribution $f(x|y, s) := \mathbb{P}(X_t = x \mid Y_t = y, S = s)$, We assume that there exists a policy that takes at each time step t binary decisions D_t (e.g., whether to admit) based on X_t and (potentially) S and decides with probability $\pi(d \mid x, s) := \mathbb{P}(D_t = d \mid X_t = x, S = s)$.

Consider now dynamics in which the decision D_t made at one time step t, directly impacts an individual's qualifications at the next step, Y_{t+1}. Assume the transition from the current qualification state Y_t to the next state Y_{t+1} is determined by the current qualification state Y_t, decision D_t and (potentially) sensitive attribute S. For example, upon receiving a positive admission decision, an individual may be very motivated and increase their qualifications. However, due to structural discrimination, the extent of the qualification change may be influenced by the individual's sensitive attribute. We denote the probability of an individual with $S = s$ changing from qualification $Y_t = y$ to $Y_{t+1} = k$ in the next step in response to decision $D_t = d$ as dynamics $g(k \mid y, d, s) := \mathbb{P}(Y_{t+1} = k \mid Y_t = y, D_t = d, S = s)$. Crucially, the next step qualification state (conditioned on the sensitive attribute) depends only on the present state qualification and decision, and not on any past states.

Dynamical System. We can now describe the evolution of the group-conditional qualification distribution $\mu_t(y|s)$ over time t. The probability of a qualification change from y to k in the next step given s is obtained by marginalizing out decision D_t, resulting in

$$\begin{aligned} &\mathbb{P}(Y_{t+1} = k \mid Y_t = y, S = s) \\ &= \sum_{xd} g(k \mid y, d, s)\pi(d \mid x, s) f(x|y, s). \end{aligned} \tag{35}$$

These transition probabilities together with the initial distribution over states $\mu_0(y|s)$ define the behavior of the dynamical system. In our model, we assume that the dynamics $g(k \mid y, d, s)$ are time-independent, meaning that the qualification changes in response to the decision, the previous qualification and the sensitive attribute remain constant over time. We also assume that the distribution of the non-sensitive features conditioned on an individual's qualification and sensitive attribute $f(x|y, s)$ does not change over time (e.g., individuals need a certain qualification to generate certain non-sensitive features). Additionally, we assume that the policy $\pi(d \mid x, s)$ can be chosen by a policymaker and may depend on time. Under these assumptions, the probability of a feature change depends solely on the policy π and sensitive feature S.