Extended Abstract

Motivation It has been long established that racial disparities are systematically enforced through current mortgage lending systems in the United States. The historical impacts of redlining on Black and African American populations remain bitingly evident in trends of lower home mortgage lending approval rates, capping social mobility and further disenfranchising vulnerable communities through the fingerprints of racist legislation that permeates policies across the U.S. Since the turn of the century, research has increased into algorithmic fairness in static environments; however, there is still a prominent gap in research analyzing the potential of RL policies to reduce disparities in multi-step environments. This work aims to address that gap by showing the benefits of using RL to learn more equitable policies across long horizons.

Method We simulate a mortgage lending market over 50 rounds using a custom Gymnasium environment. Six groups are initialized across Race (White, Black/African American) and Income Level (Low, Middle, Upper). Each round, the agent sets approval thresholds for each group based on the current credit distribution. Approved borrowers who repay gain $+\delta_{repay}$, those who default lose $-\delta_{default}$, and 5% of the population turns over. We use two baselines—one profit-maximizing policy and one per-step demographic parity policy—and three reward functions—profit only, profit penalized by population-weighted variance in approval rates, and profit penalized by population-weighted variance in group credit score means. The fairness penalty weight is swept across $\lambda = \{10, 100, 1000, 10000, 100000\}$.

Implementation The PPO agents were trained with an MLP policy, a discount factor of $\lambda = 0.99$, and a learning rate of 3×10^{-4} for 200,000 timesteps per run, resulting in 75 total runs (3 reward modes $\times$ 5 λ values $\times$ 5 random seeds). The code for our work can be found at: `github.com/namiecake/CS224R-project`.

Results Our results show that there are clear tradeoffs between profit and fairness in dynamic pricing that are difficult to overcome. The PPO agents consistently outperformed our ProfitMax baseline while maintaining or improving on equity scores. This is encouraging, as it signals that RL has clear potential in improving fairness scores. However, the fairness improvements shown in these results are relatively minimal, with small penalties demonstrating little impact and large penalties resulting in training instability. Thus, while some runs achieved a more optimal profit-equity tradeoff, the improvements were not consistent across seeds. Hence, our findings suggest that reinforcement learning is a promising approach to reduce biases in dynamic pricing contexts, but methods to achieve reliable fairness need more research.

Discussion Our results illustrate that PPO discovers strategies that is able to surpass one-step baselines on both profit and fairness, but more work with reward shaping is needed to reduce inequality. Further, the PPO agent trained only on maximizing profit was able to improve on profit compared to ProfitMax while keeping inequality about the same. This shows that even without a fairness signal, reinforcement learning methods are able to learn a greedy policy that achieves more profit than one-step methods. We also consider some limitations; first, our simulated environment may not accurately reflect real-world dynamics. Additionally, we used populated-weighted variance as our one metric of fairness, but other fairness metrics exist and may yield better reward functions or different conclusions.

Conclusion To conclude, we evaluated PPO-based reinforcement learning agents on fairness in lending and compared them against baselines that only prioritized profit or only prioritized fairness. Overall, our findings suggest that reinforcement learning is a promising approach for dynamic lending, but improving drastically on fairness while keeping profit about the same is still to be explored. Future work could explore alternative fairness formulations and more stable optimization methods.

The Long Game: A Long-Horizon RL Study of Fairness in Financial Lending

Naomi Boneh
Department of Computer Science
Stanford University
naomicyb@stanford.edu

Christelle Millos-Lopez
Department of Computer Science
Stanford University
crissyl@stanford.edu

Brydie Sigg
Department of Computer Science
Stanford University
brysigg@stanford.edu

