SIMULATING EUTOPIA: LONG-TERM FAIRNESS WITH OUTCOMES, PERFORMATIVITY, AND DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

As AI-driven Decision Makers (ADMs) influence our socioeconomic reality, their roles in both enhancing efficiency and amplifying the social biases have drawn significant attention. In this paper, we revisit the nuances of *long-term fairness* achievable by an ADM– specifically, through the lens of a loan approver inducing a population-level wealth dynamics. The literature generally (a) considers passive environments, i.e. the decisions of an ADM does not change the population's behaviour, and (b) measures bias in terms of disparity in the instantaneous decisions rather than downstream equity. Modern ADMs challenge both the notions. To address these caveats, we first formalise the wealth dynamics induced by a loan approving ADM interacting with a multi-demographic population as a Performative Markov Decision Process. Then, we mitigate the absence of such a performative test-bed by developing Eutopia: a lending-process simulator enabled with a novel performative data generator to learn long-term fair strategies. With Eutopia, we test reinforcement learning algorithms with different fairness-aware utilities dependent on approval decisions and downstream wealth. Results show that performative dynamics-aware learning with fairness-aware utility that incorporates the downstream outcomes induce better long-term equity and inclusivity.

1 INTRODUCTION

Modern AI-driven Decision Makers (ADMs) are not passive. We live in a symbiosis with these digital companions. While they bring efficiency to our high and low-stake daily decisions, they also change our socio-politico-economic behaviours over time. ADMs bring efficiency in applications ranging from credit scoring and loan underwriting (Fuster et al., 2022) to hiring (Raghavan et al., 2020), and criminal justice (Mehrabi et al., 2021). In tandem, they influence how we conduct job or loan applications, how we do medical diagnosis, how we even write papers and conduct research. This reactive effect induced by an algorithm is called *performativity* (Perdomo et al., 2020; Góis et al., 2024) and ADM's ability to influence human behaviour is measured by its *performative power* (Hardt et al., 2022). This phenomenon invokes the question:

What are the long-term effects of performative interactions between ADMs and society, in terms of equity?

Fairness in Short and Long-term. *Efficiency* and *equity* are not the same. Since ADMs learn from historical data mirroring the existing social inequalities, an unexamined pursuit of predictive accuracy can reproduce, even amplify the structural disadvantages faced by marginalised groups (Obermeyer et al., 2019; Angwin et al., 2022). This has propelled the field of *algorithmic fairness*. Researchers developed different metrics to quantify bias (Corbett-Davies et al., 2023), auditors to measure bias induced by ADMs (Ajarra et al., 2025), and algorithm designs to mitigate bias of predictive systems (Barocas et al., 2023). The initial works evaluated fairness at a *single decision point*, treating the population of applicants as a fixed, exogenous distribution (D'Amour et al., 2020). But decision making in real-world is a continual, interactive process. For example, a lending decision is not a passive observation, rather an intervention. Approving a loan changes borrower's wealth (Figure 1) that affects their future creditworthiness and also the dynamics of future applicants. Liu et al. (2018); Hashimoto et al. (2018) show that learning to maximise *short-term fairness* (e.g. demographic parity in one-shot lending) can *reduce* long-run welfare of disadvantaged groups by denying them the very opportunities. Hence, *long-term fairness is studied as sequential problem over multiple interactions*.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This has led to Reinforcement Learning (RL) driven study of long-term fairness in ADMs. Jabbari et al. (2017) formalise fairness in *Markov Decision Process* (MDP) setting– requiring no action to make any group worse off relative to a baseline. Subsequent works developed general fairness constraints, and computational-statistical techniques to ensure better efficiency-equity trade-offs in long-term (Hu & Zhang, 2022; Deng et al., 2022; Ju et al., 2023; Wang et al., 2024). But an overarching assumption persists that the *transition dynamics of the population are exogenous*. The human population is assumed to evolve according to fixed, policy-independent rules, which is not true for modern ADMs.

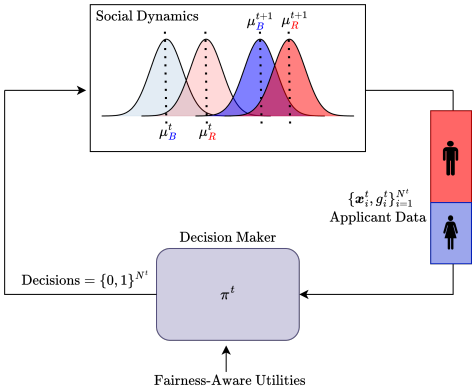


Figure 1: Sequential Credit Lending Problem

Contribution 1: A Performative Formalism

We formalise sequential credit lending as a *Performative MDP* (Section 2), i.e., the deployed policy shapes both wealth dynamics and future applications. It extends the MDP-centric long-term fairness literature.

In this model, the state consists of features from the cohort of arriving applicants, the action is a vector of accept/reject decisions, and the reward is ADM's profit (Eq. (2)). Approval increases an applicant's wealth if the loan is repaid (Eq. (1)), and a group's application rate responds to the decisions it has observed throughout. The ADM optimises discounted cumulative reward over interactions, which forces it to reason about the long-term steady state. Additionally, we show precisely why this algorithmic feedback loop cannot be absorbed simply by enlarged state-augmentation (Remark 1).

A formalism defines the algorithmic feedback loop but to use it for real-life sequential decisions, one has to realise the shift in population dynamics under a deployed policy. No standard empirical resource allows this. *Static credit* datasets are snapshots taken under a single, fixed acceptance rule. They contain no repeated interaction, or variation in the deployed policy, which are exactly the ingredients needed to identify performative responses (Perdomo et al., 2020). In contrast, *live deployment* captures the dynamics but allows little space to experiment. Measuring long-term harm to a protected group means deploying rules one expects to fail, which is neither ethically defensible nor compatible with fair-lending regulations. Simulation is the established answer to this bind (D'Amour et al., 2020), and it is missing for the performative sequential lending problem.

Contribution 2: Eutopia: a Performative Lending Simulator with Deomgraphics

We build Eutopia (Section 3), which realises the Performative MDP as a closed loop of *Decision*, *Evaluation*, and *Social Dynamics* blocks fed by a data generator. To our knowledge, it is the first simulator for long-term lending in a performative environment.

In *Decision block*, an ADM deploys a policy trained on a chosen fairness-aware utility to guide decisions for the arriving cohort. *Evaluation block* realises the consequences of the decisions taken. Loans are repaid or defaulted, as the ADM obtains its profit. *Social Dynamics block* closes the loop as wealth evolves, and a Hawkes process decides who applies next. The data generator grounds all of this on real data (Becker & Kohavi, 1996), and supplies the observations as per its response to the policy. *We deploy Eutopia to compare standard RL against performative RL learners across outcome and decision-level fairness measures with three ways of turning them into a constraint.*

Contribution 3: Long-term fairness is achievable with outcome parity and not decision-level parity

Long-term equity requires deploying a performativity-aware learner trained with a dynamic fairness-aware utility measured on outcomes.

Four empirical findings drive this conclusion (Section 4). *(a) Performative learning matters*: Performative RL agent beats its standard RL counterparts on all the metrics in every scenario, while the RL agent collapses to approving almost everyone it receives. *(b) Decision-level parity renders delayed impact of fairness*: An ADM trained for Equality of Opportunity meets its criterion early. After which its constraint stops binding and it behaves like a profit-driven utilitarian agent– reproducing the *delayed impact of fairness* (Liu et al., 2018) inside a performative environment. *(c) FL under*

Equality of Outcome renders best equity: Fairness Lagrangian under Equality of Outcome closes the wealth gap while retaining two-thirds of the maximal profit. *(d) Equality of Opportunity is selective, Outcome is inclusive*: Under Equality of Opportunity, 15 to 26% of each group never receives a loan and women gradually stop applying, whereas under Equality of Outcome nearly everyone is served.

2 Sequential Lending as an Average Reward Performative MDP

First, we formalise the sequential credit lending problem (Figure 1) that induces a social wealth dynamics. We consider a population of N applicants– each with a feature vector $\mathbf{x}_i \in \mathbb{R}^d$, and a demographic group $g_i \in \{\text{red}, \text{blue}\}$. We use the first feature $\mathbf{x}_i[1]$ to denote an applicant's wealth.[1] Each applicant requests for a loan $l_i \in \mathbb{R}_{>0}$ and yields a binary repayment outcome $Y_i \in \{0, 1\}$, where $Y_i = 1$ denotes successful repayment of loan. Importantly, Y_i is observed *only for approved applicants*. The ADM never learns what would have happened if a denied applicant would have received a loan, known as *selective label bias* (Lakkaraju et al., 2017).

Interaction Dynamics. At each time-step $t \in \mathbb{N}$, a set of $N^t \in [N]$ applicants arrive with their corresponding probability of loan application $\{p_{\text{apply}}{}_i^t\}_{i=1}^{N^t}$. Thus, the **state** is defined as the full context of all arriving applicants as $\mathbf{S}^t = \{\mathbf{x}_i, l_i, g_i\}_{i=1}^{N^t}$. Upon observing the state, the ADM takes a batch of binary **actions** $\mathbf{A}^t \in \{\perp, \not\perp\}^{N^t}$, where $\perp$ and $\not\perp$ denote approval and denial. The expected **transition** of applicant i's wealth under a deployed policy $\boldsymbol{\pi}$ is expressed as:

$$\mathbb{E}\left[\mathbf{X}_i^{t+1}[1]\right] = \underbrace{(\mathbf{x}_i^t[1] + \kappa_i) \cdot p_{\text{apply}}{}_i^t \cdot \mathbb{P}_{\boldsymbol{\pi}}\left(\mathbf{A}_i^t = \perp\right) \cdot \mathbb{P}(Y_i^t = 1)}_{\text{successful repayment}}$$
$$+ \underbrace{\mathbf{x}_i^t[1] \cdot p_{\text{apply}}{}_i^t \cdot \mathbb{P}_{\boldsymbol{\pi}}\left(\mathbf{A}_i^t = \perp\right) \cdot \mathbb{P}(Y_i^t = 0)}_{\text{default on repayment}} + \underbrace{\mathbf{x}_i^t[1] \cdot p_{\text{apply}}{}_i^t \cdot \mathbb{P}_{\boldsymbol{\pi}}\left(\mathbf{A}_i^t = \not\perp\right)}_{\text{loan denied}}. \quad (1)$$

Equation (1) encapsulates three disjoint events: (a) the expected wealth gain ($\kappa_i > 0$) upon successful repayment that depends on l_i; (b) no expected wealth gain if applicant i fails to repay the loan amount l_i; and finally, (c) no expected wealth gain if the loan is denied.

ADM's Decision Making as a PeMDP. Since the standard MDP formalism (Sutton et al., 1998) of RL remains oblivious towards the *algorithmic feedback loop* causing *endogenous* shifts in applicants' behaviour based on the ADM's decisions in this wealth process, we model long-term fair lending as a *Performative Markov Decision Process* (PeMDP) (Mandal et al., 2023b). Given a policy set $\Pi : \mathcal{S} \to \Delta(|\mathcal{A}|)$ mapping states to (distribution over) actions, the PeMDP is defined as the set of MDPs $\mathcal{M}(\Pi) \triangleq \{\mathcal{M}(\boldsymbol{\pi}) \mid \boldsymbol{\pi} \in \Pi\}$, where each MDP is a tuple $\mathcal{M}(\boldsymbol{\pi}) \triangleq (\mathcal{S}, \mathcal{A}, \mathbf{P}_{\boldsymbol{\pi}}, r_{\boldsymbol{\pi}})$. $\mathbf{P}_{\boldsymbol{\pi}}$ and $r_{\boldsymbol{\pi}}$ denote the transitions (Eq. (1)) and rewards (Eq. (2)) corresponding to a deployed policy $\boldsymbol{\pi}$. Specifically, the ADM's per-step expected reward at interest rate $\tau_{\text{interest}} \in (0, 1]$ is

$$\mathbb{E}_{\boldsymbol{\pi}}\left[r_{\boldsymbol{\pi}}^t\right] = \sum_{i=1}^{N_t} \Big(\tau_{\text{interest}}\, \mathbb{P}(Y_i^t = 1) - \mathbb{P}(Y_i^t = 0)\Big) \cdot \mathbb{P}_{\boldsymbol{\pi}}(\mathbf{A}_i^t = \perp) \cdot l_i^t \cdot p_{\text{apply}}{}_i^t. \quad (2)$$

The ADM adopts a infinite-horizon discounted objective and uses a Reinforcement Learning (RL) algorithm to maximise the value $V_{\boldsymbol{\pi}} \triangleq \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=0}^{\infty} \zeta^t r_{\boldsymbol{\pi}}^t\right]$. Optimising $V_{\boldsymbol{\pi}}$ needs reasoning about the long-term steady-state behaviour, and not just immediate returns.

A *non-performative* MDP setting assumes $p_{\text{apply}}{}^t$ to be stationary i.e., the application behaviour remains unaffected by the ADM's historical behaviour. In contrast, a *performative* PeMDP setting models the application rate based on the observed approval patterns: $p_{\text{apply}}{}^{t+1} = f_\psi\left(p_{\text{apply}}{}^t, \mathbf{S}^t, \mathbb{P}_{\boldsymbol{\pi}}(\mathbf{A}^t = \perp | \mathbf{S}^t)\right)$, where f_ψ is a parametric response function, and $\mathbb{P}_{\boldsymbol{\pi}}(\mathbf{A}^t = \perp | \mathbf{S}^t)$ is the approval probability in the current state $\mathbf{S}^t$ under the deployed policy $\boldsymbol{\pi}$. Intuitively, this captures the effect that if a demographic group is consistently rejected, they will gradually stop applying, making it harder for the ADM to learn a fair policy for that group. This feedback, where the policy shapes application behaviour, and in turn, constrains what future policies can learn, *is absent from the existing MDP-based long-term fairness works* (Jabbari et al., 2017; Hu & Zhang, 2022).

[1] *Notations.* X, x denote a random variable and its realisation, respectively. $y, \mathbf{y}$ denote a scalar and a vector, respectively. Whenever we write z_i^t, it corresponds to applicant i. Otherwise, z^t denotes $\{z_i^t\}_{i=1}^{N^t}$.

Remark 1 (Why PeMDP and not Augmented MDP?). *An alternative formalism is to absorb the feedback loop into an augmented state* $\bar{\mathbf{S}}^t \triangleq (\mathbf{S}^t, {p_{\text{apply}}}^t)$. *Since* ${p_{\text{apply}}}^t$ *is not explicitly observed, it becomes a POMDP (Smallwood & Sondik, 1973), and standard discounted reward RL algorithms do not work. In addition, under sustained denial, if a certain group's application rate collapses, no data from that group would be generated. Yet the dynamics continues to depend on* $\boldsymbol{\pi}$*'s impact on the applicants. Thus, two policies agreeing on the visited region of* $\mathcal{S}$ *may still induce different wealth trajectories while being indistinguishable from any collected trajectory. Repairing this by state augmentation would require encoding* $\boldsymbol{\pi}$ *itself in the state, which is not a well-posed RL problem.*

3 Eutopia: Simulating Long-term Impact of Algorithmic Decisions

In this section, we develop the fair lending simulator Eutopia that numerically deploys the sequential lending PeMDP of Section 2. Eutopia realises $\mathcal{M}(\boldsymbol{\pi}^t)$ and its long-term evolution as a closed loop of three blocks (Figure 2): (1) *Decision block* that allows an ADM to learn a fairness-aware policy $\boldsymbol{\pi}^t$ using an RL algorithm of choice and yields the lending decisions $\mathbf{A}^t$; (2) *Evaluation block* that realises the consequences of credit lending actions; and (3) *Social Dynamics block* that allows the ADM's actions to reshape the population dynamics that the ADM encounters in the next step.

Interaction protocol. Eutopia yields a single continual trajectory to accumulate the long-term effects of algorithmic and performative feedback loop. Thus, cumulative counters such as wealth and application histories are never reset. At every step t:

I. Decisions: The Data Generator generates the arriving applicant cohort $\mathbf{S}^t = \{\mathbf{x}^t, g^t, l^t\} = \{(\mathbf{x}_i^t, g_i, l_i)\}_{i=1}^{N^t}$. The ADM observes $\mathbf{S}^t$ provided by the Data Generator, and deploys a history-aware policy $\boldsymbol{\pi}^t$ rendering decisions $\mathbf{A}^t \in \{\perp, \not\perp\}^{N^t}$.

II. Evaluation: As repayment $\{Y_i^t\}$ of each approved applicant is observed, the ADM accumulates reward $r_{\boldsymbol{\pi}}^t$ according to Equation (2), and each repaying borrower gains κ_i in wealth.

III. Social dynamics: The observed wealth gains $\{\kappa_i\}_{i=1}^{N^t}$ update population wealth to $\{\mathbf{x}_i^{t+1}[1]\}$ (Eq. (1)). Shifted wealth, and the approval history dictate each individual's probability of next application ${p_{\text{apply}}}_i^{t+1}$ and the next cohort size N^{t+1}. Finally, the Data Generator yields the observation space $\mathbf{S}^{t+1}$ for the new cohort of applicants.

Binding the loop: Learning to lend fairly. After $T/\Delta t$ steps, referred as an episode, the chosen RL algorithm updates $\boldsymbol{\pi}^t$ such that it optimises a chosen fairness-aware utility under the shifted social dynamics.

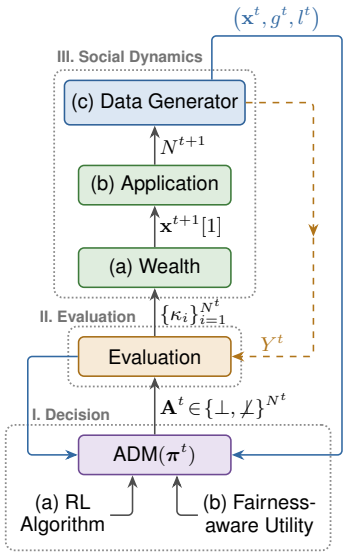


Figure 2: Eutopia : Interaction pipeline.

Steps 2–3 realise the transition kernel $\mathbf{P}_{\boldsymbol{\pi}}$. Holding the application probability constant removes its dependence on $\boldsymbol{\pi}$ and recovers the non-performative MDP. Thus, Eutopia remains compatible for direct comparison. We refer to Appendix D for the detailed pseudo-codes. In the next three subsections, we present each block of Eutopia and its components, in details.

3.1 Decision Block: How the ADM Acts

The Decision Block comprises the ADM, i.e. (a) the RL algorithm used to train decision-making policy $\boldsymbol{\pi}^t$, and (b) the fairness-aware utility used to train.

(a) RL Algorithm: Learning to decide. The RL algorithm takes a cohort of applicants' data, and its history of loan application and repayment as an input and computes an approval policy $\boldsymbol{\pi}^t$ across applicants. The decisions are drawn as $\mathbf{A}^t \sim \boldsymbol{\pi}^t(\cdot \mid \mathbf{S}^t, \mathcal{H}^{t-1})$, where $\mathcal{H}^{t-1}$ consists of the application and repayment history of the cohort till time $t-1$. We choose to work with Beta policies to have scores for and against our yielded decisions, and keep them differentiable for further algorithm design. After each episode, we update the policy to maximise an average fairness-aware utility (Equation (3)) estimated with the historical trajectory.

Table 1: The constraint C^t instantiated under each fairness perspective.

Fairness Constraint	Equality of Outcome	Equality of Opportunity
Social Welfare (SW)	$\mu^t_{\text{blue}} + \mu^t_{\text{red}}$	$\text{TPR}^t_{\text{blue}} + \text{TPR}^t_{\text{red}}$
Rawlsian Max-Min (RMM)	$\min\{\mu^t_{\text{blue}}, \mu^t_{\text{red}}\}$	$\min\{\text{TPR}^t_{\text{blue}}, \text{TPR}^t_{\text{red}}\}$
Fairness Lagrangian (FL)	$\mid \mu^t_{\text{blue}} - \mu^t_{\text{red}} \mid$	$\mid \text{TPR}^t_{\text{blue}} - \text{TPR}^t_{\text{red}} \mid$

(b) Fairness-aware utility. The decision behaviour of a learning agent is guided by the utility function it aims to maximise. Thus, we make this choice explicit along two axes. The first axis is the *fairness perspective* that leads to the question:

Should the ADM seek parity in the decisions it takes, or in the actual outcomes after deployment?

This invokes two perspectives on fairness (Table 1): (a) *Equality of Outcome* asks whether the two groups end up equally wealthy, captured by the group's mean wealth till time t, i.e., μ^t_g, and (b) *Equality of Opportunity* (Hardt et al., 2016) asks whether equally creditworthy applicants are approved at equal rates regardless of the group, captured by group-specific true positive rate till time t, i.e., TPR^t_g. These two fairness perspectives need not agree. A policy can equalise approval rates among creditworthy applicants while still producing sharply diverging wealth trajectories. We study both perspectives separately to understand their individual impacts in long-term.

The second axis is the *fairness constraint* that turns the two groups' wealth/opportunity statistics into the quantity C^t (Table 1): Social Welfare (SW) aims to maximise the aggregated wealth/opportunity (Harsanyi, 1955), Rawlsian Max-Min (RMM) considers maximising wealth/opportunity of the worse-off protected group (Rawls, 1971), and the Fairness Lagrangian (FL) (Xu et al., 2024) penalises discrepancies between the two protected groups.

The ADM aims to maximise discounted cumulative reward subject to the constraint C^t, and thus, deploys an RL algorithm that episodically optimises the dual objective $\max_{\boldsymbol{\pi}\in\Pi} \mathcal{L}(\boldsymbol{\pi}, \gamma)$, where

$$\mathcal{L}(\boldsymbol{\pi}, \gamma) \triangleq \frac{1}{T}\sum_{t=1}^{T}\Big(\mathbb{E}_{\boldsymbol{\pi}}\left[r^t_{\boldsymbol{\pi}}\right] + s\gamma C^t\Big), \qquad s = \begin{cases} +1 & \text{SW, RMM,} \\ -1 & \text{FL,} \end{cases} \quad \gamma \geq 0. \tag{3}$$

The Lagrangian multiplier γ is learned using dual ascent. It rises while the constraint is violated and decays once it is satisfied (Algo. 2). Setting $\gamma \equiv 0$ recovers the solely profit-driven MAXUTIL agent (D'Amour et al., 2020). For each fairness constraint and perspective, we deploy and compare both RL (Policy Gradient and Soft Actor-Critic) and performative (PePG) RL algorithms (Section 4.1).

3.2 Evaluation Block: What the Decisions Realise

Evaluation block turns decisions into outcomes. For each approved applicant i, repayment is drawn as $Y^t_i \sim \text{Bernoulli}(1-d_i)$. d_i is the applicant's default probability obtained from the data generator, which the ADM never observes. Denied applicants generate no label emulating the selective-label bias (Lakkaraju et al., 2017) in a controlled setting. The ADM receives the reward according to Equation (2): $\tau_{\text{interest}} l_i$ on repayment, $-l_i$ on default, and 0 on denial. The applicant gains wealth $\kappa_i = \tau_{\text{gain}} l_i Y^t_i$, where $\tau_{\text{gain}} > 0$ is the return from investing the loan net of interest.

Thus, a repaid loan grows ADM profit and applicant's wealth at once, while a default costs the ADM and leaves the applicant no better off. This is the tension our RL agent must navigate while learning any of the utility functions (Equation (3)). The realisations are sent to the Dynamics block that forms the group statistics μ^t_g and TPR^t_g from which C^{t+1} is computed, and to the Decision block that keeps track of the cumulative counts of loans and defaults by applicants as $\mathcal{H}^t$.

3.3 Social Dynamics Block: How the Population Responds

Social Dynamics block closes the feedback loop of Eutopia. Equipped with realisations from the Evaluation block, it updates wealth, decides on who applies next, and realises both with real data.

(a) Wealth evolution. Each borrower's gain is added to their wealth, $\mathbf{x}^{t+1}_i[1] = \mathbf{x}^t_i[1] + \kappa_i$, depending on whether i applies, is approved and repays recovers Equation (1). For simplicity, we keep the financial type (d_i, l_i, κ_i) fixed for each applicant in experiments but Eutopia provides the flexibility to vary them over time.

(b) Application dynamics. At each time step $t \in \mathbb{N}$, loan applications arrive in cohorts of N^t. In real-life lending systems, an approval signals access to credit and encourages further applications, while sustained denial discourages them. This phenomenon is known as "Matthew effect" (Merton, 1968). Thus, we model each group's application intensity as a Hawkes process (Hawkes, 1971):

$$\lambda_g^t = f(\mu_g^t) + \alpha_g \sum_{t_j \in \mathcal{H}_g^t} e^{-\beta_g (t - t_j)}. \tag{4}$$

The base rate $f(\mu_g^t)$ increases with the group's mean wealth relative to the population that means people from wealthier groups apply more. The second term adds an excitation for each past approval $t_j \in \mathcal{H}_g^t$ decaying at rate β_g. We set $\alpha_g = 0.3$ and $\beta_g = 2.0$ for both groups. Thus, any asymmetry among groups that emerges over time is caused by the policy, not built into the application arrivals. Individual i then applies with probability $p_{\text{apply}_i}^t \propto \frac{\lambda_{g_i}^t}{N_{g_i}}$. This deployment realises f_ψ of Section 2, i.e., the policy influences future applications through the approval history $\mathcal{H}_g$ and through wealth. Thus, a group that is consistently denied loses both, applies less, and generates less of the data a ADM would need to serve it fairly. Conversely, loan approvals encourages people from a protected group to apply more in future, accumulate more wealth, hence preserving the Matthew effect in a sequential lending system in a performative environment.

(c) Data Generator: realising the state. For the ADM to act, the state $\mathbf{S}^t$ must be populated with applicants. The Data Generator supplies them. Unlike a static dataset, it produces both an initial population and the parameters governing its response to the policy. It builds on the UCI Adult Income dataset (Becker & Kohavi, 1996). Hence, the initial population carries real demographic and economic heterogeneity. For each individual i, it generates a wealth $\mathbf{x}_i[1]$; a protected group g_i (gender– red for male and blue for female as per the dataset's documented wealth disparity); a fixed financial type (l_i, d_i, κ_i), i.e. loan amount and *fixed* default probability, and gain on repayment respectively. At each step $t \in \mathbb{N}$, Data Generator passes the arriving applicants' data $\{\mathbf{x}_i^t, g_i, l_i, \kappa_i\}_{i=1}^{N^t}$ to the ADM except $\{d_i\}_{i=1}^{N^t}$ that Evaluation block uses to tally repayments $\{Y_i^t\}_{i=1}^{N^t}$ following loan approvals. With this 12-dimensional state space yielded by the data generator, the ADM's policy combines the applicants' features and repayment history $\mathcal{H}^t$ for further decision making. Appendix E.1 and E.2 provide further details on the data generator and Eutopia.

4 Design Choices Matter for Achieving Long-term Equity

We now use Eutopia to study which learner and which utility can actually deliver long-term equity. We first fix the setting, and then answer three questions.

Experimental setting: which learners, and why. We deploy two RL and a performative RL algorithm to compare the effect of performative learning in long-term fairness-utility trade-off.

The RL algorithm: PG. As RL algorithms, we test vanilla Policy Gradient (PG) (Williams, 1992) and Soft Actor-Critic (SAC) (Haarnoja et al., 2018). In the main text, we describe PG as it emerges as the better baseline. SAC learns a fixed score cutoff in every scenario. Its Lagrangian multiplier sits at the floor for most of deployment, and four out of six scenarios end at the same point (Appendix F.4). As SAC ignores its fairness-aware utility, it cannot tell us how these utilities shape long-term outcomes.

The performative RL algorithm: PePG. Most of the performative RL algorithms (Mandal et al., 2023b; Rank et al., 2024b) aim to converge to *performative stability* i.e., a policy that is optimal for the environment it induces. A stable policy can be far from the optimal one (Perdomo et al., 2020; Miller et al., 2021). ADM aims to deploy the policy that maximises $V_{\boldsymbol{\pi}}$ over the whole family $\mathcal{M}(\Pi)$, i.e. achieving *performative optimality*. Performative Policy Gradient (PePG) (Basu et al., 2025) exactly solves this. It differentiates through the population's response to the policy. It is the performative version of PG, thus making the comparison fair and accurate. *We call them RL and PeRL agents for the rest of the paper.* Note that there exists no performative variant for SAC.

Training details. Both agents use a Beta policy over the approval probability, a two-layer MLP (width 128) and AdamW optimiser (Loshchilov & Hutter, 2017) (learning rate 10^{-3}). Each group has $N_M = N_F = 12{,}000$ individuals. An episode spans 100 time units in 200 decision steps. Each agent is trained for 1,000 episodes, and then deployed for 3,000 episodes in Eutopia. The multiplier γ^t is set via dual ascent (step 0.1). Every experiment is run and averaged across 10 seeds. Further training details and hyperparameters are described in Appendix E.3.2.

Table 2: End-of-deployment Performance of RL and PeRL agents across fairness-aware utilities (mean ± std over 10 seeds after 3000 episodes). For each metric, the better agent is in **bold**. ** denotes the winners for each metric.

Perspective	Constraint	Cumulative Profit ($\times 10^6$)		|Wealth Gap|		Inequality Ratio ρ	
		RL	PeRL	RL	PeRL	RL	PeRL
Outcome	FL	0.374 ± 0.385	$\mathbf{4.909 \pm 0.579}$	745.0 ± 1288.3	$\mathbf{46.5 \pm 82.1}^{**}$	0.190 ± 0.132	$\mathbf{1.045 \pm 0.032}^{**}$
	RMM	1.096 ± 0.096	$\mathbf{1.267 \pm 0.058}$	3190.0 ± 68.9	$\mathbf{3081.9 \pm 26.5}$	10.198 ± 0.032	$\mathbf{0.101 \pm 0.000}$
	SW	1.103 ± 0.071	$\mathbf{4.721 \pm 1.346}$	3169.8 ± 82.5	$\mathbf{1956.0 \pm 1012.6}$	10.294 ± 0.082	$\mathbf{0.138 \pm 0.048}$
Opportunity	FL	1.131 ± 0.062	$\mathbf{7.443 \pm 0.410}$	3154.8 ± 103.9	$\mathbf{1086.8 \pm 332.2}$	10.328 ± 0.045	$\mathbf{3.981 \pm 0.568}$
	RMM	1.105 ± 0.100	$\mathbf{6.711 \pm 1.229}$	3177.3 ± 55.4	$\mathbf{207.9 \pm 1328.7}$	10.231 ± 0.043	$\mathbf{1.229 \pm 0.643}$
	SW	1.118 ± 0.099	$\mathbf{6.757 \pm 1.137}$	3115.9 ± 136.1	$\mathbf{535.2 \pm 1187.0}$	10.304 ± 0.079	$\mathbf{1.776 \pm 0.878}$
MaxUtil		1.092 ± 0.101	$\mathbf{7.495 \pm 0.194}^{**}$	3137.6 ± 125.1	$\mathbf{1181.2 \pm 110.0}$	10.369 ± 0.105	$\mathbf{4.672 \pm 0.195}$
Winner		**PeRL+Opportunity+FL**		**PeRL+Outcome+FL**		**PeRL+Outcome+FL**	

Performance metrics. We track three macro metrics while simulating the long-term interactions between the RL and PeRL agents and Eutopia across the two fairness perspectives and three fairness constraints (Table 1). *(a) Wealth gap*: $\Delta_\mu^t \triangleq |\mu_M^t - \mu_F^t|$. $\Delta_\mu^t = 0$ indicates outcome parity. *(b) Inequality ratio*: $\rho^t \triangleq (\mu_M^t - \mu_M^0)/(\mu_F^t - \mu_F^0)$. $\rho^t = 1$ means both groups grow at the same rate, and $\rho^t > 1$ means the male group grows faster. *(c) Cumulative profit*: $\mathrm{CP}^t \triangleq \sum_{\tau \le t} r_{\boldsymbol{\pi}}^\tau$. This is the ADM's utilitarian gain, while the other two metrics capture equity. Specifically, we ask three empirical questions:

Q1. Does being reactive to the performative shifts in the environment yields better equitable outcomes for the applicants in long-term?
Q2. Which fairness-aware utility function (fairness perspective + constraint) should be chosen by a practitioner to ensure fairness-utility trade-off in long-term?
Q3. How much different fairness-aware utilities can ensure intra-group fairness in long-term?

4.1 Q1: Performative RL for Better Long-term Fairness-Utility Trade-off

To address Q1, we report the best-performing agents per metric achieved by the end of deployment in Table 2. Trajectory-level results and detailed discussions are in Appendix F.2.

Performative RL leads to better fairness-utility trade-off. PeRL performs significantly better in terms of higher cumulative profit, smaller wealth gap, and an inequality ratio closer to one than RL (Table 2). Under Equality of Opportunity, PeRL earns 6–7× RL's profit while cutting the wealth gap by a factor of 3–15, and under FL with Equality of Outcome, it reduces the wealth gap 16× relative to RL, with an inequality ratio $\rho = 1.045$ against RL's 0.19.

Outcome- and performativity-oblivious RL converges to approve most applicants. Except the case of FL under Equality of Outcome, RL's end-of-deployment performance metrics are surprisingly indistinguishable from one another. These are, to within seed noise, the metrics demonstrated by Always Approve rule (Appendix E.3.1). Under RMM and SW with either perspective, and under FL with Equality of Opportunity, RL eventually settles on approving nearly every applicant of both groups. A lender who approves everyone treats the groups identically by construction. This leads to lower approval rate disparity, while the Matthew effect drives the wealth gap and ρ to similar values as a heuristic policy. Only in case of FL under Equality of Outcome, it recalibrates. As a result, the Female group ends up growing five times faster than the red one ($\rho = 0.19$).

4.2 Q2: A Practitioner's Guide to Designing a Fairness-aware Utility

Section 4.1 establishes the fact that deploying performative agents yields better long-run equitable outcomes in society. Now, using performative RL, we investigate which combination of *fairness perspective and constraint* (Table 1) should be learnt for obtaining the best fairness-utility trade-off.

FL under Equality of Outcome renders best long-term equity. This is the only utility function that closes the wealth gap. It leads to a gap of 46.5 and inequality ratio $\rho = 1.05$, which are the best values across the grid in Table 2. Moving from evaluating FL under Opportunity vs. Outcome shrinks the wealth gap 23× and brings down ρ from 3.98 to 1.05. It still yields two thirds of the MaxUtil's profit (4.91 against 7.50 Million). Thus, the utility trade-off is modest.

FL is the only constraint that penalises the gap between the protected groups rather than optimising an aggregate of their statistics. SW, on the other hand maximises a sum, oblivious to how a fixed total is distributed across groups (Harsanyi, 1955). RMM is flat in the better-off group's statistic (Rawls, 1971), so the learning signal is one-sided and vanishes wherever the $\arg\min$ is stable. The resulting behaviour is visible in terms of ρ being far away from 1, and huge wealth gaps.

Equality of Opportunity reproduces the delayed impact of fairness. Equality of Opportunity agents fail in long-term exactly as the literature predicts. FL under Opportunity ends with a gap of $\sim 1{,}087$ and $\rho = 3.98$. This is close to MAXUTIL (1,181.2, $\rho = 4.67$), i.e., an agent that ignores fairness altogether. The reason is that TPR parity is a decision-level criterion. As soon as the agent meets it early: approval-rate disparity falls to 0.01 (Appendix F.2.2), and the constraint no more binds. But equal approval rates are not equal outcomes. As the male group starts wealthier, they apply more. Over deployment, they file four times more applications and receives four times the loans per person (Appendix F.3). Its wealth pulls ahead. This is known as the *"delayed impact of fairness"*– a criterion satisfied at the decision-level harms long-term equity (Liu et al., 2018). In the non-performative case, D'Amour et al. (2020) observed that an Equal-Opportunity (EO) agent ensuring instantaneous decision-level parity causes long-term harm in outcomes of the disadvantageous group, while MAXUTIL agent does not. Eutopia reproduces the results of EO agent for end-of-deployment macro metric and trajectory-level micro metrics (Section 4.3), while demonstrating MAXUTIL agent to be equally harmful in long-term. *FL under Outcome escapes it by constraining the outcome disparity itself as the constraint cannot be satisfied while the harm accumulates.*

Only FL under Equality of Outcome keeps the fairness-utility trade-off alive. Figure 3 tracks the multiplier γ^t quantifying the balance between profit and fairness. For four out of six utility functions, it decays to floor as t grows. At this point, the fairness constraint no longer binds and the agent maximises profit alone– essentially becoming a utilitarian MAXUTIL agent. Its end-of-deployment performance is then guided by the bias acquired early and further amplified by the Matthew effect. FL and RMM under Outcome are the exceptions. While RMM's Lagrangian multiplier converges around 500 episodes, γ^t for FL keeps evolving throughout. This persistent balancing between profit and fairness is precisely what a fairness-aware utility is meant to reflect. While the TPR gap is met once approval rates match, the wealth gap works differently. The Matthew effect keeps pushing it to open, and thus, the outcome-based constraint must keep binding like FL under Outcome.

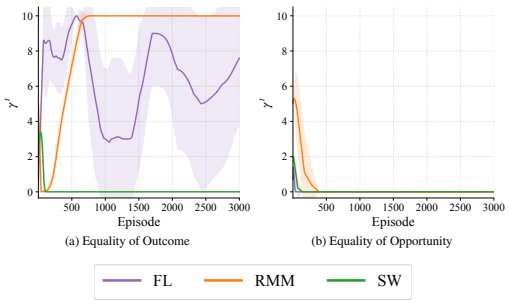


Figure 3: Evolution of the multiplier γ^t over deployment for PeRL agent.

Another concern might surface from Table 2: for RMM and SW, every macro metric is better under Opportunity than Outcome unlike FL. A practitioner choosing utilities by macro metrics alone might prefer them. Here, we suggest looking at how credit is distributed inside the groups and over time.

4.3 Q3: Equality of Outcome Matters for Intra-group Fairness

Metrics such as wealth gap and inequality ratio measure fairness between groups, but they are oblivious towards how credit access is distributed within groups. A policy can achieve near-perfect group-level parity while systematically concentrating loans among the same high-wealth individuals in each group. To surface this, we conduct a micro-analysis. We report *Lorenz curves* to track cumulative loan concentration within groups over deployment (Figure 4), where the diagonal represents that the whole population in a group gets loan at least once. In Appendix F.2.4, we also report the *reach rate* plots for tracking the fraction of unique recipients per episode relative to the full group size, and cumulative loan distributions that reveal the within-group skewness of loans received in short-term vs. the full deployment horizon.

Equality of Outcome is inclusive; Equality of Opportunity is selective. Under Equality of Outcome, almost everyone in each group gets an opportunity. Almost no individual goes without a chance of getting loan under FL or RMM, and only 6% of men miss under SW. Under Opportunity, 15 to 26% of each group never receive a loan rendering the flat segments in corresponding Lorenz curves. The reason is what each perspective pays for. An Opportunity-driven utility pays for approving the

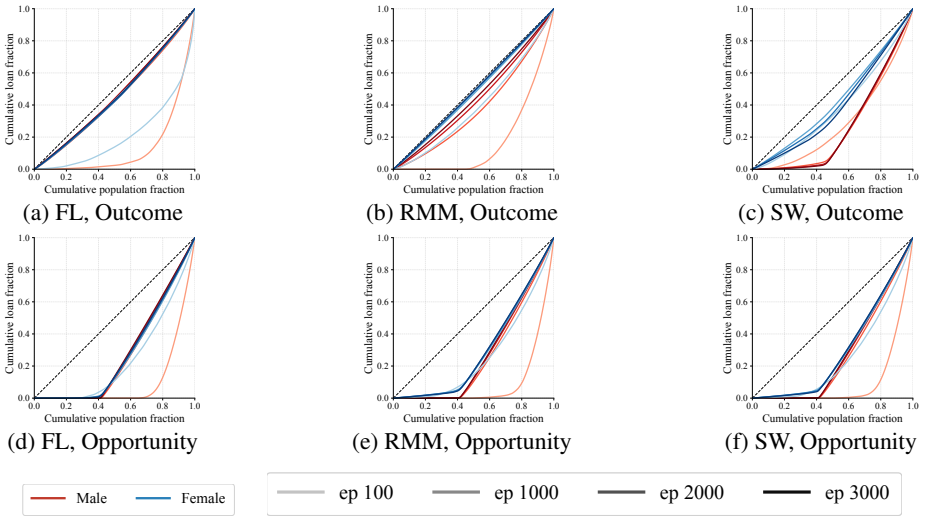


Figure 4: **Lorenz curves of cumulative loans for PeRL agent across utilities and over time.**

creditworthy and is oblivious to the rest. An Outcome-driven utility pays for wealth created across the group. Hence, leaving part of it unserved is costly.

Under Opportunity, the disadvantaged group stops applying. The female reach rate peaks early, at $\sim 5-7\%$, then declines to 1.8% under FL, 3.5% under SW and 5% under RMM. The male reach rate rises throughout, to 9-12%. TPR parity is therefore maintained on a shrinking pool of applicants. This is representation disparity (Hashimoto et al., 2018), and it is the micro-level trace of the decaying Lagrangian multiplier (Sec. 4.2). A decision-level criterion constrains who is approved among those who apply. It does not consider harms on the ones yet to apply (Zhang et al., 2019).

Macro metrics place RMM under Opportunity second based on end-of-deployment outcomes (207.9, $\rho = 1.23$). Micro-analysis shows that it leaves a quarter of men and a sixth of women unserved, and lets women's access decay. *Only FL under Outcome is equitable between groups, within groups, and over time.* An aggregated end-of-deployment evaluation cannot surface this effect (D'Amour et al., 2020). Thus, we advocate for FL under Equality of Outcome as *the* fairness-aware utility, since it is the one utility which pushes the ADM to be fair even when its instantaneous criterion is met, and hence, finally refutes the delayed impact of fairness (Liu et al., 2018).