Abstract

Standard machine learning fairness criteria treat decisions as independent events, enforcing constraints like demographic parity at a single point in time. However, in deployed systems such as mortgage lending, decisions reshape the populations they act on: approved borrowers who repay improve their financial standing, while denied applicants lose the opportunity to build credit. Prior theoretical work has shown that one-step fairness constraints can paradoxically worsen long-run outcomes for the very groups they aim to protect, but this analysis has largely been confined to simple threshold policies in stylized environments. We investigate whether a reinforcement learning agent optimizing over long horizons with equity-aware reward shaping can discover lending policies that produce better long-run outcomes for disadvantaged groups than static fairness baselines. Using the Home Mortgage Disclosure Act (HMDA) dataset to ground demographic distributions, approval rates, and loan characteristics in real lending data, we build a population-dynamics simulator in which creditworthiness evolves over time as a function of lending decisions and repayment outcomes. We train PPO agents under several reward functions such as profit-only, profit with a demographic parity penalty, and a blended long-run equity objective and compare them against profit-maximizing and per-step demographic parity baselines. We evaluate policies over 50 simulated rounds along four axes: long-run inter-group credit gaps, total profit, the equity-efficiency tradeoff frontier, and learned per-group threshold behavior. We hypothesize that one-step fairness baselines either fail to close or actively widen inter-group gaps over time, while long-horizon RL with fairness-shaped rewards discovers strategic policies that achieve stronger equity outcomes. Beyond lending, we expect our findings to inform sequential decision-making in other high-stakes domains such as healthcare resource allocation and criminal justice. The code for our work can be found at: github.com/namiecake/CS224R-project.

1 Introduction

When a creditworthy borrower is denied a loan, their financial standing stagnates. When they are approved and repay the loan, it improves. When these individual outcomes repeatedly appear across thousands of applicants over time, the results compound, which results in shifts in the credit score distributions of entire demographic groups. This means that policies that satisfy standard fairness

criteria only at individual decision points still entrench, or even widen, the gap between demographic groups. In some cases, this actively causes harm to the groups they aim to protect Liu et al. (2018). The resulting feedback dynamic (early decisions shape who gets selected from later applicant pools) is largely overlooked when running analyses with static fairness frameworks—the exact frameworks that currently dominate the literature on algorithmic fairness.

In this work, we investigate whether a reinforcement learning agent with rewards that incentivize fairness can discover lending policies that produce better long-term outcomes for disadvantaged groups. Specifically, we are addressing the question: Can a reinforcement learning agent with rewards that incentivize fairness discover lending policies that produce better long-term outcomes for disadvantaged groups?

To study this, we built a Gymnasium environment to simulate lending behavior over 50 rounds. We define six demographic groups by race (White and Black/African American) and income level (Low, Middle, Upper) which we initialized using 2022 HMDA mortgage data. We store persistent credit score distributions over these groups, and these distributions change in response to lending decisions each round. The agent then observes each group's credit distribution and sets per-group approval thresholds. We train the PPO agents using three reward functions: profit only, profit penalized by population-weighted variance in approval rates, and profit penalized by population-weighted variance in group credit score means. The results are compared against two baselines: a profit-maximizing threshold policy and a demographic parity policy that equalizes approval rates across groups each round.

Our key finding is that PPO agents can achieve higher profit than the profit-maximizing baseline while also maintaining similar fairness levels, and that how much fairness costs in terms of profit depends largely on how strongly the reward penalizes the unfairness (λ). However, none of the policies we studied significantly closed the inter-group credit score gap over time, suggesting that more experimentation into effective reward shaping is necessary.

2 Related Work

The majority of research into algorithmic fairness defines criteria like demographic parity and equalized odds over static populations, treating each decisions as independent and ignoring the interaction between past and future outcomes. Hardt et al. (2016) define the term "equalized odds" as the requirement that true positive and false positive rates are equivalent across demographic groups when conditioned on the outcome, and offer "equal opportunity" as a less restrictive requirement that only equalizes the treatment of applicants in the "advantaged" group—in a loaning settings, being applicants who repay thier loans. Zafar et al. (2017) build on this, proposing a method for training classifirs that can be tuned to create a tradeoff between fairness and accuracy, and Mitchell et al. (2021) survey the many different definitions of fairness that are used throughout the literature. These, and other, frameworks have generated and synthesied rigorous tools that are highly applicable in the real world, but they have a well-documented limitation. D'Amour et al. (2020) observes that most approaches focus on static settings, failing to consider long-term dynamics. This limitaiton comes fundamentally from ML frameworks, which require fixed datasets and single-timestep objectives.