5 Discussions, Limitations, and Open Questions

We revisit the long-term fairness achieved by an Algorithmic Decision Maker (ADM) interacting with a population. We built the first performative fair lending simulator Eutopia to observe: (a) a performative RL agent beats standard RL counterparts on profit, wealth gap and inequality ratio under all six fairness-aware utilities; (b) to ensure long-term equity, fairness must be measured based on outcomes rather than decisions; (c) while Equality of Opportunity based fairness objectives leave a fifth to a quarter of each group without a single loan and let the disadvantaged group's access decay throughout deployment, FL under Outcome is the only long-term utility ensuring equity between groups, within groups, and along the trajectory. It is, therefore, *the* utility design that escapes delayed impact of fairness in long-term (Liu et al., 2018).

Anticipating outcomes of long-term interactions is hard even in the simplest setups. Liu et al. (2018) show that a one-step feedback model already suffices for a static criterion to harm the group it protects. D'Amour et al. (2020) show that single-step analyses fail to predict long-term behaviour even in a two-group lending model with a fixed applicant pool. Each additional feedback channel broadens the possibilities between what is modelled and what can be claimed. We, therefore, position Eutopia as a *concrete test-bed for simulating long-term interaction between a ADM and a reactive population*, and not a universal recipe for ensuring long-term fairness. Rather, its purpose is to provide a comparative guideline for algorithms aiming for long-term equity: specifically, which objectives and learners sustain/degrade under the algorithmic feedback loop, in what way, and at what cost. In addition, Eutopia serves as a test-bed for performative RL algorithms. Our findings also call for further theoretical and algorithmic investigations of sequential decision making with outcome-based fairness measures.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods, with everything being reviewed by the authors. We have **not** used generative AI tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments. Additionally, we used generative AI tools to create and edit code. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our study primarily works on a scale of simulations. The dataset used in the simulation study is publicly available. No protected data has been used. All the fairness policies are mainly meant to be used on a theoretical scale, and the environment is meant to be only deployed in theoretical research scale.

REPRODUCIBILITY STATEMENT

In order to ensure reproducibility, all our experiments are run on fixed seeds. We provide all experimentation details to reproduce the results in Appendix E. Anonymous link to our code repository is https://anonymous.4open.science/r/LoanSimulator-D831/.

REFERENCES

Ayoub Ajarra, Bishwamittra Ghosh, and Debabrota Basu. Active fourier auditor for estimating distributional properties of ml models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 15330–15338, 2025. 1

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pp. 254–264. Auerbach Publications, 2022. 1

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023. 1

Debabrota Basu, Udvas Das, Brahim Driss, and Uddalak Mukherjee. Performative policy gradient: Optimality in performative reinforcement learning, 2025. URL https://arxiv.org/abs/2512.20576. 6, 16, 22, 37

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20. 2, 6, 21

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017. 16

Sam Corbett-Davies, Johann D Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. The measure and mismeasure of fairness. *Journal of Machine Learning Research*, 24(312):1–117, 2023. 1

Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, David Sculley, and Yoni Halpern. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 525–534, 2020. 1, 2, 5, 8, 9

Zhun Deng, He Sun, Zhiwei Steven Wu, Linjun Zhang, and David C Parkes. Reinforcement learning with stepwise fairness constraints. *arXiv preprint arXiv:2211.03994*, 2022. 2, 16

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018. 16

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012. 16

Andrew Estornell, Sanmay Das, Yang Liu, and Yevgeniy Vorobeychik. Group-fair classification with strategic agents. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pp. 389–399, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594006. URL https://doi.org/10.1145/3593013.3594006. 16

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

António Góis, Mehrnaz Mofakhami, Fernando P Santos, Gauthier Gidel, and Simon Lacoste-Julien. Performative prediction on games and mechanism design. *arXiv preprint arXiv:2408.05146*, 2024. 1

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. Pmlr, 2018. 6

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016. 5, 16

Moritz Hardt, Meena Jagadeesan, and Celestine Mendler-Dünner. Performative power. *Advances in Neural Information Processing Systems*, 35:22969–22981, 2022. 1

John C Harsanyi. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of political economy*, 63(4):309–321, 1955. 5, 8

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pp. 1929–1938. PMLR, 2018. 1, 9

Alan G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 04 1971. ISSN 0006-3444. doi: 10.1093/biomet/58.1.83. URL https://doi.org/10.1093/biomet/58.1.83. 6

Yaowei Hu and Lu Zhang. Achieving long-term fairness in sequential decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 9549–9557, 2022. 2, 3, 16

Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in reinforcement learning. In *International conference on machine learning*, pp. 1617–1626. PMLR, 2017. 2, 3

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. *Advances in neural information processing systems*, 29, 2016. 16

Peizhong Ju, Arnob Ghosh, and Ness Shroff. Achieving fairness in multi-agent mdp using reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2023. 2, 16

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016. 16

Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 275–284, 2017. 3, 5

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pp. 3150–3158. PMLR, 2018. 1, 2, 8, 9, 16

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

Debmalya Mandal, Stelios Triantafyllou, and Goran Radanovic. Performative reinforcement learning, 2023a. URL https://arxiv.org/abs/2207.00046. 16

Debmalya Mandal, Stelios Triantafyllou, and Goran Radanovic. Performative reinforcement learning. In *International Conference on Machine Learning*, pp. 23642–23680. PMLR, 2023b. 3, 6, 16

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021. 1

Robert K. Merton. The matthew effect in science. the reward and communication systems of science are considered. *PubMed*, 1968. 6

John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*, pp. 6917–6926. PMLR, 2020. 16

John P Miller, Juan C Perdomo, and Tijana Zrnic. Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learning*, pp. 7710–7720. PMLR, 2021. 6

Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P Dickerson. Fairness through robustness: Investigating robustness disparity in deep learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 466–477, 2021. 16

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. 1

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pp. 7599–7609. PMLR, 2020. 1, 2, 6, 16

Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 469–481, 2020. 1

Ben Rank, Stelios Triantafyllou, Debmalya Mandal, and Goran Radanovic. Performative reinforcement learning in gradually shifting environments, 2024a. URL https://arxiv.org/abs/2402.09838. 16

Ben Rank, Stelios Triantafyllou, Debmalya Mandal, and Goran Radanovic. Performative reinforcement learning in gradually shifting environments. *arXiv preprint arXiv:2402.09838*, 2024b. 6

John Rawls. *A Theory of Justice: Original Edition*. Harvard University Press, 1971. ISBN 9780674880108. URL http://www.jstor.org/stable/j.ctvjf9z6v. 5, 8

Richard D Smallwood and Edward J Sondik. The optimal control of partially observable markov processes over a finite horizon. *Operations research*, 21(5):1071–1088, 1973. 4

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998. 3

Shufan Wang, Guojun Xiong, and Jian Li. Online restless multi-armed bandits with long-term fairness constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15616–15624, 2024. 2, 16

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, May 1992. ISSN 1573-0565. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696. 6, 22

Jiduan Wu, Rediet Abebe, Moritz Hardt, and Ana-Andreea Stoica. Policy design in long-run welfare dynamics. *arXiv preprint arXiv:2503.00632*, 2025. 16

Sierra Wyllie, Ilia Shumailov, and Nicolas Papernot. Fairness feedback loops: training on synthetic data amplifies bias. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2113–2147, 2024. 16

Yuancheng Xu, Chenghao Deng, Yanchao Sun, Ruijie Zheng, Xiyao Wang, Jieyu Zhao, and Furong Huang. Adapting static fairness to sequential decision-making: bias mitigation strategies towards equal long-term benefit rate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024. 5

Xueru Zhang, Mohammadmahdi Khaliligarekani, Cem Tekin, et al. Group retention when using machine learning in sequential decision making: the interplay between user dynamics and fairness. *Advances in neural information processing systems*, 32, 2019. 9

Appendix

Table of Contents

A Notations **15**

B Detailed Related Works **16**

C Reactivity of a Learner **17**

D Pseudo-codes **19**

E Experimental Setting **21**
- E.1 Data Generator . . . 21
- E.2 Eutopia: Details on our proposed Simulator . . . 21
- E.3 Algorithmic Details . . . 22

F Experimental Results **24**
- F.1 Diagnosing Eutopia with Heuristic Policies . . . 24
- F.2 Passive vs. Performative Models of RL . . . 25
- F.3 The Best of Both Worlds: A ADM's Policy for Equity and Profit . . . 35
- F.4 Why a Policy-Gradient Learner: PG versus SAC . . . 37
- F.5 Robustness to the Simulator's Settings . . . 39
- F.6 Trajectories for Lagrangian Multiplier . . . 40

G Computational Resources **40**

A NOTATIONS

Table 3: Notation used throughout the paper.

Symbol	Description
Population and applicants	
N, N_g	Total population size, and size of protected group g
N^t	Number of applicants arriving at step t (cohort size)
$g_i \in \{\text{red}, \text{blue}\}$	Protected group of applicant i (male, female)
$\mathbf{x}_i \in \mathbb{R}^d$, $\mathbf{x}_i[1]$	Feature vector of applicant i, and its first coordinate, the wealth
$l_i \in \mathbb{R}_{>0}$	Loan amount requested by applicant i, fixed across its applications
$d_i \in (0, 1)$	Default probability of applicant i; latent, never observed by the ADM
$\kappa_i = \tau_{\text{gain}} l_i$	Wealth gained by applicant i upon repayment
$c_i \in \{0, 1\}$	Creditworthiness label; used only to evaluate TPR, never observed
$Y_i^t \in \{0, 1\}$	Repayment outcome, $Y_i^t = 1$ on repayment; observed only if approved
Interaction and decisions	
t, T, Δt	Decision step, episode horizon in time units, and time increment
$\mathbf{S}^t = \{\mathbf{x}_i^t, g_i, l_i\}_{i=1}^{N^t}$	State: the context of all applicants arriving at step t
$\mathbf{A}^t \in \{\perp, \not\perp\}^{N^t}$	Action vector; $\perp$ denotes approval and $\not\perp$ denial
$p_i \in [0, 1]$	Approval probability output by the policy for applicant i
π_t, Π	Deployed policy at step t, and the policy set
${p_{\text{apply}}}_i^t$	Probability that applicant i applies at step t
f_ψ	Performative response function governing ${p_{\text{apply}}}_i^t$
$\mathcal{M}(\pi) = (\mathcal{S}, \mathcal{A}, \mathbf{P}^\pi, r^\pi)$	MDP induced by the deployed policy π
$\mathcal{M}(\Pi) = \{\mathcal{M}(\pi) \mid \pi \in \Pi\}$	The PeMDP, i.e. the family of induced MDPs
τ_{interest}, τ_{inv}	Lender's interest rate and applicant's return on investing the loan
$\tau_{\text{gain}} = \tau_{\text{inv}} - \tau_{\text{interest}}$	Net gain rate of a repaid loan, strictly positive
r_π^t, V_π, ζ	Lender's per-step reward, and its average-reward objective, discount factor
Social dynamics	
μ_g^t, $\bar{\mu}^t$	Mean wealth of group g at step t, and population mean wealth
λ_g^t	Hawkes application intensity of group g
$f(\mu_g^t)$	Base arrival rate, increasing in the group's relative wealth (Matthew effect)
α_g, β_g	Excitation strength and decay rate of the Hawkes process
$\mathcal{H}_g$	History of past approvals of group g
Fairness-aware utility	
TPR_g^t	True positive rate: creditworthy applicants of group g approved by step t
C^t	Constraint statistic (Table 1)
SW, RMM, FL	Social Welfare, Rawlsian Max-Min, and Fairness Lagrangian constraints
$\gamma^t \geq 0$, $s = \pm 1$	Lagrange multiplier, learned by dual ascent, and its sign per constraint
$\mathcal{L}(\pi, \gamma)$	Fairness-aware utility optimised by the ADM
MAXUTIL	Profit-only agent, recovered by $\gamma^t \equiv 0$
Evaluation metrics	
$\Delta_\mu^t = \lvert\mu_M^t - \mu_F^t\rvert$	Wealth gap; 0 indicates parity of outcomes
$\rho^t = \frac{\mu_M^t - \mu_M^0}{\mu_F^t - \mu_F^0}$	Inequality ratio; $\rho^t = 1$ means equal growth rates
$\text{CP}^t = \sum_{\tau \leq t} r_\pi^\tau$	Cumulative profit of the lender
Reach rate	Fraction of a group receiving at least one loan in an episode
Gini, Lorenz curve	Within-group concentration of cumulative loans

B Detailed Related Works

The study of fairness in machine learning has evolved from static, snapshot-based metrics toward a dynamic understanding of how algorithms reshape the populations they serve.

Fairness in modern Machine Learning. Early foundational work established formal definitions of algorithmic fairness, including individual fairness (Dwork et al., 2012) and group fairness criteria such as equalized odds and calibration (Hardt et al., 2016). A key theoretical insight from this era is that many natural fairness criteria are mutually exclusive except in degenerate cases (Chouldechova, 2017; Kleinberg et al., 2016). More critically, these definitions treat fairness as a property of a single decision point, ignoring feedback dynamics inherent in repeated deployment. Short-term fairness metrics in lending can reduce the long-term welfare of disadvantaged groups by denying them opportunities to build credit history (Liu et al., 2018) . Similarly, instantaneous fairness constraints create feedback loops: a policy trained on historically biased data may perpetuate exclusion, as the population distribution at time $t+1$ depends on decisions made at time t (Wyllie et al., 2024) . These findings motivate reasoning about fairness across extended temporal horizons.

Fairness in Sequential Decision making. To address temporal dependencies, researchers recast fair decision-making as a sequential problem. Initial work in the bandit setting (Joseph et al., 2016; Nanda et al., 2021) introduced fairness constraints that account for shifting outcome distributions. The natural generalization is to formalise the problem as a Markov Decision Processes (MDPs), where the agent must balance immediate fairness requirements against long-term consequences encoded in state dynamics. Subsequent work developed tractable algorithms: Lyapunov-based methods for high-probability constraint satisfaction (Hu & Zhang, 2022), primal-dual approaches for online fair policy learning (Deng et al., 2022; Wang et al., 2024), and sample-efficient methods for unknown dynamics (Ju et al., 2023). A common assumption across this literature, however, is that transition dynamics are exogenous– the environment evolves according to fixed, policy-independent rules. This assumption breaks down when the population adapts its behaviour in response to the deployed policy (Miller et al., 2020; Estornell et al., 2023), or the underlying state dynamics of the MDP depends on the deployed policy directly (Mandal et al., 2023b).

Fairness in Performative Environment and Long-term Equity. The intersection of long-term fairness and performative prediction (Perdomo et al., 2020) represents the current frontier. Performative prediction recognises that deploying a model changes the distribution it was trained on: a loan approval model affects who applies for loans in future, and a crime recidivism predictor influences defendant behaviour (Dressel & Farid, 2018). Furthermore, a foundational analysis of different policy objectives demonstrates the steady-state distribution of welfare across demographic groups (Wu et al., 2025). It is important to note that the choice of objective has profound implications for long-run inequality. However, pre-existing frameworks do not account for the *performative* shift in transition dynamics that arises when populations respond strategically to deployed policies. Recent theoretical advances have begun to address this gap by establishing conditions for performative stability in discrete MDPs (Mandal et al., 2023a; Rank et al., 2024a), as well as by developing policy gradient methods with convergence guarantees for performative optimality (Basu et al., 2025).

C REACTIVITY OF A LEARNER

We propose a structural classification of reactivity for a learning algorithm operating in a performative environment. Standard analyses of long-term fairness compare policies by their cumulative profit, wealth gap, or inequality ratio, but none of these isolate how a learner lets information from the environment propagate into its decision rule. Reactivity is the missing axis: it asks whether the learner's responsiveness coefficient is itself updated as the environment shifts, or remains pinned to its training-time value.

Model. We model the learner-environment interaction as a coupled linear system with exogenous input Z_t and i.i.d. zero-mean Gaussian noises ξ_t (environment) and ν_t (learner):

$$X_{t+1} = m\,X_t + q\,U_t + \chi\,Z_t + \xi_t, \tag{5}$$
$$U_{t+1} = a\,U_t + b\,X_t + \nu_t, \tag{6}$$

with $|m|, |a| < 1$ for boundedness. X_t denotes the environment state and U_t the learner's decision variable. The coefficients (m, q, χ) of the environment and (a, b) of the learner are specific to this section.

Trajectory unrolling. Substituting (5) into (6) recursively and gathering terms by source yields a three-part decomposition,

$$U_{t+1} \;=\; a\,U_t + q\underbrace{\sum_{s=0}^{t-1} m^{t-1-s} b_s U_s}_{\text{(Past)}} \;+\; \underbrace{\sum_{s=0}^{t-1} m^{t-1-s} b_s \chi_s Z_s \;+\; b_0\,m^t X_0}_{\text{(Data)}} \;+\; \underbrace{\sum_{s=0}^{t-1} m^{t-1-s} b_s \xi_s + \nu_t}_{\text{(Noise)}}. \tag{7}$$

The (Past) term sums contributions from earlier learner states; the (Data) term is the running sum of inputs $\{Z_s\}_{s<t}$ weighted by the per-step responsiveness b_s, plus the propagated initial condition $b_0 m^t X_0$; and finally the (Noise) term collects the noise from both the environment and the learner. *Reactivity is governed by the (Data) term, and specifically by where b_s enters the temporal sum.*

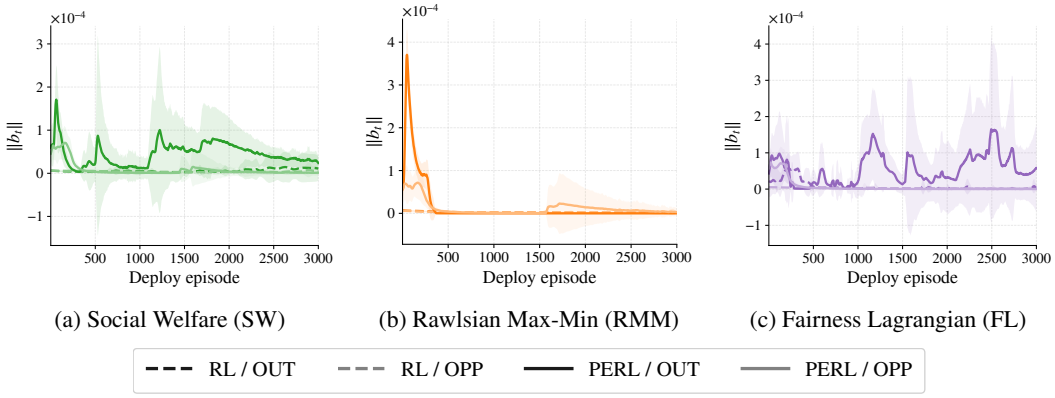
Local sensitivities. The two channels through which information moves are the partial derivatives $\chi_t \triangleq \frac{\partial X_t}{\partial Z_t}$, $b_t \triangleq \frac{\partial U_t}{\partial X_t}$, both allowed to be time-varying in (5)–(6). The qualitative behaviour of the learner is determined by whether b_t remains pinned to its initial value b_0 or evolves with the dynamics.

Definition 1 (Reactivity classes)**.** *Under (5)–(6), a learner is (i) **non-reactive** if $b_t \equiv 0$; (ii) **passive** if $b_t = \Phi(X_t)$ for some t-independent function Φ; (iii) **reactive** if $b_t = \Phi_t(X_t)$ with Φ_t depending genuinely on t.*

Empirical evolution. For each agent we set $b_t = \nabla_\mu \pi_t$, the gradient of the policy's expected approval probability (the Beta mean) with respect to the two group-mean wealths $(\mu^{\text{red}}, \mu^{\text{blue}})$, estimated by auto-grad on the batch of states seen during deploy episode t (same procedure for both agents). We record b_t for the six cells of Table 1 over the $E = 3{,}000$ deploy episodes and 10 seeds of the main experiments. Figures 5 and 6 show the per-episode magnitude $\|b_t\|$ (25-episode moving average) and its running sum $\sum_{s\le t} \|b_s\|$, one panel per fairness constraint and one curve per (agent, perspective). Three observations follow:

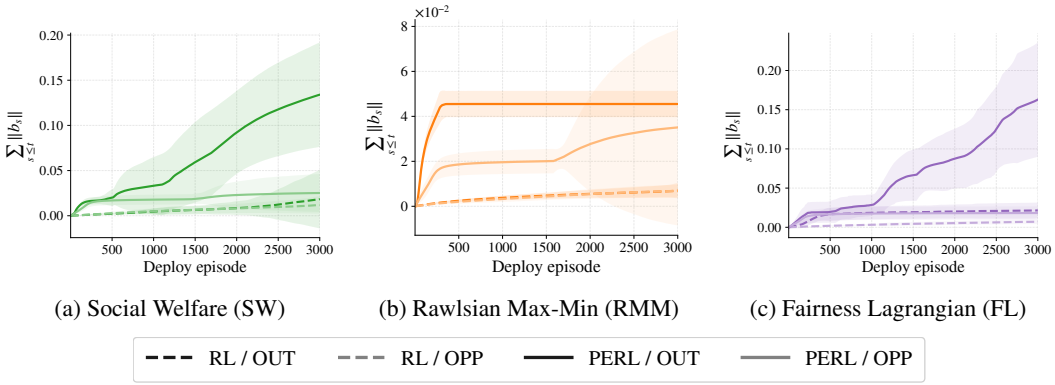
1. PeRL is more reactive than RL in every scenario (Figure 6): 6–7.5$\times$ more under Equality of Outcome and 2–5$\times$ more under Equality of Opportunity. RL's running sum grows slowly and linearly in every cell: its sensitivity to the group wealths stays where training left it, the *passive* class of Definition 1. PeRL, whose gradient carries the dependence of the dynamics on the policy, is the *reactive* learner.

2. PeRL reacts, but fairness perspective matters (Figure 5). Under Equality of Opportunity every constraint shows one burst within the first 200 episodes and is flat afterwards: true-positive rates can be equalised once and the constraints stops binding. FL under Equality of Outcome the constraint acts on a wealth gap that the Matthew dynamics keep re-opening, and PeRL keeps reacting, and shows the largest peak of the study near episode 2,500.



(a) Social Welfare (SW) (b) Rawlsian Max-Min (RMM) (c) Fairness Lagrangian (FL)

Figure 5: **Episodic reactivity.** $\|b_t\|$ over $E = 3{,}000$ deploy episodes, one panel per fairness constraint, RL (dashed) versus PeRL (solid) under Equality of Outcome (OUT, dark) and Equality of Opportunity (OPP, light). Mean over 10 seeds, 25-episode moving average, bands $\pm$ one standard deviation across seeds.



(a) Social Welfare (SW) (b) Rawlsian Max-Min (RMM) (c) Fairness Lagrangian (FL)

Figure 6: **Cumulative reactivity.** Running sum $\sum_{s \le t} \|b_s\|$ over $E = 3{,}000$ deploy episodes; same panels, curves and bands as Figure 5, without smoothing.

3. Reactivity can vanish exactly. For RMM under equality of Outcome every PeRL seed reaches $\|b_t\| = 0$ by episode $\approx 1{,}000$: the policy has converged to the approve-everyone corner, where the guards of Appendix E.3.2 clip the input gradient, and has moved from *reactive* to *non-reactive* in the terms of Definition 1. As approve-everyone is the welfare-optimal fixed point of that scenario (Section E.3.1), this is the end of learning rather than a failure of it; but a flat cumulative curve must be read together with where the policy ends up.

D PSEUDO-CODES

Algorithm 1 gives Eutopia's interaction protocol, shared by every agent and utility in the paper; Equation (8) the per-applicant reward; and Algorithm 2, with Table 4, the dual-ascent update of the multiplier γ^t for the six cells of Table 1. Notation follows Sections 2 and 3; π_θ is the deployed policy π^t, with parameters θ updated once per episode. Apart from wealth $\mathbf{x}_i[1]$, every per-individual quantity (c_i, g_i, l_i, d_i, κ_i) is fixed by the Data Generator (Appendix E.1) and never changes. The ADM's creditworthiness estimate $\hat{c}_i^t$ uses scorer coefficients (ϕ_g, ϕ_X, b) fit once on the initial population, applied to $\tilde{\mathbf{x}}_i^t[1]$, the individual's current wealth standardised by the initial population's mean and standard deviation; since neither is refit, $\hat{c}_i^t$ rises with wealth over deployment. Algorithm 1 is the deployment protocol, which PeRL also follows during pre-training; RL's pre-training environment instead resets wealth, the Hawkes history and the counters at the start of every episode (Appendix E.3.2).

Algorithm 1 Eutopia interaction protocol

Input: groups $g \in \{\text{red}, \text{blue}\}$ of size N_g; per-individual $(\mathbf{x}_i[1], c_i, g_i, l_i, d_i, \kappa_i)$ from the Data Generator; scorer coefficients (ϕ_g, ϕ_X, b); policy π_θ; multiplier γ^t; horizon T, timestep Δt; Hawkes parameters α_g, β_g; Matthew coupling K; application sensitivity s_g; reference population N_{ref}
Initialise $\mathbf{x}_i[1]$ from the Data Generator; $t \leftarrow 0$; $\mathcal{H}_g \leftarrow \emptyset$; $\text{loan}_g, \text{def}_g \leftarrow 0$
for each of the E episodes **do**
  $(\text{tp}_g, \text{fn}_g) \leftarrow 0$; $\mu_g^{\text{start}} \leftarrow \mu_g^t$ *(all other state carries over)*
  for $T/\Delta t$ timesteps **do**
    $\mu_g^t \leftarrow \frac{1}{N_g} \sum_{i \in g} \mathbf{x}_i^t[1]$; $\bar{\mu}^t \leftarrow \frac{1}{2}(\mu_{\text{red}}^t + \mu_{\text{blue}}^t)$; $\bar{d}_g^t \leftarrow \text{def}_g / \max(\text{loan}_g, 1)$
    $\hat{c}_i^t \leftarrow \sigma\big(\phi_g \mathbf{1}[g_i = \text{red}] + \phi_X \tilde{\mathbf{x}}_i^t[1] + b\big)$ for every individual i *(creditworthiness estimate)*
    $f(\mu_g^t) \leftarrow \dfrac{N_g}{N_{\text{ref}}} \max\Big(0.5,\ 2\big(1 + K\,(\mu_g^t/\bar{\mu}^t - 1)\big)\Big)$
    $\lambda_g^t \leftarrow f(\mu_g^t) + \alpha_g \sum_{t_j \in \mathcal{H}_g} e^{-\beta_g (t - t_j)}$ *(Equation (4))*
    ${p_{\text{apply}}}_i^t \leftarrow \text{clip}\Big(\dfrac{\lambda_{g_i}^t \Delta t}{N_{g_i}} \big(1 + s_{g_i}\, \hat{c}_i^t\big),\ 0,\ 1\Big)$ for every individual i
    $\mathcal{N}^t \leftarrow \{\, i : u_i < {p_{\text{apply}}}_i^t,\ u_i \sim \mathcal{U}[0,1] \,\}$ *(the cohort; $|\mathcal{N}^t| = N^t$)*
    for each applicant $i \in \mathcal{N}^t$ **do**
      $o_i^t \leftarrow \big(\mathbf{x}_i^t[1]/100,\ g_i,\ \mu_{\text{red}}^t/100,\ \mu_{\text{blue}}^t/100,\ \lambda_{\text{red}}^t,\ \lambda_{\text{blue}}^t,\ \bar{d}_{\text{red}}^t,\ \bar{d}_{\text{blue}}^t,\ \hat{c}_i^t,\ \phi_g,\ \phi_X,\ l_i\big)$
      $p_i \sim \pi_\theta(\cdot \mid o_i^t)$; $\mathbf{A}_i^t \leftarrow \perp$ with probability p_i, else $\not\perp$
    end for
    $r_i^t \leftarrow$ Equation (8) for each $i \in \mathcal{N}^t$, from the state before this cohort's decisions
    for each applicant $i \in \mathcal{N}^t$ with $\mathbf{A}_i^t = \perp$ **do**
      $Y_i^t \sim \text{Bernoulli}(1 - d_i)$; $\kappa_i^t \leftarrow \kappa_i Y_i^t$; $\mathbf{x}_i[1] \leftarrow \mathbf{x}_i[1] + \kappa_i^t$
      Append t to $\mathcal{H}_{g_i}$; loan_{g_i} += 1; def_{g_i} += $1 - Y_i^t$
    end for
    tp_g += $\#\{i \in \mathcal{N}^t \cap g : c_i = 1,\ \mathbf{A}_i^t = \perp\}$; fn_g += $\#\{i \in \mathcal{N}^t \cap g : c_i = 1,\ \mathbf{A}_i^t = \not\perp\}$
    $t \leftarrow t + \Delta t$
  end for
  $\mu_g^{\text{end}} \leftarrow \mu_g^t$; update θ to increase the discounted return $\sum_t \zeta_{\text{discount}}^t \sum_{i \in \mathcal{N}^t} r_i^t$ over the episode *(Appendix E.3.2)*
  Update γ^t from $\mu_g^{\text{start}}, \mu_g^{\text{end}}, \text{tp}_g, \text{fn}_g$ *(Algorithm 2)*
end for

Reward. Each applicant's reward is its expected profit (Equation (2)) plus the fairness term of its cell of Table 1:

$$r_i^t \;=\; \underbrace{p_i\big[(1 - d_i)\,\tau_{\text{interest}} - d_i\big]\, l_i}_{\text{expected profit}} \;+\; s\,\gamma^t \times \begin{cases} C^t / N^t & \text{Equality of Opportunity,} \\ u_i\; p_i\,(1 - d_i)\,\kappa_i / \bar{\kappa} & \text{Equality of Outcome,} \end{cases} \tag{8}$$

with s as in Equation (3), C^t the Opportunity constraint of Table 1 evaluated on the true positive rates before the cohort's decisions, $\bar{\kappa}$ the population-mean gain on a repaid loan, and $u_i = 1$ (SW),

Table 4: Constraint value C, measured at the end of each episode, and threshold ϑ used by Algorithm 2 for each cell of Table 1. SW and RMM require $C \geq \vartheta$, FL requires $C \leq \vartheta$. $\Delta W_g = N_g(\mu_g^{\text{end}} - \mu_g^{\text{start}})$ is the wealth created in group g during the episode.

	Equality of Outcome		Equality of Opportunity	
Constraint	C	ϑ	C	ϑ
SW	$(\Delta W_{\text{red}} + \Delta W_{\text{blue}})/\big(\bar{\kappa}\,(N_{\text{red}} + N_{\text{blue}})\big)$	0.035	$\text{TPR}_{\text{red}} + \text{TPR}_{\text{blue}}$	1.5
RMM	$\Delta W_{\text{poorer}}/(\bar{\kappa}\,N_{\text{poorer}})$	0.025	$\min(\text{TPR}_{\text{red}}, \text{TPR}_{\text{blue}})$	0.75
FL	$\lvert\mu_{\text{red}} - \mu_{\text{blue}}\rvert/\bar{\mu}$	0.12	$\lvert\text{TPR}_{\text{red}} - \text{TPR}_{\text{blue}}\rvert$	0.10

$\mathbf{1}[\mu_{g_i} < \mu_{\bar{g}_i}]$ (RMM), $\text{sign}(\mu_{g_i} - \mu_{\bar{g}_i})$ (FL), where $\bar{g}_i$ is the other group. Under Equality of Opportunity the constraint enters as its level, split evenly over the N^t applicants of the step, so that each step contributes one copy of C^t. Under Equality of Outcome the level is a function of the group means, which one step's approvals move by about 10^{-5} relative, too little to carry a per-applicant training signal. Each applicant is instead credited with its expected contribution to the change of the constraint: $p_i(1 - d_i)\kappa_i/\bar{\kappa}$ is the wealth the approval is expected to create, and u_i assigns it to the constraint, so that the credits of an episode are, in expectation, proportional to the episode's change in C^t. Dividing by $\bar{\kappa}$ puts γ^t on the profit scale: it reads as the profit the ADM gives up per average loan's worth of borrower wealth.

Algorithm 2 Dual ascent on γ^t, after every episode

Require: current γ^t; the episode's C and threshold ϑ (Table 4); step $\eta = 0.1$; floor $\epsilon = 10^{-4}$; cap $\gamma_{\max} = \max(10, 2\gamma^0)$

1: $v \leftarrow s\,(\vartheta - C)/|\vartheta|$ *(s as in Equation (3); $v > 0$ iff the constraint is violated)*

2: $\gamma^t \leftarrow \text{clip}\big(\gamma^t + \eta\, v,\ \epsilon,\ \gamma_{\max}\big)$

Dual ascent. After every episode, Algorithm 2 raises γ^t while the constraint is violated and lowers it once it is satisfied. Under Equality of Outcome the constraint is measured per episode in normalised form: SW and RMM by the wealth the episode created, per capita and in units of $\bar{\kappa}$, the same quantity the per-applicant credits of Equation (8) add up; FL by the wealth gap relative to mean wealth. Dividing the violation by $|\vartheta|$ gives the step η the same meaning in every cell. The thresholds are set against a coin-flip ADM (approval probability 0.5) in the deployment environment, so that every constraint binds against profit without being out of reach. γ^t starts at $\gamma^0 = 2$ (SW), 5 (RMM) and 0.5 (FL) and is reset to γ^0 at the start of deployment, so that both agents enter deployment with the same multiplier. Formally, Algorithm 2 is projected gradient descent in γ on the Lagrangian of the constraint $sC \geq s\vartheta$; the threshold ϑ does not change the policy's maximiser in Equation (3), which is why the main text omits it.

E EXPERIMENTAL SETTING

E.1 DATA GENERATOR

The Data Generator of Section 3.3 builds the initial population from the UCI Adult Income dataset (Becker & Kohavi, 1996) and fixes the per-individual parameters that govern its response to lending.

Wealth and creditworthiness. A logistic regression predicts $\mathbf{1}[\text{income}_i > \$50\text{K}]$ from age, years of education, weekly working hours, and log capital gains and losses; gender is not among the features. Its output $sc_i \in [0,1]$ sets both the initial wealth, $\mathbf{x}_i[1] = 10 + 190\, sc_i \in [10, 200]$ (in \$K), and the creditworthiness label $c_i = \mathbf{1}[sc_i \geq 0.5]$, which the agent never observes and which is used only to compute true positive rates. Individuals are split by gender into the red (male) and blue (female) groups, with $N_M = N_F = 12{,}000$.

Financial type. Each individual's financial type (l_i, d_i, κ_i) is fixed (Section 3.3). The loan amount is $l_i \sim \mathcal{N}(30, 10^2)$, clipped to $[10, 100]$ (\$K), and the gain on repayment is $\kappa_i = \tau_{\text{gain}}\, l_i = 0.17\, l_i$, with $\tau_{\text{inv}} = 0.35$ and $\tau_{\text{interest}} = 0.18$. The default probability is set by rank: within each group, individuals are ranked by initial wealth and assigned d_i linearly from $\delta_{\max}$ (lowest) to $\delta_{\min}$ (highest), with $[\delta_{\min}, \delta_{\max}] = [0.05, 0.25]$ in deployment. Ranking keeps the mean default probability at $(\delta_{\min} + \delta_{\max})/2$ whatever the shape of the wealth distribution, and ties risk to an observable feature; an independent random d_i would give a policy nothing to learn. Repayment is drawn afresh on every approval, $Y_i^t \sim \text{Bernoulli}(1 - d_i)$.

ADM's scorer. The scorer behind $\hat{c}_i^t$ in Algorithm 1 is a second logistic regression, fit once on the Adult records to predict $\mathbf{1}[\text{income}_i > \$50\text{K}]$ from the group indicator and standardised wealth. Unlike sc_i, it uses group membership, and neither its coefficients (ϕ_g, ϕ_X, b) nor its standardisation are refit during deployment.

E.2 Eutopia: DETAILS ON OUR PROPOSED SIMULATOR

Algorithm 1 gives the protocol; this subsection gives its constants and the reasoning behind its choices.

Time. An episode has horizon $T = 100$ simulated time units at timestep $\Delta t = 0.5$, i.e. 200 decision steps. Simulated time, wealth, the Hawkes history and the loan and default counters carry over from one episode to the next, so a run is a single continual trajectory. Each agent is pre-trained for 1,000 episodes and then deployed for 3,000, continuing to learn throughout deployment; Appendix E.3.2 provides details for the pre-training environments.

Applications. Arrivals follow the Hawkes process of Equation (4): each approval raises its group's intensity by $\alpha_g = 0.3$, decaying at rate $\beta_g = 2.0$ (under 0.1% of its initial value after 3.5 time units), while denials and defaults add nothing. The base rate is

$$f(\mu_g^t) = \frac{N_g}{N_{\text{ref}}} \max\Big(0.5,\ 2\big(1 + K\,(\mu_g^t/\bar{\mu}^t - 1)\big)\Big), \qquad K = 2, \quad N_{\text{ref}} = 3{,}000.$$

It depends on a group's wealth *relative* to the population mean $\bar{\mu}^t$: a base rate proportional to absolute wealth would make wealth grow exponentially, whereas the relative form lets the ratio of group wealths diverge, which is the inequality under study, while absolute wealth stays bounded in rate. The factor N_g/N_{ref} makes arrivals per capita. Without it the cohort size would not grow with N_g, each approval would move μ_g by κ_i/N_g, and the performative drift would vanish as the population grows; with it, a larger population reduces noise without weakening the dynamics. N_{ref} is the population size at which the constants were calibrated. An individual then applies with the probability of Algorithm 1, which scales the group rate by $1 + s_g\, \hat{c}_i^t$ with $s_g = 0.5$ for both groups, so that individuals the ADM is more likely to approve apply more often.

Actions. The policy outputs an approval probability $p_i \in [0,1]$ and the decision is drawn as $\text{Bernoulli}(p_i)$, so the policy's output is continuous and differentiable while the realised decision stays binary. The 12-dimensional observation is o_i^t of Algorithm 1.

E.3 Algorithmic Details

E.3.1 Rule-based Policies

We evaluate seven rule-based policies, summarised in Table 5, as fixed reference points for the learned agents. None of them learns during deployment. Six have no parameters; the seventh, Pattern Prediction, uses the ADM's scorer, whose coefficients are fit once before deployment (Appendix E.1).

Table 5: Rule-based reference policies. p_i is the approval probability for applicant i, d_i its default probability and $\hat{c}_i^t$ the ADM's creditworthiness estimate (Algorithm 1).

Policy	Symbol	Decision rule
Oracle	$\mathbb{O}$	$p_i = \mathbf{1}\left[d_i < \tau_{\text{interest}}/(1+\tau_{\text{interest}})\right]$
Always Approve	$\mathbb{A}_a$	$p_i = 1$
Always Reject	$\mathbb{A}_r$	$p_i = 0$
Uniform Acceptance	$\mathbb{A}_U$	$p_i \sim \mathcal{U}(0,1)$
Rich getting Richer	$\mathbb{A}_{R\uparrow}$	$p_i = 1$ if $g_i = \text{red}$, else $p_i \sim \mathcal{U}(0,1)$
Reverse Rich getting Richer	$\mathbb{A}_{B\uparrow}$	$p_i = 1$ if $g_i = \text{blue}$, else $p_i \sim \mathcal{U}(0,1)$
Pattern Prediction	$\mathbb{A}_H$	$p_i = \sigma(\hat{c}_i^t)$

Oracle ($\mathbb{O}$). The Oracle observes each applicant's true default probability d_i, which no learned agent sees, and approves exactly when the loan is profitable in expectation: $(1-d_i)\,\tau_{\text{interest}}\,l_i - d_i\,l_i > 0$, i.e. $d_i < \tau_{\text{interest}}/(1+\tau_{\text{interest}}) \approx 0.153$. It is the per-decision profit optimum under perfect knowledge of risk. Because this break-even point lies inside the deployment range $[0.05, 0.25]$, the Oracle is selective: it approves the 51% least risky members of each group.

Pattern Prediction ($\mathbb{A}_H$). Pattern Prediction approves with the ADM's own creditworthiness estimate, $p_i = \sigma(\hat{c}_i^t)$, read directly from the observation. The scorer is never refit (Appendix E.1), so the rule is a static learned heuristic: its output moves with the applicant's current wealth and so responds to the population's drift, but it never learns from it.

E.3.2 Learning Agents

RL and PeRL share the policy network, optimiser, discount, entropy bonus, gradient clipping and dual ascent (Table 6), and differ in how the policy gradient is estimated. Both are deployed under identical conditions.

Policy. $\boldsymbol{\pi}_\theta(\cdot \mid o_i^t)$ is a Beta distribution over the approval probability p_i, whose two concentration parameters are the two heads of an MLP with two hidden layers of width 128 and ReLU activations; its input is the observation o_i^t of Algorithm 1. The concentrations are bounded above and below to keep the policy away from degenerate approve-all and reject-all solutions. Both agents use AdamW with learning rate 10^{-3}, weight decay 0.1 on the heads, gradient clipping at unit norm, and an entropy bonus of 10^{-3}.

RL. The RL agent is REINFORCE (Williams, 1992). Per-applicant rewards (Equation (8)) are summed over each timestep's cohort and normalised online by their running mean and standard deviation. The return-to-go is discounted per timestep with ζ_{discount}, and a baseline, an exponential moving average of past returns-to-go at the same timestep index, is subtracted; every applicant in a cohort receives its cohort's advantage. RL treats the transition as fixed and follows the score-function gradient.

PeRL. The PeRL agent is Performative Policy Gradient (Basu et al., 2025), which differentiates through the population's response to the policy. Each episode it runs, from the same starting state as the simulator, a differentiable mean-field shadow rollout of the group means μ_g and Hawkes intensities λ_g: at every step, a batch of individuals drawn from the current population passes through the policy with a reparameterised Beta sample, and their expected outcomes update μ_g, λ_g and the

reward of Equation (8) with the current γ^t. Backpropagating through the unrolled episode credits each decision with its effect on every later state and reward, which is the performative term absent from the RL gradient. The shadow rollout only supplies the gradient; the simulator itself advances the population and produces every reported metric, exactly as for RL.

Table 6: Hyperparameters, shared by RL and PeRL agents.