While there is robust literature studying static fairness criteria, a growing body of work has identified that satisfying these criteria at each decision point doesn't always promote, and even potentially undermines, equitable outcomes over time. Liu et al. (2018) note that even with a one-step feedback model, traditional fairness criteria haven't been shown to improve fairness over time, and occasionally cause harm "in cases where an unconstrained objective would not". This demonstrates that demographic parity and equal opportunity can lead to a decline in creditworthiness for the groups they aim to protect. D'Amour et al. (2020) build on this line of inquiry, advocating for the use of simulation to study these dynamics in order to give a more complete picture of long-term consequences of ML-based decision systems. However, both works are still limited: Liu et al. (2018) use a one-step feedback model with fixed threshold policies, and D'Amour et al. (2020) study the behavior of fixed, rule-based agents instead of training adaptive agents.

Despite the potential of applying RL to fairness problems, this area of research remains underdeveloped. Jabbari et al. (2017) establish that it is computationally expensive to enforce fairness constraints exactly in RL and show that approximate fairness is much more efficient. Reuel and Ma (2024) offer a survey of the landscape of research, and note that most fairness analysis has been on one-shot

classification tasks, which doesn't accurately represent real-world systems. Additionally, there is a notable lack of research using simulators on real demographic data, especially with applications in lending systems.

In this work, we study algorithmic fairness within the social context of unequal mortgage lending. Racial disparities in this domain are well-documented; Lewis-Faupel and Tenev (2024) find significant disparities in approval rates between racial groups in U.S. mortgage data even after accounting for factors like income and debt-to-income ratio. This demonstrates evidence of persistent sysstemic barriers for disadvantaged racial categories. Prior work studying how lending policies can be deisgned to reduce these gaps over time is, however, limited. Altman (1999)'s *Constrained Markov Decision Processes* provides the theoretical framework applied here for MDPs that optimize one objective while satisfying constraints on others.

Our work directly builds on the constraints visible in prior work in this domain by training deep RL agents inside a persistent-population lending simulator using real Consumer Financial Protection Bureau (2025) mortgage data. To our knowledge, no prior work has combined deep RL with a lending simulator to study long-horizon fiarness dynamics in this way.

3 Method

We simulate a mortgage lending market over 50 rounds using a custom Gymnasium environment. Six demographic groups, initialized from 2022 HMDA data, are defined across Race (White/Black) and Income level (Low/Mid/Upper). Each round, the agent observes each group's credit distribution and sets approval thresholds per group. Approved borrowers who repay their loan gain $+\delta$, those who default lose $+\delta$, and 5% of each group is replaced to simulate population turnover. We use two baselines: `ProfitMax` which aims to maximize expected profit, and `DemographicParity` which aims to equalize approval rates between groups. These are used to compare three reward functions: profit only, profit penalized by population-weighted variance in approval rates, and profit penalized by population-weighted variance in group credit score means. We did a sweep across the fairness penalty weight $\lambda \in \{10, 100, 1000, 10000, 100000\}$ to visualize the tradeoff between profit and fairness. The PPO agents are trained for 200k timesteps across 75 total runs (3 reward functions x 5 λ x 5 seeds).