Hyperparameter	Value	Note
Hidden layers	2×128	ReLU
Policy	Beta	Over the approval probability p_i
Optimiser	AdamW	Weight decay 0.1 on the heads
Learning rate	10^{-3}	
Discount ζ_{discount}	0.99	Per timestep
Entropy coefficient	10^{-3}	
Gradient clipping	1.0	Global norm
Baseline decay (RL)	0.9	Per timestep index
Dual-ascent step η	0.1	Algorithm 2
γ^0	2 / 5 / 0.5	SW / RMM / FL; reset at deployment
γ^t floor, cap	10^{-4}, $\max(10, 2\gamma^0)$	Algorithm 2
Episodes	1,000 + 3,000	Pre-training + deployment
Seeds	10	Per experiment

F Experimental Results

F.1 Diagnosing Eutopia with Heuristic Policies

We deploy the seven rule-based policies of Appendix E.3.1 for 3,000 episodes and 10 seeds each, under the learned agents' deployment setting. None of them adapts, so what happens to the population is the simulator's response to a fixed rule. Figure 7 reports the wealth gap, inequality ratio and cumulative profit of Section 4.1, and the blue group's welfare $R_F^t = (\mu_F^t - \mu_F^0)/t$, with PeRL under FL and Equality of Outcome (Table 2) overlaid for reference. Seed variation of the rules is negligible at this population size, so we quote means.

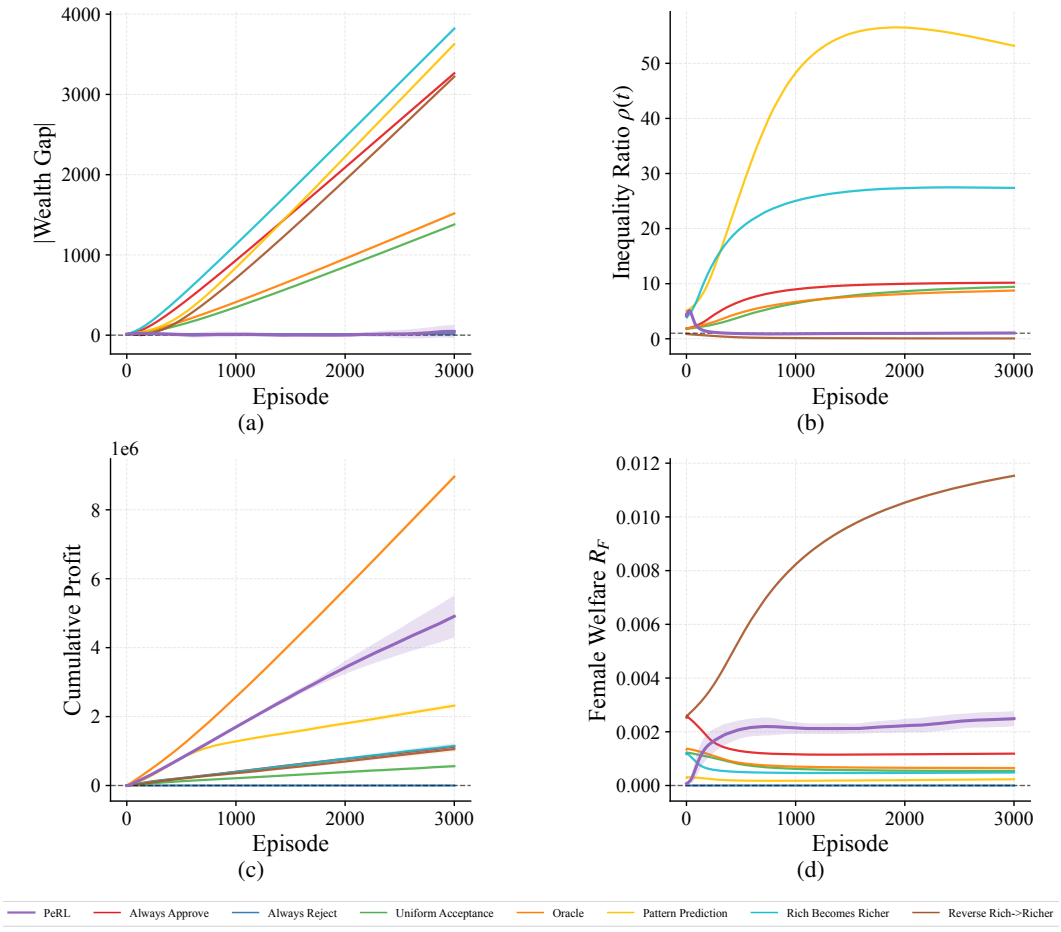


Figure 7: **Diagnosing Eutopia with rule-based policies.** (a) wealth gap Δ_μ^t, (b) inequality ratio ρ^t, (c) cumulative profit CP^t, (d) blue-group welfare R_F^t, over 3,000 deployment episodes, mean over 10 seeds. PeRL under FL and Equality of Outcome is overlaid for reference; its band is ± 1 std across seeds.

Wealth gap. Under Always Reject no loan is issued and the gap stays at its initial 13.2. Every other rule widens it without saturating. The biased rules reach the largest gaps, 3,820 for Rich Becomes Richer and 3,221 in the opposite direction for Reverse Rich Becomes Richer. The group-blind rules widen it too: Always Approve to 3,263, Oracle to 1,517, Uniform Acceptance to 1,379. This is the Matthew effect of the arrival process (Section 3.3): the red group starts wealthier, applies more, receives more loans and pulls further ahead, and a fixed rule has nothing that counteracts it. Pattern Prediction ends at 3,627. PeRL's gap never exceeds 46.5, below every rule except Always Reject.

Inequality ratio. The group-blind rules drift to $\rho \approx 9$–10 (Always Approve 10.2, Uniform Acceptance 9.4, Oracle 8.8): the red group's wealth grows about nine times faster although the ADM ignores group membership. Rich Becomes Richer reaches 27 and Reverse Rich Becomes Richer 0.06. The two are mirror images in what the ADM does (approval-rate disparity ± 0.50) but not in outcome, because the blue group also starts poorer and faces the arrival process working against it. Pattern Prediction peaks at 57 around episode $1{,}900$ and ends at 53. PeRL settles at 1.05 after dipping to 0.87 around episode 770. ρ is undefined under Always Reject.

Profit. Oracle earns the most, 8.96M, since it alone sees each applicant's default risk. Its break-even point $d_i < 0.153$ is close to the mean default probability of 0.15, so approving everyone is only marginally profitable: Always Approve earns 1.13M, the two biased rules 1.15M and 1.06M, Uniform Acceptance 0.56M. Pattern Prediction reaches 2.32M, starting well but flattening after about 700 episodes as the population drifts away from the one its scorer was fit on. PeRL earns 4.91 ± 0.58M, second only to Oracle.

Female welfare. Under every group-blind rule R_F^t falls over deployment, from 0.0022 at episode 100 to 0.0012 under Always Approve and to about 0.0006 under Oracle and Uniform Acceptance: as the red group pulls ahead, the blue group applies less and receives a shrinking share of the loans. It keeps rising only under Reverse Rich Becomes Richer, to 0.0115, at the cost of the inverted inequality above. Under PeRL it rises to 0.0022 by episode 600 and 0.0025 by the end, twice Always Approve, with ρ near 1.

Pattern Prediction. The scorer-based rule is the closest analogue to a deployed credit model, and it is the least fair rule without an explicit bias: approval-rate disparity 0.72, the second-largest wealth gap and ρ above 50. Its group coefficient ϕ_g encodes the correlation of gender and income in the Adult data, so it approves red applicants at a far higher rate; the resulting loans raise red wealth and hence red scores, while blue scores rise more slowly. A static learned rule amplifies the drift it induces.

What the diagnostics establish. The rules behave as their definitions predict: nothing moves under Always Reject, the biased rules bias, and Oracle is the most profitable. Beyond that, even group-blind lending turns the initial gap into a nine-fold difference in growth rates, so equal treatment does not yield equal outcomes, and the most profitable rule ends at $\rho \approx 9$. PeRL is the only policy in the figure that keeps ρ near 1 and the gap small, and it earns more than every rule except Oracle.

F.2 Passive vs. Performative Models of RL

No fixed rule keeps profit and fairness together once the population responds to its decisions (Appendix F.1). We now ask whether a learning agent can, and whether it matters that the agent models that response. We compare RL, which follows the ordinary policy gradient and treats the population as fixed, with PeRL, which differentiates through the population's response to its own policy (Appendix E.3.2). Each is pre-trained for $1{,}000$ episodes in its own environment, static for RL and performative for PeRL, and both are deployed identically for $3{,}000$ episodes in all six cells of Table 1, with 10 seeds. The tables abbreviate Equality of Outcome and Equality of Opportunity as OUT and OPP. We report the macro metrics of Section 4 and the approval-rate disparity at the end of deployment, then follow each metric over deployment. Throughout this section the multiplier is learned by dual ascent; the fixed-multiplier variant of PeRL under FL and Equality of Outcome reported in Table 2 is examined in Appendix F.3.

F.2.1 PeRL versus RL

Table 7 places the two agents side by side in every cell. Tables 8 and 9 give, for each agent, the best constraint under each perspective, and Tables 10 and 11 the best perspective under each constraint.

PeRL is better wherever the objective trades profit against fairness. PeRL earns more, leaves a smaller wealth gap and keeps ρ closer to one than RL in all six cells: twenty of the twenty-four comparisons in Table 7. Under Equality of Opportunity it earns six to seven times RL's profit and

Table 7: **RL vs. PeRL across all six cells** at the end of deployment (mean ± std over 10 seeds). For each metric the better of the two agents is in bold: higher profit, smaller wealth gap, smaller approval-rate disparity, ρ closer to 1. †: RL approves nearly every applicant of both groups in this cell, so the disparity is small because approval is near-universal, not because the groups are balanced.

Perspective	Constraint	Cumulative Profit ($\times 10^6$)		|Wealth Gap|		|Approval Rate Disparity|		Inequality Ratio ρ	
		RL	PeRL	RL	PeRL	RL	PeRL	RL	PeRL
OUT	FL	0.374 ± 0.385	$\mathbf{1.304 \pm 0.974}$	745.0 ± 1288.3	$\mathbf{71.9 \pm 298.0}$	0.0135 ± 0.5070	$\mathbf{0.0119 \pm 0.2550}$	0.190 ± 0.132	$\mathbf{1.059 \pm 0.094}$
OUT	RMM	1.096 ± 0.096	$\mathbf{1.267 \pm 0.058}$	3190.0 ± 68.9	$\mathbf{3081.9 \pm 26.5}$	$\mathbf{0.0024 \pm 0.0202}^{\dagger}$	0.1086 ± 0.0030	10.198 ± 0.032	$\mathbf{0.101 \pm 0.000}$
OUT	SW	1.103 ± 0.071	$\mathbf{4.721 \pm 1.346}$	3169.8 ± 82.5	$\mathbf{1956.0 \pm 1012.6}$	$\mathbf{0.0088 \pm 0.0481}$	0.2081 ± 0.1897	10.294 ± 0.082	$\mathbf{0.138 \pm 0.048}$
OPP	FL	1.131 ± 0.062	$\mathbf{7.443 \pm 0.410}$	3154.8 ± 103.9	$\mathbf{1086.8 \pm 332.2}$	0.0109 ± 0.0459	$\mathbf{0.0097 \pm 0.0129}$	10.328 ± 0.045	$\mathbf{3.981 \pm 0.568}$
OPP	RMM	1.105 ± 0.100	$\mathbf{6.711 \pm 1.229}$	3177.3 ± 55.4	$\mathbf{207.9 \pm 1328.7}$	$\mathbf{0.0049 \pm 0.0257}$	0.0543 ± 0.0792	10.231 ± 0.043	$\mathbf{1.229 \pm 0.643}$
OPP	SW	1.118 ± 0.099	$\mathbf{6.757 \pm 1.137}$	3115.9 ± 136.1	$\mathbf{535.2 \pm 1187.0}$	$\mathbf{0.0085 \pm 0.0556}$	0.0427 ± 0.0734	10.304 ± 0.079	$\mathbf{1.776 \pm 0.878}$
PeRL better in		6/6		6/6		2/6		6/6	

Table 8: **RL: best fairness constraint under each fairness perspective**, per metric, at the end of deployment (mean ± std over 10 seeds; ρ is the ratio of seed-mean growths, std by bootstrap). Bold: best value in the table for that metric. †: RL approves nearly every applicant of both groups in this cell, so the disparity is small because approval is near-universal, not because the groups are balanced.

Perspective	Cumulative Profit ($\times 10^6$)		|Wealth Gap|		|Approval Rate Disparity|		Inequality Ratio ρ	
	Best	Mean ± Std	Best	Mean ± Std	Best	Mean ± Std	Best	Mean ± Std
OUT	SW	1.103 ± 0.071	FL	$\mathbf{745.0 \pm 1288.3}$	RMM	$\mathbf{0.0024 \pm 0.0202}^{\dagger}$	FL	$\mathbf{0.190 \pm 0.132}$
OPP	FL	$\mathbf{1.131 \pm 0.062}$	SW	3115.9 ± 136.1	RMM	0.0049 ± 0.0257	RMM	10.231 ± 0.043

Table 9: **PeRL: best fairness constraint under each fairness perspective**, per metric, at the end of deployment (mean ± std over 10 seeds; ρ is the ratio of seed-mean growths, std by bootstrap). Bold: best value in the table for that metric.

Perspective	Cumulative Profit ($\times 10^6$)		|Wealth Gap|		|Approval Rate Disparity|		Inequality Ratio ρ	
	Best	Mean ± Std	Best	Mean ± Std	Best	Mean ± Std	Best	Mean ± Std
OUT	SW	4.721 ± 1.346	FL	$\mathbf{71.9 \pm 298.0}$	FL	0.0119 ± 0.2550	FL	$\mathbf{1.059 \pm 0.094}$
OPP	FL	$\mathbf{7.443 \pm 0.410}$	RMM	207.9 ± 1328.7	FL	$\mathbf{0.0097 \pm 0.0129}$	RMM	1.229 ± 0.643

cuts the wealth gap by a factor of three to fifteen; under FL with Equality of Outcome it reaches a gap of 72 against RL's 745, with $\rho = 1.06$ against 0.19. The four comparisons RL wins are all on approval-rate disparity, and are explained next.

RL tends to approve most applicants. In five of the six cells RL ends at nearly the same point: profit 1.10 to 1.13M, wealth gap 3,116 to 3,190, ρ between 10.2 and 10.3, and approval-rate disparity below 0.011. These are, within seed noise, the numbers of Always Approve (1.13M, 3,263, 10.2, 0; Appendix F.1). Under RMM and SW in both perspectives, and under FL with Equality of Opportunity, RL approves nearly every applicant of both groups: its disparities are small because approval is near-universal, and the Matthew effect drives the gap and ρ to the values of the approve-all rule. Under FL with Equality of Outcome, the gap penalty is strong enough to move RL, which then under-lends to the red group, so that the blue group grows five times faster ($\rho = 0.19$).

PeRL's results depend on the utility. PeRL's profit ranges from 1.27M to 7.44M and its wealth gap from 72 to 3,082 across the six cells, as expected of an agent that responds to each constraint. Tables 9 and 11 summarise this: FL under Equality of Outcome gives the smallest gap and the ρ closest to one, FL under Equality of Opportunity the highest profit and smallest disparity, and for RMM and SW the Opportunity perspective is better on every metric. Appendix F.2.2 follows each metric over deployment.

F.2.2 How PeRL Trades: Metric by Metric

Table 7 reports end points; this section follows each metric over deployment for PeRL (Figures 8 and 10; RL in Figures 9 and 11), with the three constraints overlaid in each panel. Two patterns recur. Every PeRL cell begins deployment by favouring the blue group: γ^t is reset to γ^0 at deployment, and early on the fairness term outweighs profit. What follows depends on whether the constraint pushes back once the blue group catches up: FL penalises the gap in either direction, while RMM

Table 10: **RL: best fairness perspective under each fairness constraint**, per metric, at the end of deployment (mean ± std over 10 seeds). Bold: best value in the table for that metric. †: RL approves nearly every applicant of both groups in this cell, so the disparity is small because approval is near-universal, not because the groups are balanced.

Constraint	Cumulative Profit ($\times 10^6$)		|Wealth Gap|		|Approval Rate Disparity|		Inequality Ratio ρ	
	Best	Mean ± Std	Best	Mean ± Std	Best	Mean ± Std	Best	Mean ± Std
FL	OPP	$\mathbf{1.131 \pm 0.062}$	OUT	$\mathbf{745.0 \pm 1288.3}$	OPP	0.0109 ± 0.0459	OUT	$\mathbf{0.190 \pm 0.132}$
RMM	OPP	1.105 ± 0.100	OPP	3177.3 ± 55.4	OUT	$\mathbf{0.0024 \pm 0.0202}^{\dagger}$	OUT	10.198 ± 0.032
SW	OPP	1.118 ± 0.099	OPP	3115.9 ± 136.1	OPP	0.0085 ± 0.0556	OUT	10.294 ± 0.082

Table 11: **PeRL: best fairness perspective under each fairness constraint**, per metric, at the end of deployment (mean ± std over 10 seeds). Bold: best value in the table for that metric.

Constraint	Cumulative Profit ($\times 10^6$)		|Wealth Gap|		|Approval Rate Disparity|		Inequality Ratio ρ	
	Best	Mean ± Std	Best	Mean ± Std	Best	Mean ± Std	Best	Mean ± Std
FL	OPP	$\mathbf{7.443 \pm 0.410}$	OUT	$\mathbf{71.9 \pm 298.0}$	OPP	$\mathbf{0.0097 \pm 0.0129}$	OUT	$\mathbf{1.059 \pm 0.094}$
RMM	OPP	6.711 ± 1.229	OPP	207.9 ± 1328.7	OPP	0.0543 ± 0.0792	OPP	1.229 ± 0.643
SW	OPP	6.757 ± 1.137	OPP	535.2 ± 1187.0	OPP	0.0427 ± 0.0734	OPP	1.776 ± 0.878

and SW do not, and under Equality of Outcome the arrival dynamics (Section 3.3) then amplify whichever group is ahead.

Wealth gap. Under Equality of Outcome the constraints separate completely (Figure 8a). FL keeps $|\Delta_\mu^t|$ below 100 throughout: the gap closes by episode 170, opens in the blue group's favour to 50 around episode 600, closes again at episode 1,150, and ends at 72 with the red group ahead, after peaking at 82. RMM and SW open the gap in the blue group's favour and never close it: RMM to 3,082 with negligible seed variation, close to the magnitude RL's approve-all behaviour produces in the other direction, and SW to 1,956 (±1,013). Under Equality of Opportunity (Figure 10a) the gap stays below 150 for the first 900 (FL), 1,700 (SW) and 2,700 (RMM) episodes, and then grows in the red group's favour, to 1,087 for FL, 535 for SW and 208 for RMM. The RMM and SW bands are the widest in the figure (±1,329 and ±1,187), because seeds differ in which group ends up ahead. RL follows the approve-all diagonal in five cells; under FL with Equality of Outcome it opens a gap of 745 in the blue group's favour by under-lending to the red group.

Inequality ratio. ρ^t makes the direction explicit (Figures 8b and 10b; values in the first few episodes are ratios of very small changes). Under Equality of Outcome, FL overshoots and recovers: it falls to 0.49 by episode 200, crosses 1 at episode 1,235, peaks at 1.12 near episode 1,750 and ends at 1.06, pulled back by the two-sided gap penalty. RMM and SW have no such pull: RMM's ρ peaks at 0.29 and decays to 0.10, SW's peaks at 0.66 around episode 150 and decays to 0.14, so the blue group grows seven to ten times faster, a mirror image of the fixed rules (Appendix F.1). Under Equality of Opportunity all three cross 1 and keep rising: FL at episode 220, to 3.98; SW at 530, to 1.78; RMM at 1,730, to 1.23. Equalising true positive rates constrains who is approved among the creditworthy, not how much wealth each group accumulates, so the Matthew effect reopens the growth-rate gap once the constraint is met; the blue group's welfare R_F^t (not shown) peaks between episodes 450 and 1,800 and then declines as its application rate falls.

Profit. Profit separates by perspective (Figures 8c and 10c). Under Equality of Opportunity all three constraints earn 6.7 to 7.4M, six to seven times RL, at a nearly constant rate (about 1.0M by episode 500 and 2.0 to 2.2M by 1,000): approving the creditworthy of both groups at similar rates is close to what a profit-maximising policy does. Under Equality of Outcome FL and RMM earn 1.30M (±0.97M) and 1.27M, a fifth of their Opportunity counterparts, at a near-linear rate. SW under Outcome earns 4.72M on a convex curve (0.53M at episode 500, 1.08M at 1,000, 2.57M at 2,000), accelerating as the blue group's wealth, and with it its application rate, grows.