3.1 Simulation Environment

We model a mortgage lending market as a Markov Decision Process over 50 rounds. Groups are generated by fitting truncated Normal distributions to 2022 HMDA mortgage application data pulled from five U.S. states (CA, TX, FL, NY, IL), which is filtered to select home-purchase loans from White non-Hispanic and Black/African American non-Hispanic applicants. HMDA does not report credit scores, so we constructed an estimate from income, debt-to-income ratio, and loan-to-value ratio (LTV):

$$s = 0.50 \cdot \frac{\log(1 + \text{income}_k)}{\log(1 + 500)} + 0.30 \cdot \left(1 - \frac{\text{DTI}}{0.65}\right) + 0.20 \cdot \left(1 - \frac{\text{LTV}}{100}\right) \tag{1}$$

The LTV column is frequently missing in the HMDA dataset, so in many cases the LTV term is dropped and weights are normalized to $0.625/0.375$. Income levels are generated using the `income` and `ffiec_msa_md_median_family_income`, with group cutoffs determined by the Federal Financial Institutions Examination Council (2026) published income level buckets:

Table 1: FFIEC AMI Income Levels

Income Level	Percentage Relative to AMI
Low - Moderate	> 0% and < 80% of the AMI
Middle	$\geq$ 80% and < 120% of the AMI
Upper	$\geq$ 120% of the AMI

Each round, a per-group approval threshold is set by the agent $\tau_g \in [0, 1]$, and all applicants with a creditworthiness score above said threshold are approved. Repayment is handled stochastically, with the probability that an applicant repays given by:

$$P(repay|s) = \sigma(8(s - 0.5)) \tag{2}$$

After each round, scores are updated in the following ways: the scores of applicants who were denied don't change, those of applicants who repayed their loan increase by $\delta_+ = 0.05$ and those of applicants who defaulted decrease by $\delta = 0.10$ (clipped to [0, 1]). The group's per-step profit is as below:

$$\pi_g = u_+ \cdot n_{repaid,g} - u_- \cdot n_{defaulted,g}, u_+ = 1.0, u_- = 2.0 \tag{3}$$

Defaults are penalized with a rate that is twice as harsh as the rate that repayments are rewarded.

Each step, the mean μ_g, standard deviation σ_g, and population size N_g are computed from the current population for each of the six groups.

3.2 Baselines

We established two baseline policies that operate analytically on the truncated Normal distribution implied by the current state (μ_g, σ_g, N_g), rather than on the empirical population. These baselines search over a grid of candidate thresholds at resolution 0.01.

ProfitMax treats each group independently where we have that for group g it selects the threshold

$$\tau_g^* = \arg\max_{\tau} \ \mathbb{E}[\text{profit}_g(\tau)], \tag{4}$$

where expected profit is computed by integrating the per-applicant profit density:

$$N_g \cdot f_g(s) \cdot \big(u_{\text{repay}} \cdot p_{\text{repay}}(s) - u_{\text{default}} \cdot (1 - p_{\text{repay}}(s))\big) \tag{5}$$

over $[\tau, 1]$ under the group's truncated Normal, with a repayment probability of $p_{\text{repay}}(s) = \sigma(k(s - s_0))$. This policy maximizes single-round profit with no fairness constraint.

DemographicParity then looks at a common target approval rate $r^* \in [0.01, 0.99]$ and for each r^* computes per-group thresholds τ_g such that

$$P(s \geq \tau_g) = r^* \quad \forall\, g, \tag{6}$$

via the truncated Normal quantile function. It then selects the r^* that maximizes total expected joint profit across all groups subject to the equal-approval-rate constraint. This helps enforce a strict one-step demographic parity criterion for each round. Neither baseline uses the reward signal or any information beyond the current state.