Approval-rate disparity. The cumulative disparity $\mathrm{app}_{\mathrm{red}} - \mathrm{app}_{\mathrm{blue}}$ (Figures 8d and 10d; negative when the blue group is approved at the higher rate) shows the early tilt: at episode 100 every PeRL cell favours the blue group, by 0.17 (FL) to 0.69 (RMM) under Outcome and by 0.34 to 0.42 under Opportunity. Under Opportunity it then decays to between 0.01 and 0.05 as the true posi-

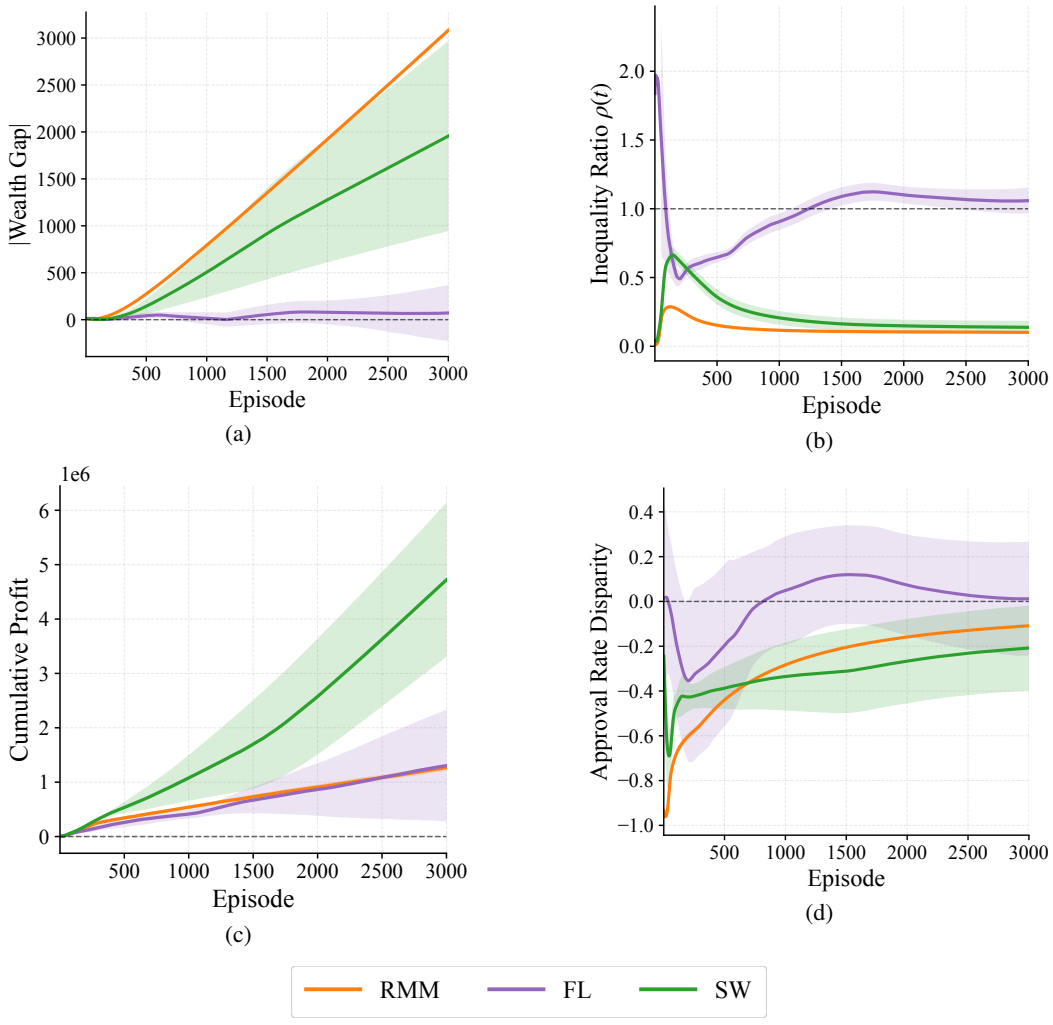


Figure 8: **PeRL: Equality of Outcome (OUT).** (a) wealth gap, (b) inequality ratio, (c) cumulative profit, (d) approval-rate disparity (red minus blue), for PeRL under FL, RMM and SW with the Equality of Outcome perspective, over 3,000 deployment episodes. Mean ± std across 10 seeds.(Section 4).

tive rates equalise and γ^t falls. Under Outcome, FL reaches -0.35 around episode 200 and $+0.12$ around episode 1,500 before ending at 0.01, the same restoring pattern as its gap, while RMM and SW end at -0.11 and -0.21. RL's disparities stay within 0.01 of zero in its approve-all cells; under FL with Outcome it starts at -0.80 and comes within 0.01 of zero only after about 1,650 episodes.

Summary. Under Equality of Outcome the constraint sets the profile: FL closes the gap and keeps ρ near 1, RMM and SW let the blue group overtake, and only SW turns that into profit. Under Equality of Opportunity the perspective sets it: all three constraints earn the same high profit and differ mainly on the wealth axes, where RMM and SW end closest to balance and FL furthest from it. A ADM prioritising the wealth gap is served by FL under Outcome, at a cost in profit; one prioritising profit, by any constraint under Opportunity, at a cost in wealth equality. Appendix F.2.3 compares the six profiles, and Appendix F.3 examines a single policy that delivers both.

F.2.3 Macro-analysis: Pareto Frontiers of Utilities

Figure 12 folds the four metrics of Table 7 into one profile per PeRL cell; RL is omitted, since its five approve-all cells give nearly the same polygon. Each axis is normalised over the six cells, so the rim is the best value any cell reached and the centre the worst, and the lower-is-better metrics are inverted so that outward is always better. Two cautions apply. The scale is relative to these six

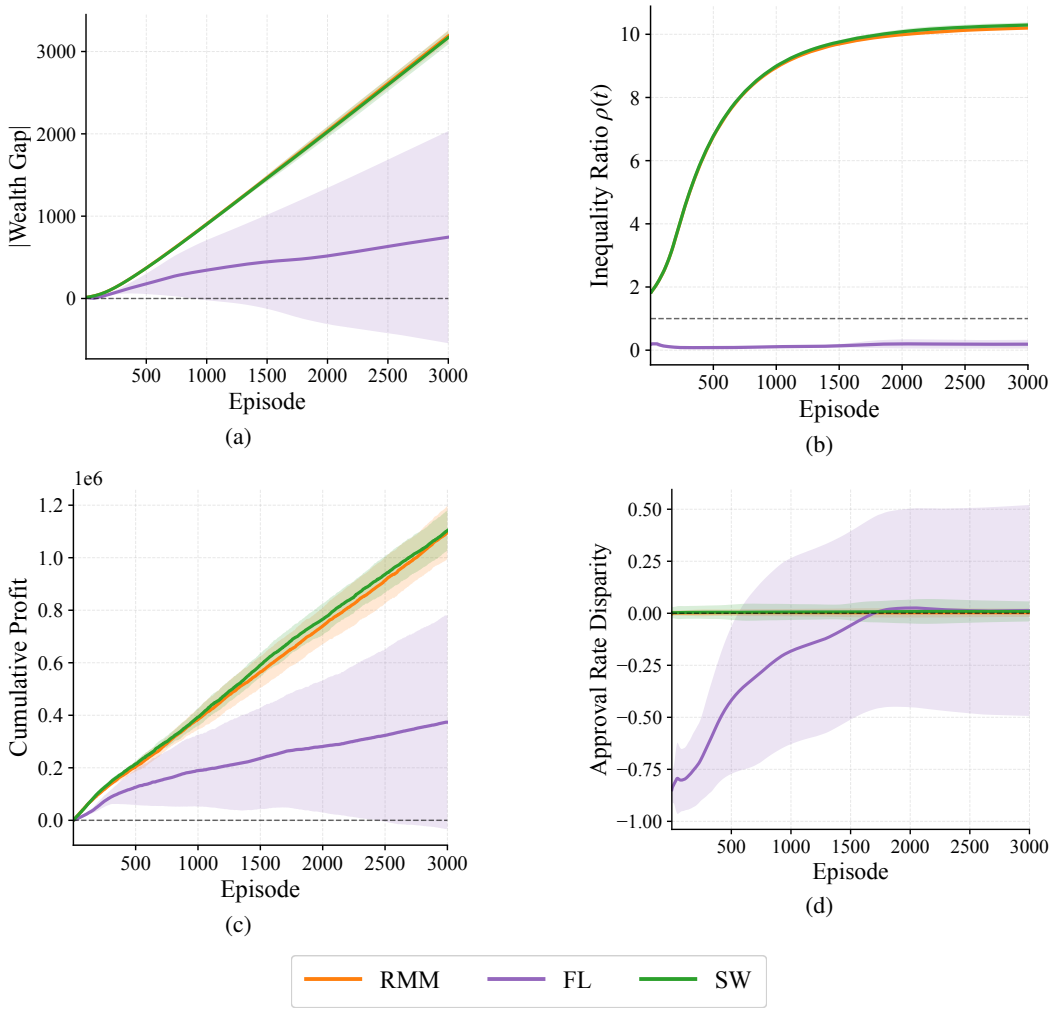


Figure 9: **RL: Equality of Outcome (OUT).** (a) wealth gap, (b) inequality ratio, (c) cumulative profit, (d) approval-rate disparity (red minus blue), for RL under FL, RMM and SW with the Equality of Outcome perspective, over 3,000 deployment episodes. Mean ± std across 10 seeds. (Section 4).

cells, not absolute. And the inequality axis uses the linear distance $|\rho - 1|$, so $\rho = 0.10$ (the blue group growing ten times faster) sits at 0.70 while $\rho = 3.98$ sits at the centre: the axis compresses inequality in the blue group's favour.

Equality of Outcome. The three polygons do not overlap (Figure 12a). FL reaches the rim on all three fairness axes (wealth gap 0.98, inequality 0.98, disparity 0.94) and the centre on profit (0.18, 1.30M): it is the fairest cell in the grid on every fairness measure. RMM sits at the centre on wealth gap (0.04, the largest gap of the six cells, 3,082) and at 0.48 on disparity, with a profit no higher than FL's; its 0.70 on inequality is the compression just described. SW is the only Outcome cell with a substantial profit vertex (0.63, 4.72M), but it is the worst cell in the grid on disparity (0.04, a cumulative disparity of 0.21 in the blue group's favour). Under Outcome, then, the choice is between FL's fairness and SW's profit.

Equality of Opportunity. All three polygons reach the rim on profit (0.90 to 1.00, i.e. 6.7M to 7.4M; Figure 12b), because equalising true positive rates costs the ADM little (Appendix F.2.2); the wealth axes separate them. RMM has the largest and most even polygon in either chart: wealth gap 0.93, profit 0.90, inequality 0.92, disparity 0.74, i.e. a gap of 208, 6.71M, $\rho = 1.23$ and a disparity of 0.054. SW has a slightly smaller version of the same shape (0.83, 0.91, 0.74, 0.79). FL is the most profitable cell (1.00) and nearly the best on disparity (0.95), but sits at the centre on inequality: $\rho = 3.98$, the largest inequality in the red group's favour of any cell. Under Opportunity

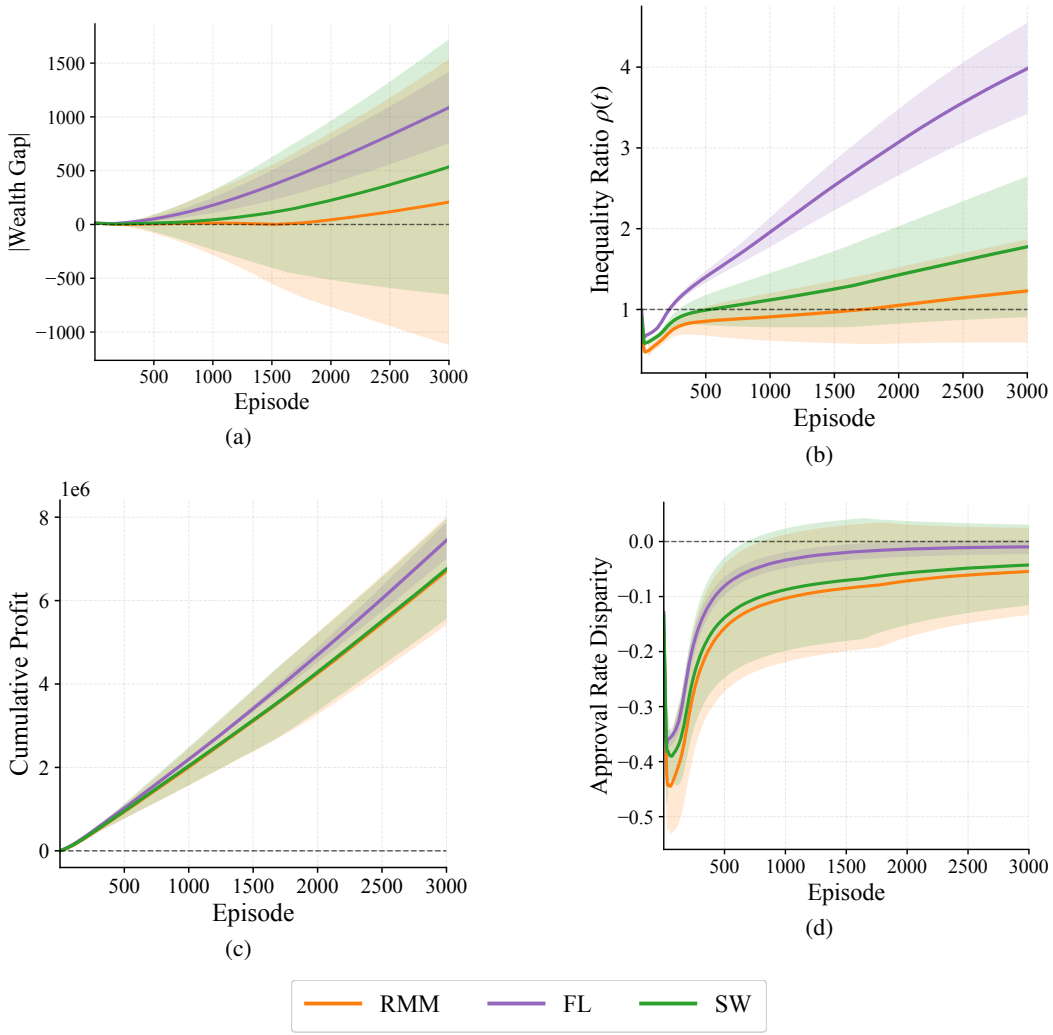


Figure 10: **PeRL: Equality of Opportunity (OPP).** (a) wealth gap, (b) inequality ratio, (c) cumulative profit, (d) approval-rate disparity (red minus blue), for PeRL under FL, RMM and SW with the Equality of Opportunity perspective, over 3,000 deployment episodes. Mean ± std across 10 seeds.(Section 4).

the perspective secures profit and disparity, and the constraint decides the wealth axes: RMM and SW cover them, FL does not.

The frontier. Treating the six cells as candidate policies over the four metrics, two are dominated: RMM and SW under Outcome are each worse than the same constraint under Opportunity on all four. The other four are undominated: FL under Opportunity for profit; SW under Opportunity for a slightly higher profit and lower disparity than RMM, at a wider gap; RMM under Opportunity for the wealth axes among the Opportunity cells; and FL under Outcome for the best wealth gap, inequality ratio and disparity of any cell. Since our objective is the fairness outcome, with profit a constraint rather than a target, FL under Outcome is the reference point. It keeps the two groups growing at the same rate, holds the gap near zero and treats the groups alike at the point of decision, and its ρ has settled by the end of deployment, whereas RMM under Opportunity, its nearest competitor on the wealth axes, crosses 1 only around episode 1,730 and is still rising. What FL under Outcome lacks is profit, a fifth of the Opportunity cells'. The frontier thus has FL at both ends: the same constraint gives the most profitable cell under Opportunity (7.44M) and the fairest under Outcome (1.30M), at opposite corners of the radar. The two cells differ only in what the constraint is applied to and in how the multiplier evolves during deployment (Appendix F.6). Appendix F.3 therefore keeps the FL constraint and the Outcome perspective but fixes the multiplier, to see whether one policy can take the fairness of the first cell and the profit of the second.

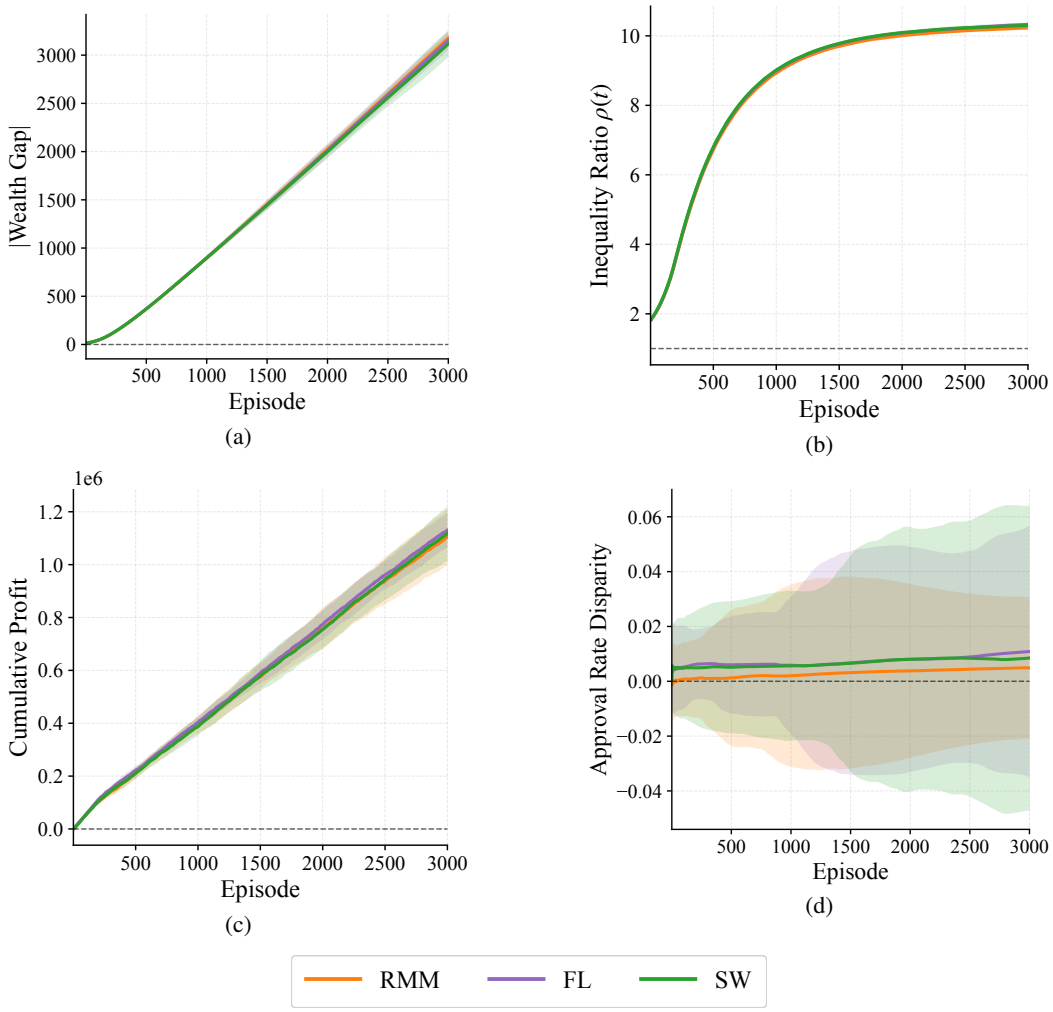


Figure 11: **RL: Equality of Opportunity (OPP).** (a) wealth gap, (b) inequality ratio, (c) cumulative profit, (d) approval-rate disparity (red minus blue), for RL under FL, RMM and SW with the Equality of Opportunity perspective, over 3,000 deployment episodes. Mean $\pm$ std across 10 seeds. (Section 4).

F.2.4 Micro-analysis: Who Receives the Loans

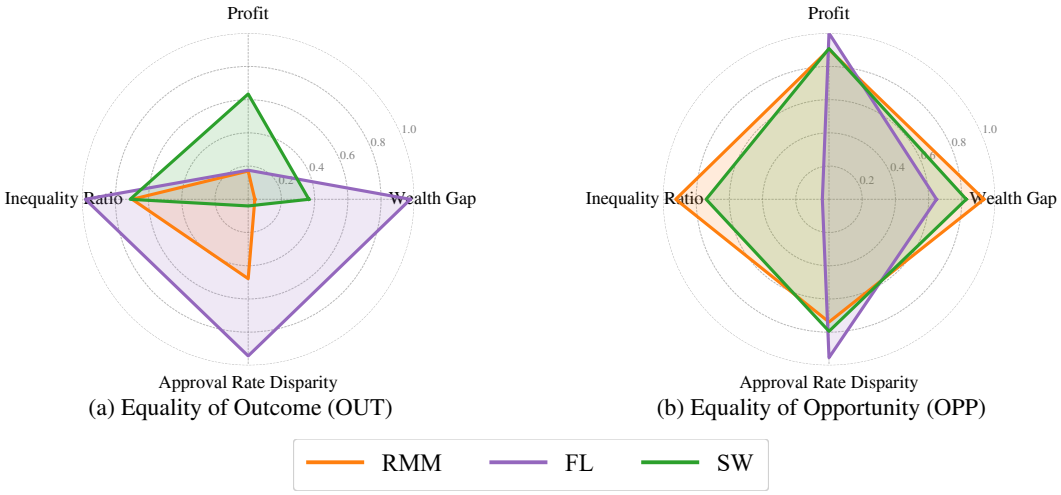
A wealth gap of zero says that the two groups' mean wealth coincide, not how the loans behind it were spread within each group. For each PeRL cell we record how many loans each of the 12,000 individuals per group received, and read that record in three ways:

- **Lorenz curves** (Figure 4, also expanded in Figure 13), at episodes 100, 1,000, 2,000 and 3,000: the cumulative share of a group's loans held by its least-served members. The diagonal is equal shares; the further a curve sags below it, the more concentrated the group's credit, and a flat stretch on the left is the part of the group that received almost nothing. The Gini coefficient summarises a curve in one number, 0 for equal shares.

- **Reach rate** (Figure 14): the share of a group that receives a loan in each episode, i.e. the breadth of access over time.

- **Loan distributions** (Figures 15 and 16): total loans per individual at the end of deployment, from the most to the least served, against the equal-share line.

Table 12 collects the corresponding end-of-deployment numbers.



(a) Equality of Outcome (OUT)

(b) Equality of Opportunity (OPP)

Figure 12: **PeRL: performance profiles.** End-of-deployment values of the four metrics of Table 7 for PeRL under FL, RMM and SW, one chart per fairness perspective. Each axis is normalised over all six cells, with 1 the best value observed and 0 the worst, and oriented so that outward is better (profit as is; wealth gap, $|\rho - 1|$ and approval-rate disparity inverted). The two charts share one scale, so polygons are comparable across them. Larger and more even is better.

Table 12: **Who receives the loans under PeRL, at the end of deployment.** Loans per individual; share of the group never served; share of the group's loans held by its most-served 10% (0.10 means equal shares); Gini coefficient (0 means equal shares); and reach rate over the last 100 episodes. R: red (male), B: blue (female). Means over 10 seeds.

		Loans / individual		Never served		Top-10% share		Gini		Reach rate	
Persp.	Constr.	R	B	R	B	R	B	R	B	R	B
OUT	FL	243	227	0%	0%	0.12	0.13	0.08	0.09	0.086	0.074
OUT	RMM	80	790	0%	0%	0.13	0.11	0.10	0.02	0.028	0.257
OUT	SW	71	512	6%	0%	0.21	0.15	0.47	0.22	0.020	0.152
OPP	FL	318	80	24%	21%	0.18	0.20	0.43	0.44	0.120	0.018
OPP	RMM	231	191	26%	15%	0.19	0.18	0.44	0.40	0.089	0.051
OPP	SW	265	152	25%	17%	0.19	0.18	0.44	0.40	0.103	0.035

Equality of Outcome. The three Outcome cells reach everyone: no individual goes without a loan under FL or RMM, and only 6% of the red group does under SW (Table 12). They differ in how the loans are shared.

Under FL the two groups end alike on every count. The Lorenz curves (Figure 4a) start far from the diagonal at episode 100 (Gini 0.73 for the red group, 0.58 for the blue), when few loans have been issued; by episode 1,000 they are at 0.11 and 0.10, and they end at 0.08 and 0.09. The reach rate (Figure 14a) rises from 2% of each group per episode to 7 to 9% by episode 500 and stays there. The loan distributions are the flattest of the twelve panels (Figures 15a and 15d): the most-served tenth of each group holds 12 to 13% of its loans, and the two groups receive nearly the same volume (243 and 227 loans per individual). FL under Outcome is the one cell that shares loans evenly within each group and in equal volume across the two.

Under RMM each group is served as evenly as under FL, but in very different volumes. The blue group's Lorenz curve is on the diagonal from episode 1,000 on (Gini 0.04, falling to 0.02), and the red group's approaches it more slowly (0.66 at episode 100, 0.10 at the end). The blue group receives 790 loans per individual and the red 80, ten to one, and the reach rates show the same split over time: 26% of the blue group per episode against 3% of the red. This is the micro-level counterpart of RMM's inverted inequality ratio ($\rho = 0.10$, Appendix F.2.2).

SW is the one Outcome cell with within-group concentration. The red group's Lorenz curve (Figure 4c) has a flat segment over the first 45% of the population and a Gini of 0.47 from episode 1,000 on: nearly half the red group receives almost nothing (median 35 loans, mean 71) while the rest is

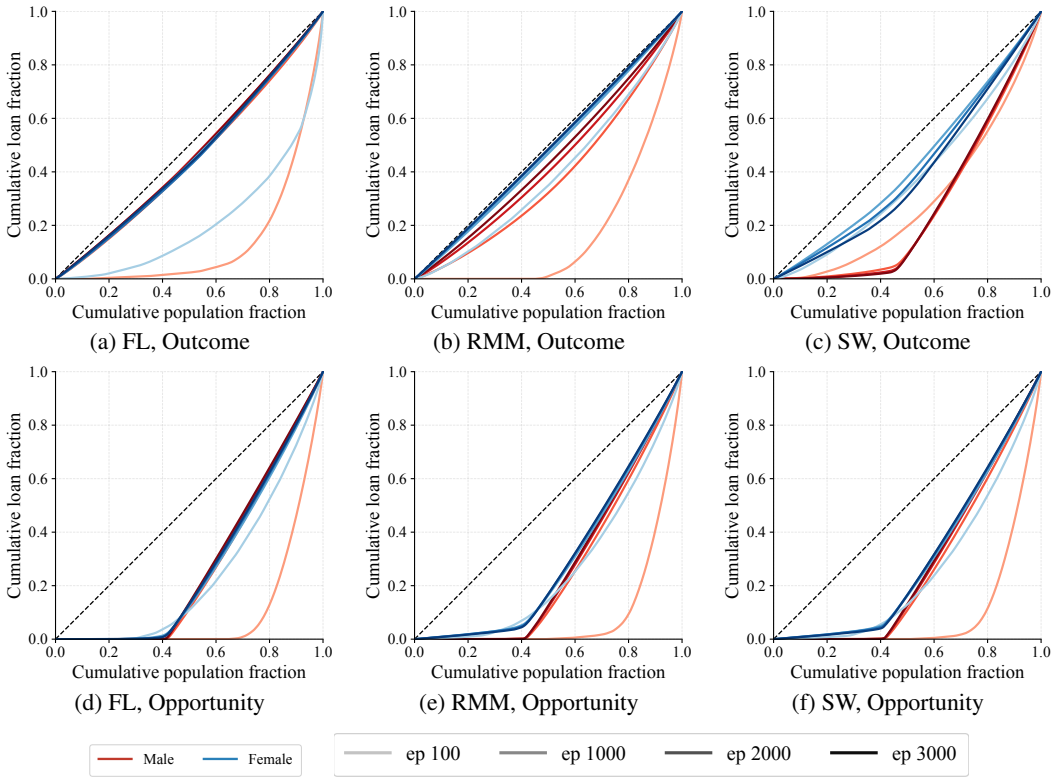


Figure 13: **PeRL: Lorenz curves of cumulative loans.** Cumulative loan fraction against cumulative population fraction within each group (red: male, blue: female), at episodes 100 to 3,000 (light to dark), for PeRL under FL, RMM and SW with the Equality of Outcome (top) and Equality of Opportunity (bottom) perspectives. Curves are averaged over 10 seeds. The dashed diagonal is perfect within-group equality.

served evenly. The blue group's curve moves away from the diagonal over deployment (Gini 0.15 at episode 1,000, 0.22 at the end), and its distribution slopes down from 640 to 720 loans per individual over the most-served 45% to about 300 over the least-served third (Figure 15f). The blue reach rate peaks at 19% around episode 1,400 and settles at 15%; the red stays at 2%.

Equality of Opportunity. The three Opportunity cells look alike, and unlike any Outcome cell. From episode 1,000 on, every Lorenz curve in the bottom row of Figure 4 is flat over roughly the first 40% of the group and then rises straight to the corner: part of the group receives little or nothing, and the rest is served nearly uniformly. The Gini coefficients are 0.41 to 0.47 at episode 1,000 and 0.40 to 0.44 at the end. The loan distributions show the same structure as a plateau followed by a drop (Figure 16). For the red group it is the same in all three cells: the most-served 57% sit on a common plateau, the next 15% receive about one loan, and the last 25% none, in every seed (22 to 30% per seed). The blue group has the same profile under FL (21% never served). Under RMM and SW its mean profile shows a second, lower plateau of about 60 loans over the last 40%, an average over seeds: in eight RMM seeds and nine SW seeds, 14 to 21% of the blue group are never served, while one RMM seed excludes 8% and one seed of each serves the whole group.

What the constraint changes under Opportunity is the volume each group receives, not the shape. The red group's plateau stands at four times the blue group's under FL (318 against 80 loans per individual), 1.7 times under SW (265 against 152), and close to level under RMM (231 against 191). The reach rates show how this builds up (Figure 14, bottom row): the red group's rises throughout deployment, to 12% under FL, 9% under RMM and 10% under SW, while the blue group's peaks at 5 to 7% between episodes 130 and 670 and then declines, to 2%, 5% and 3.5%. The blue line starts above the red and falls below it, earliest under FL and latest under RMM.

Summary. The three views separate the perspectives more sharply than any aggregate metric. Under Equality of Outcome everyone is served, within-group shares are even, and the constraint

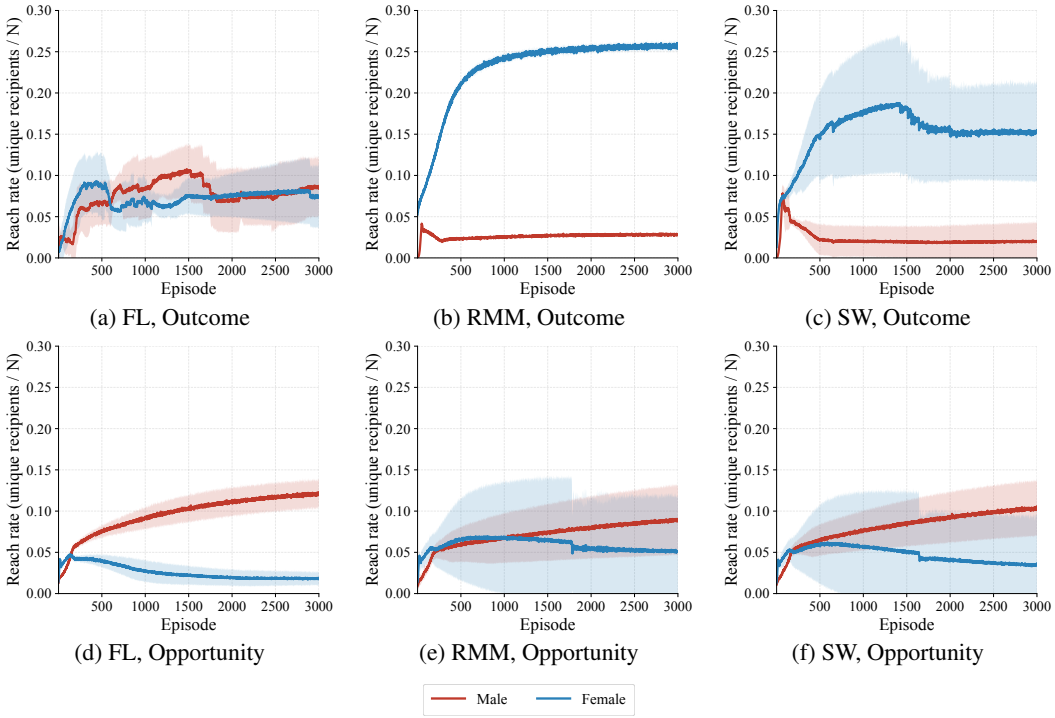


Figure 14: **How much of each group PeRL reaches over time.** Share of the red (male) and blue (female) groups receiving a loan in each episode; mean ± one standard deviation over 10 seeds. Top: Equality of Outcome; bottom: Equality of Opportunity.

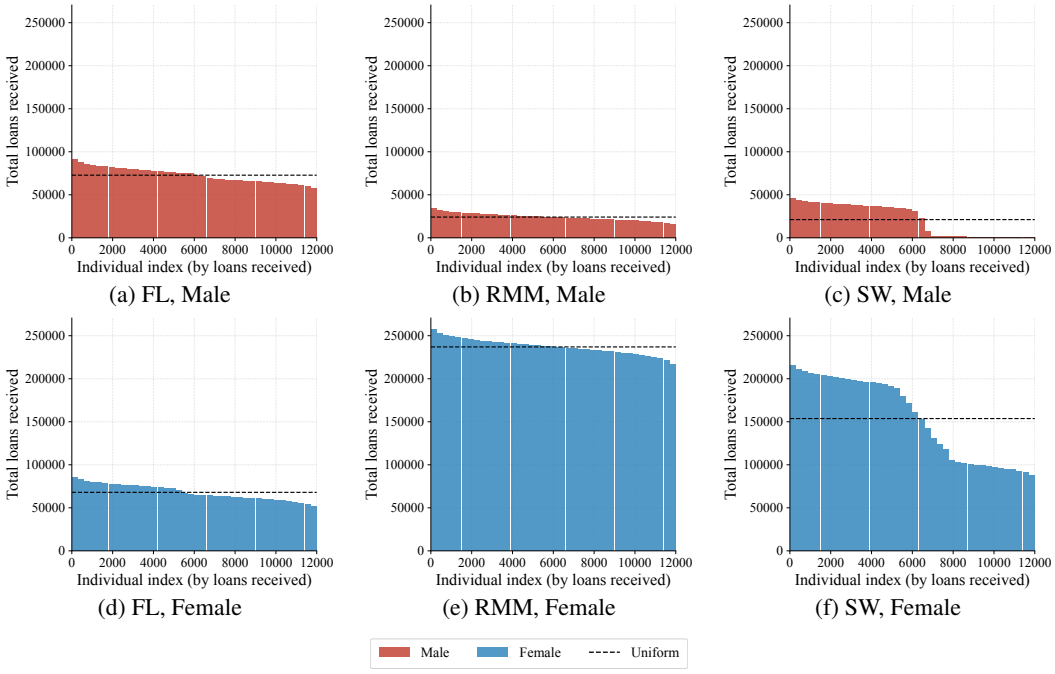


Figure 15: **Loans received by each individual, PeRL under Equality of Outcome.** Individuals are ordered from most to least served, in bins of 300; each bar is a bin's total loans over deployment, and the dashed line is what every bin would receive under equal shares. Top: male group; bottom: female group. Mean over 10 seeds.

sets the balance between groups: equal under FL, ten to one in the blue group's favour under RMM, and, under SW, a blue group served broadly against a red group of which half receives almost nothing. Under Equality of Opportunity roughly a quarter of the red group and a fifth of the blue are

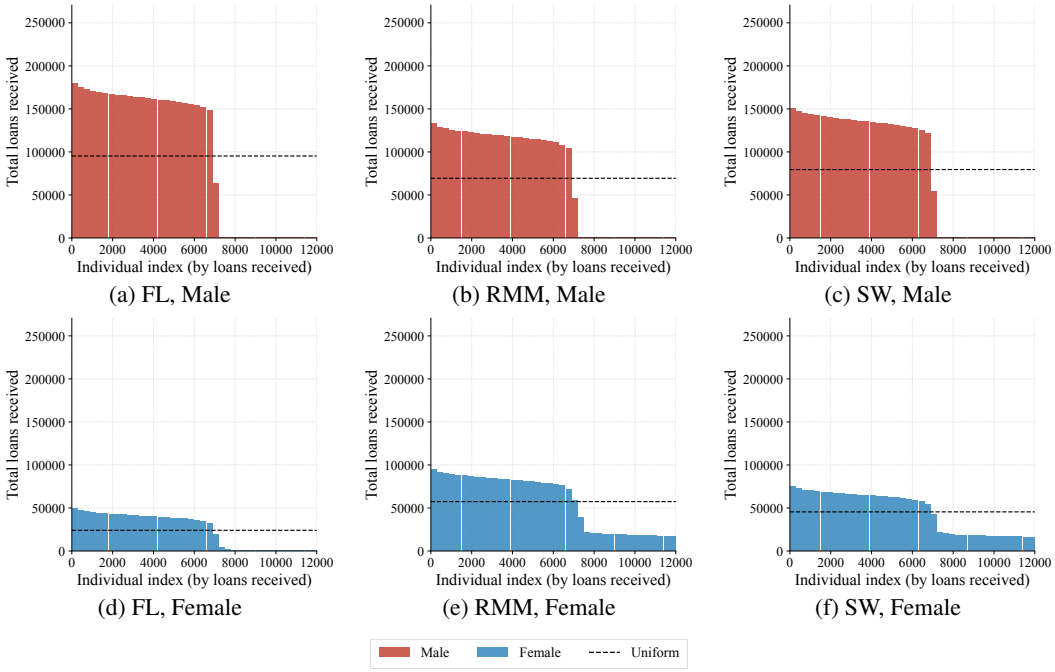


Figure 16: **Loans received by each individual, PeRL under Equality of Opportunity.** As in Figure 15, on the same axes. The drop to zero on the right of each panel marks the individuals who were never served.

never served, the rest are served uniformly, and the constraint only sets the ratio between the groups' volumes. Appendix F.3 adds the fixed-multiplier FL policy to this comparison.

F.3 The Best of Both Worlds: A ADM's Policy for Equity and Profit

Appendices F.2.3 and F.2.4 leave the ADM with a choice between two FL cells. FL under Equality of Outcome closes the wealth gap, holds ρ near 1 and serves every individual of both groups, at a profit of 1.30M. FL under Equality of Opportunity earns 7.44M, the most of any cell, but leaves a gap of 1,087, $\rho = 3.98$ and a quarter of each group never served. This section describes PeRL-$\gamma^\star$, a single policy that keeps the constraint and perspective of the first cell and recovers most of the profit of the second. It is the policy reported for FL under Equality of Outcome in Table 2.

Method. Dual ascent sets the strength of the fairness term from how slack the constraint currently is, not from the trade-off the ADM wants. For FL this drives the multiplier to one of two extremes (Appendix F.6): under Outcome it rises to the cap and then moves between floor and cap, and under Opportunity it falls to the floor within the first episodes, so the fairness term is either enforced at full strength or switched off. PeRL-$\gamma^\star$ replaces the constraint with a fixed trade-off: PeRL is trained on the reward of Equation (8) with $\gamma^t \equiv \gamma^\star$ throughout pre-training and deployment, so that it maximises a single weighted utility of profit and the FL term rather than solving a constrained problem. A fixed weight needs no threshold ϑ, and each value of $\gamma^\star$ selects one point on the trade-off between profit and parity. We take $\gamma^\star = 1.87$, the strength at which the learned FL multipliers of the Outcome and Opportunity cells meet during pre-training, a level the dual ascent itself passes through on its way to the fairness-holding and to the profit-earning extreme. Everything else is as in Appendix F.2: the PeRL agent, the FL constraint under the Outcome perspective, the hyperparameters of Table 6, and 1,000 pre-training and 3,000 deployment episodes over 10 seeds.

Aggregate outcomes. Table 13 and Figure 17 compare PeRL-$\gamma^\star$ with the two FL cells. It is the best of the seven policies on all three fairness metrics: a wealth gap of 46.5, below FL under Outcome's 71.9 and four to twenty-three times below the Opportunity cells (208 to 1,087); $\rho = 1.045$, the closest to 1 in the grid; and an approval-rate disparity of 0.004, the smallest of any cell. Its profit is 4.91M, 3.8 times FL under Outcome and two thirds of both FL under Opportunity (7.44M) and the unconstrained MaxUtil agent (7.50M).

Table 13: **PeRL-$\gamma^\star$ against the two FL cells with a learned multiplier.** End-of-deployment metrics, mean $\pm$ standard deviation over 10 seeds, as in Table 7. RMM under Opportunity is included as the Opportunity cell closest to balance. Best value per column in bold.

Policy	Cumulative Profit ($\times 10^6$)	|Wealth Gap|	|Approval Rate Disparity|	Inequality Ratio ρ
FL / OUT	1.304 ± 0.974	71.9 ± 298.0	0.0119 ± 0.2550	1.059 ± 0.094
FL / OPP	$\mathbf{7.443 \pm 0.410}$	1086.8 ± 332.2	0.0097 ± 0.0129	3.981 ± 0.568
RMM / OPP	6.711 ± 1.229	207.9 ± 1328.7	0.0543 ± 0.0792	1.229 ± 0.643
PeRL-$\gamma^\star$	4.909 ± 0.579	$\mathbf{46.5 \pm 82.1}$	$\mathbf{0.0039 \pm 0.1114}$	$\mathbf{1.045 \pm 0.032}$

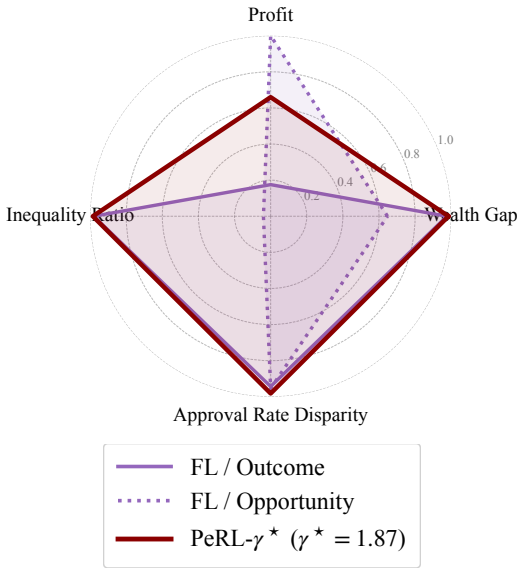
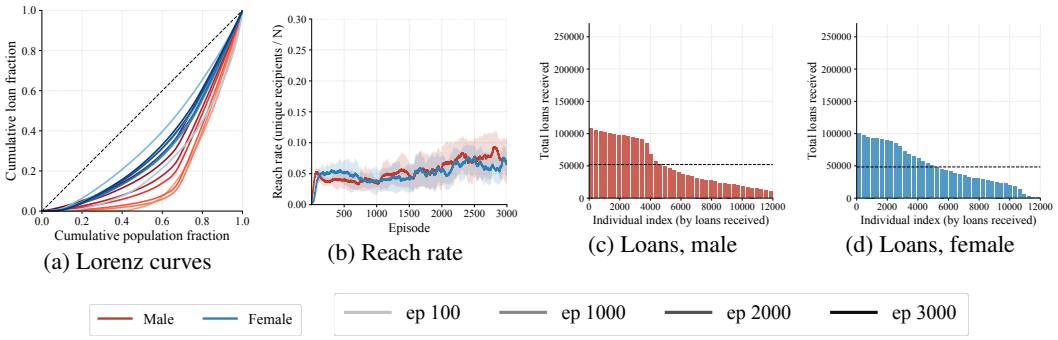


Figure 17: **PeRL-$\gamma^\star$ against the two learned-multiplier FL cells.** Axes as in Figure 12, rescaled to include PeRL-$\gamma^\star$; outward is better on every axis.

Table 14: **PeRL-$\gamma^\star$ at the micro level.** As in Table 12, for PeRL-$\gamma^\star$ and the two learned-multiplier FL cells. R: red (male), B: blue (female).

	Loans / individual		Never served		Top-10% share		Gini		Reach rate	
Policy	R	B	R	B	R	B	R	B	R	B
FL / Outcome	243	227	0%	0%	0.12	0.13	0.08	0.09	0.086	0.074
FL / Opportunity	318	80	24%	21%	0.18	0.20	0.43	0.44	0.120	0.018
PeRL-$\gamma^\star$	174	161	1%	1%	0.21	0.20	0.38	0.36	0.072	0.069

The three policies differ in what they do with an application. FL under Outcome approves three quarters of each group's applications (0.75 and 0.74), and its loans default at close to the population rate (14.5%): it holds the gap by lending to nearly everyone who applies, earning 0.27 per loan. FL under Opportunity approves 65% of each group's applications, with 11% of its loans defaulting, and earns 1.56 per loan; but since its constraint does not depend on how many apply, the red group ends up filing four times as many applications as the blue (5.8M against 1.5M over deployment) and receiving four times the loans. PeRL-$\gamma^\star$ is selective like the Opportunity cell and even-handed like the Outcome cell. It approves 59% of the red group's applications and 58% of the blue group's, 11 to 12% of its loans default, and it earns 1.26 per loan; with the same approval rate in both groups, both keep applying at the same volume (3.5M and 3.3M applications), receive the same volume of credit (2.1M and 1.9M loans), and grow their wealth at the same rate ($R_M = 0.0026$, $R_F = 0.0025$). The profit comes from lending to fewer, safer applicants; the fairness from doing so identically in both groups.



(a) Lorenz curves (b) Reach rate (c) Loans, male (d) Loans, female

Figure 18: **Who receives the loans under PeRL-$\gamma^\star$.** (a) Lorenz curves, (b) reach rate and (c, d) loans received by each individual, read as in Figure 4 and Appendix F.2.4, on the same axes. Red: male group; blue: female group.

Who receives the loans. Figure 18 and Table 14 place PeRL-$\gamma^\star$ between the two FL cells. It is as inclusive as the Outcome cell: 1% of each group is never served, against 0% under FL/Outcome and 21 to 24% under FL/Opportunity, and the two groups receive nearly the same volume (174 and 161 loans per individual). After the first 100 episodes its reach rate stays between 3.5 and 9% of each group, the two groups' curves overlapping throughout and both rising to about 7% by the end. It differs from FL under Outcome in how evenly loans are spread within each group. Its Lorenz curves end at a Gini of 0.38 (red) and 0.36 (blue), between FL under Outcome (0.08, 0.09) and FL under Opportunity (0.43, 0.44); its loan distributions show a most-served third of each group at roughly twice the equal-share line, then a step and a long decline toward zero, and the top tenth of each group holds 20 to 21% of its loans. These curves are still moving at episode 3,000, more than any in Figure 4: the red group's Gini falls from 0.65 at episode 100 to 0.50 at 2,000 and 0.38 at 3,000, and the blue group's, at 0.38 at episode 1,000 and 0.41 at 2,000, ends at 0.36.

What the ADM gets. PeRL-$\gamma^\star$ delivers the Outcome cell's fairness at the aggregate level while earning nearly four times its profit, and its inclusivity too: nobody is left out, and both groups are served in equal measure. What it does not reproduce is the Outcome cell's within-group evenness: its loans reach everyone but not in equal amounts, at a concentration between the two FL cells. A ADM whose objective is the between-group outcome, with a profit that must not collapse, is better served by PeRL-$\gamma^\star$ than by either FL cell; a ADM who also requires every individual to be served alike still needs the learned multiplier and its cost.

F.4 Why a Policy-Gradient Learner: PG versus SAC

The passive learner of the main text is vanilla policy gradient (PG) on a Beta policy, and PeRL is its performative counterpart (Basu et al., 2025), which differentiates through the population's response to the policy. A natural question is whether a stronger passive learner would make a better baseline. We therefore trained Soft Actor-Critic (SAC) through the same pipeline: the same policy network, observation, per-applicant reward and dual ascent as PG, with the update replaced by off-policy soft actor-critic (twin Q critics with Polyak targets, automatic temperature tuning, a replay buffer of 10,000 cohorts, discount factor $\zeta_{\text{discount}} = 0.99$, Polyak rate 0.005, learning rate 10^{-3}). Since there is no per-applicant next state, the critic bootstraps on the mean soft value of the next cohort. SAC is run in all six cells with 1,000 pre-training and 3,000 deployment episodes and 10 seeds. Table 15 compares the two learners at the end of deployment, and Figures 19 and 20 over deployment.

What SAC learns. SAC learns a score cutoff. In all six cells it approves 53% of the red group's applications, and in five of them 55% of the blue group's, whatever the constraint or perspective. The approved half is the lower-risk half: 11 to 12% of SAC's loans default, against 15% in the population, and the loans an individual receives track their wealth, with a correlation of 0.81 (Outcome) and 0.85 (Opportunity) between wealth at episode 100 and loans received by episode 3,000, against 0.31 and 0.32 for PG. The fairness term plays little part: in five of the six cells the multiplier is at its floor for 75 to 99% of deployment, and four cells (SW under Outcome and all three Opportunity cells) end at the same point, 6.9 to 7.0M in profit, a gap of 1,509 to 1,517 and ρ of 9.3 to 9.4. The two exceptions are Outcome cells. Under RMM, the one cell whose multiplier stays off the floor (4% of episodes

Table 15: **PG versus SAC at the end of deployment.** Approval rate of the red and blue group, cumulative profit, |wealth gap| and inequality ratio ρ, as in Table 7; means over 10 seeds.

		Approval rate R / B		Profit ($\times 10^6$)		|Wealth Gap|		Inequality Ratio ρ	
Persp.	Constr.	PG	SAC	PG	SAC	PG	SAC	PG	SAC
OUT	FL	0.32 / 0.31	0.53 / 0.55	0.37	5.26	745	8	0.19	0.97
OUT	RMM	0.98 / 0.98	0.53 / 0.87	1.10	5.44	3190	1074	10.20	4.05
OUT	SW	0.97 / 0.97	0.53 / 0.56	1.10	6.91	3170	1517	10.29	9.31
OPP	FL	0.97 / 0.96	0.53 / 0.55	1.13	7.04	3155	1511	10.33	9.30
OPP	RMM	0.98 / 0.97	0.53 / 0.55	1.11	7.03	3177	1509	10.23	9.29
OPP	SW	0.96 / 0.95	0.53 / 0.55	1.12	7.04	3116	1516	10.30	9.36

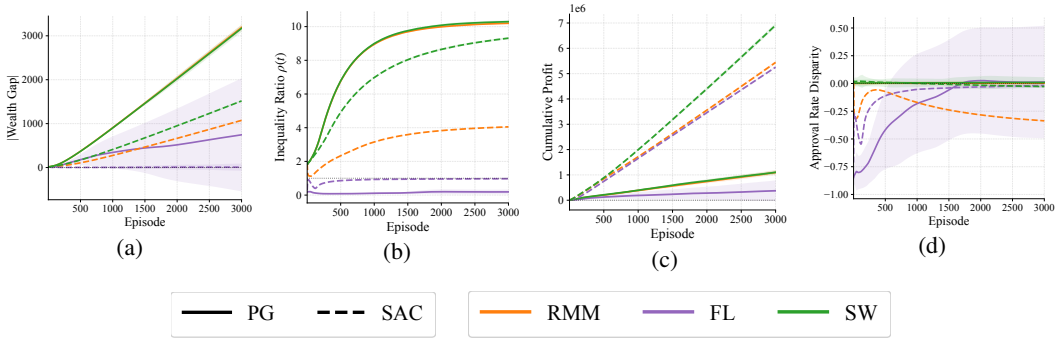


Figure 19: **PG (solid) and SAC (dashed) under Equality of Outcome.** (a) Wealth gap, (b) inequality ratio, (c) cumulative profit and (d) approval-rate disparity (red minus blue) over deployment, for each fairness constraint. Mean over 10 seeds; bands are one standard deviation across seeds, bootstrapped for ρ.

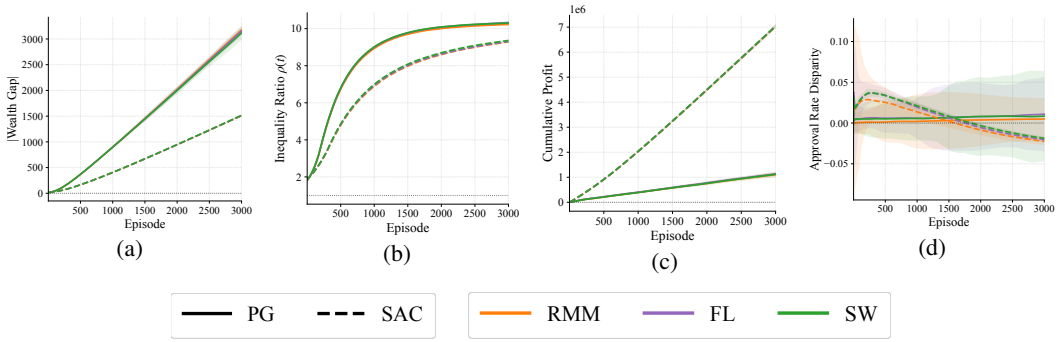


Figure 20: **PG (solid) and SAC (dashed) under Equality of Opportunity.** Panels as in Figure 19.

at the floor), the blue group's approval rate rises to 0.87 while the red group's stays at 0.53. Under FL, the cutoff is applied to both groups alike, which balances the aggregates (gap 8, $\rho = 0.97$) while loans still follow wealth within each group (Gini 0.41 in both).

Why PG is the baseline. A passive baseline is useful only if it responds to the reward it is given. SAC settles on the same cutoff under four of the six rewards and on small variations of it under the other two: a fixed score threshold of the kind the rules of Appendix E.3.1 define by hand, here recovered by a critic. The threshold denies credit at every step to the lower-scored half of each cohort, whose wealth a loan would move most, and in the four cells where SAC settles on it the red group's wealth grows at half the rate it does under PG ($R_M = 0.006$ against 0.012). PG, in contrast, responds to each reward differently, and it is what PeRL reduces to without the performative term, so the comparison of Appendix F.2 isolates that term. SAC's higher profit is the profit of a fixed cutoff, and a learner that holds a fixed cutoff does not engage with the question of this paper: what a ADM does when its decisions reshape the population it faces next.

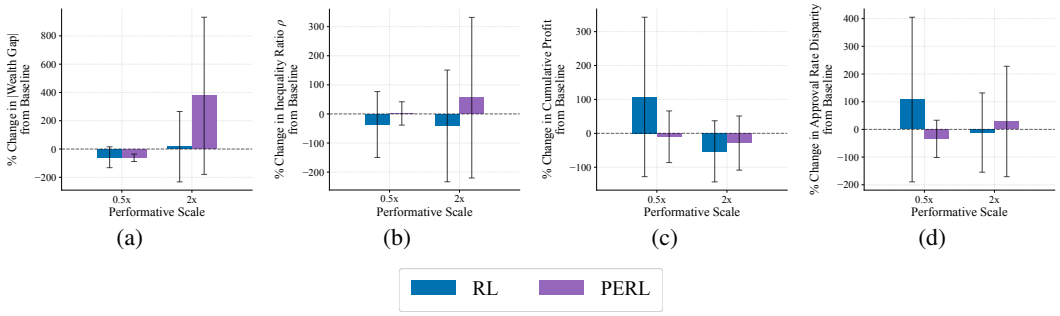


Figure 21: **Sensitivity to the strength of the population's response.** Change in (a) wealth gap, (b) inequality ratio, (c) cumulative profit and (d) approval-rate disparity at the end of deployment, relative to the default simulator, when the performative channels are halved or doubled; RL and PeRL under FL and Equality of Outcome. Error bars: one standard deviation across 10 seeds.

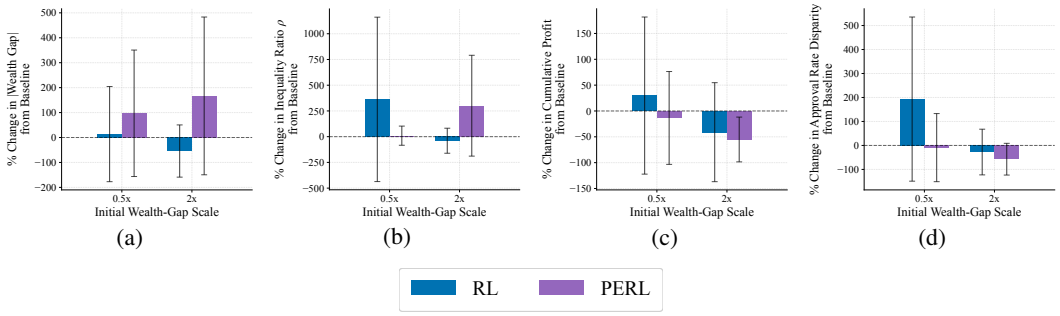


Figure 22: **Sensitivity to the initial wealth gap.** As in Figure 21, with the initial gap between the groups halved or doubled.

F.5 Robustness to the Simulator's Settings

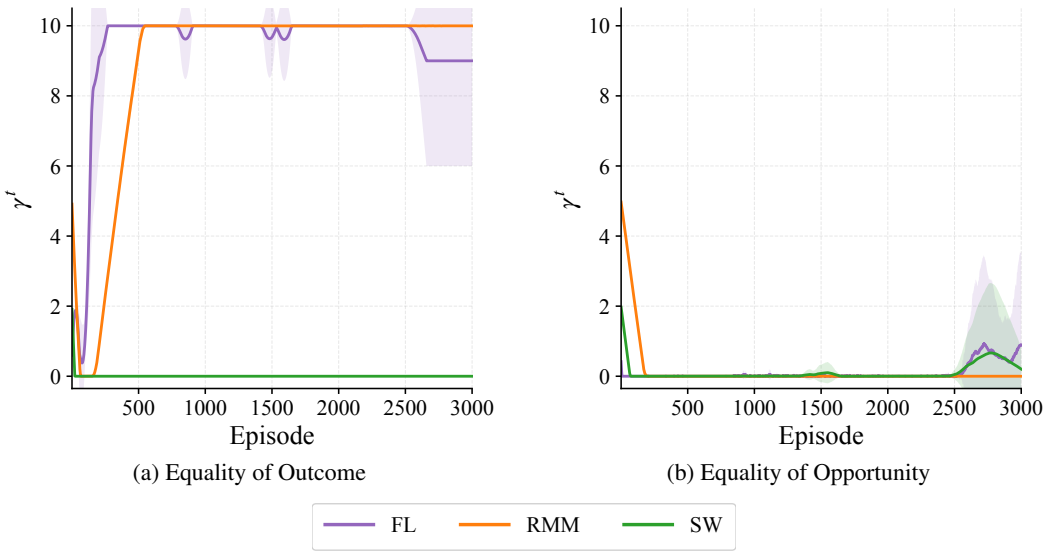
Two settings of Eutopia fix the strength of the effects the paper studies: how strongly the population responds to lending, and how unequal the groups are at the start. We vary each around its default for RL and PeRL under FL with Equality of Outcome and the learned multiplier of Appendix F.2, with 1,000 pre-training and 3,000 deployment episodes and 10 seeds per setting.

- **Performative strength** $\times 0.5$ and $\times 2$: one factor scales both channels through which decisions feed back into the population, the Hawkes excitation α_g of an approval and the wealth gain κ_i of a repaid loan. At $\times 1$ it is the simulator of Section 3.
- **Initial wealth gap** $\times 0.5$ and $\times 2$: each group's initial wealth is rescaled by one factor so that $\mu_M^0 - \mu_F^0$ is halved or doubled, keeping the population-mean wealth, and hence the arrival rates, which depend on $\mu_g/\bar{\mu}$, and each group's within-group distribution unchanged. Creditworthiness, default probabilities and loan amounts are not changed.

Figures 21 and 22 report each end-of-deployment metric as the percent change from the same agent's $\times 1$ result. Here the metrics are per-seed final values averaged over seeds, so the $\times 1$ references differ from Table 7: PeRL's wealth gap is 252 (the mean of the per-seed gaps, against the gap of the mean, 72), and ρ is the mean of the per-seed ratios (1.11 for PeRL and 0.92 for RL, against 1.06 and 0.19).

Performative strength. At half strength PeRL's wealth gap falls by 62% (96 ± 30 against 252 ± 154), with every seed ending at 120 or below, and its disparity by 34%, while its profit and ρ stay within about 10% of the default. RL's gap also falls (-58%), while its disparity and profit double, with uncertainties larger than the changes. At twice the strength PeRL's gap rises to $1{,}201 \pm 1{,}193$, nearly five times the default, some seeds holding the gap and others not; its profit falls by 29% and ρ rises to 1.7 ± 3.0. RL loses half its profit, and its ρ moves further from 1.

Initial wealth gap. Changing the initial gap in either direction widens PeRL's final gap, to 497 ± 563 when it is halved and 673 ± 683 when it is doubled, against 252; its disparity falls (-9% and -57%) and so does its profit (-14% and -55%). With the doubled gap ρ reaches 4.5 ± 5.3: the



(a) Equality of Outcome

(b) Equality of Opportunity

Figure 23: **The multiplier over deployment for RL.** γ^t for FL, RMM and SW; mean $\pm$ one standard deviation over 10 seeds. Each cell starts at its γ^0 (0.5 for FL, 2 for SW, 5 for RMM), between the floor 10^{-4} and the cap 10. PeRL's multipliers are in Figure 3.

red group pulls ahead in some seeds and not in others. RL's changes are within their uncertainty at every setting. Across both settings the seed spreads are wide, and only two shifts clear them: PeRL's smaller gap at half the performative strength and its lower profit at twice the initial gap. The others indicate directions rather than sizes.

F.6 Trajectories for Lagrangian Multiplier

Figures 3 (PeRL) and Figure 23 (RL) show the multiplier of Algorithm 2 in every cell, from its reset value γ^0 at the start of deployment. Under Equality of Outcome the three constraints behave differently. FL rises from 0.5 to the cap within the first few hundred episodes (by episode 545 for PeRL and 264 for RL); PeRL's multiplier then moves between floor and cap for the rest of deployment, at the cap in about half of all episodes, while RL's stays at the cap in over 90% of episodes. RMM first falls from 5 to the floor within 40 to 65 episodes, then climbs to the cap by episode 700 (530 for RL) and stays there. SW falls from 2 to the floor within about 100 episodes and does not leave it.

Under Equality of Opportunity all three multipliers fall to the floor and stay there for both agents: FL within 40 episodes, SW within 70 to 125, and RMM, which starts highest, within 180 to 380. The multiplier is at the floor in more than 93% of deployment episodes in every Opportunity cell; RL's small late rises under FL and SW stay below 1.

The two cells whose multiplier ends high, FL and RMM under Outcome, are the two whose constraint is written on the groups' wealth. The four whose multiplier ends at the floor have constraints on true positive rates or on total credit, which the policy satisfies early in deployment.

G Computational Resources

All experiments run on CPUs; no GPU is used. The policy networks are small, and the cost of a run is dominated by simulating the environment: 24,000 individuals over 4,000 episodes of 200 decision steps each. Every run, one agent in one cell with one seed, is an independent job on a compute cluster with 2 CPU cores and 8 GB of memory, and takes about one hour; the rule-based policies, which are only deployed, use 1 core and 4 GB. The reported experiments comprise 120 runs of RL and PeRL across the six cells, 60 of SAC, 80 for the robustness study, 10 of PeRL-$\gamma^\star$, 20 of the MaxUtil agents and 70 rule-based deployments, all over 10 seeds.