3.3 Reward Shaping

During each round the agent receives a scalar reward made up of a profit term and, for the two rewards that incorporate fairness, a penalty on the populated-weighted variance of group metrics. The inequality penalty uses population-weighted variance, defined as

$$\text{WVar}(v, N) = \sum_{g=1}^{G} w_g \big(v_g - \bar{v}\big)^2, \quad w_g = \frac{N_g}{\sum_{g'} N_{g'}}, \quad \bar{v} = \sum_{g=1}^{G} w_g v_g, \tag{7}$$

which weighs each group's deviation from the mean proportionally to its population size. We then established three reward functions:

1. $R_{\text{profit}} = \sum_g \pi_g$: maximizes total lender profit with no fairness constraint; serves as the RL analogue of the ProfitMax baseline.
2. $R_{\text{approval}} = \sum_g \pi_g - \lambda \cdot \text{WVar}(r_g, N_g)$: penalizes disparity in per-round approval rates r_g across groups, incentivising the agent to equalize access to credit each round.
3. $R_{\text{mean}} = \sum_g \pi_g - \lambda \cdot \text{WVar}(\mu_g, N_g)$: penalizes disparity in the post-update group credit score means μ_g, directly targeting the long-run equity outcome rather than the per-round decision.

λ determines how much we decide to penalize inequality, which opposes the goal of maximizing profit. We use $\lambda \in \{10, 100, 1000, 10000, 100000\}$ as our range of λs for PPO training.

3.4 RL Agent

We trained Proximal Policy Optimization (PPO) agents with an MLP policy network. PPO is well-suited to this setting because the action space is continuous ($\tau \in [0,1]^6$), we have short episodes (50 steps), and the stochastic population dynamics produce noisy rewards that benefit from PPO's clipped surrogate objective. The observation we get is an 18-dimensional vector $[\mu_g, \sigma_g, N_g]_{g=1}^6$ and the action is a 6-dimensional vector of per-group approval thresholds, both in $[0,1]$. We train each agent for 200,000 environment timesteps using the default Stable-Baselines hyperparameters (learning rate 3×10^{-4}, $n_{\text{steps}} = 2048$, $n_{\text{epochs}} = 10$, clip range 0.2). To assess the sensitivity of both the reward function and the fairness penalty weight, we ran a full factorial sweep: 3 reward functions $\times$ 5 values of $\lambda \times$ 5 random seeds, for 75 total training runs. Each trained policy was evaluated by rolling out 5 additional test episodes under a fixed seed schedule and measuring cumulative profit, final population-weighted variance in group means, and the maximum pairwise credit score gap at round 50.

4 Experimental Setup

All experiments use the six-group lending simulator described in the Methods section and were run for $T = 50$ rounds per episode. Table 2 lists the fixed environment hyperparameters used across all runs. Initial group parameters are calibrated from HMDA data as described in Section 3.1 and summarised in Table 3. All stochastic components (population initialisation, repayment draws, churn) are seeded for reproducibility.

Table 2: Fixed environment hyperparameters.

Parameter	Symbol	Value	Description
Repayment gain	δ_{repay}	0.05	Score increase on repayment
Default loss	δ_{default}	0.10	Score decrease on default
Sigmoid steepness	k	8.0	Controls repayment probability curve
Sigmoid midpoint	s_0	0.50	Score at 50% repayment probability
Repayment utility	u_{repay}	1.0	Per-applicant profit on repayment
Default utility	u_{default}	2.0	Per-applicant loss on default
Churn rate	—	5%	Fraction of population replaced per round
Episode horizon	T	50	Rounds per episode

Table 3: HMDA-calibrated initial group parameters (2022, home-purchase loans, CA/TX/FL/NY/IL, $n = 50{,}000$ records).

Group	Race	Income tier	μ_{init}	σ_{init}	N_{init}
White_Low	White	Low	0.542	0.083	7,643
White_Middle	White	Middle	0.610	0.077	8,630
White_Upper	White	Upper	0.740	0.109	24,338
Black_Low	Black	Low	0.525	0.083	2,421
Black_Middle	Black	Middle	0.581	0.072	2,057
Black_Upper	Black	Upper	0.675	0.099	2,618

Baseline evaluation. ProfitMax and DemographicParity are both evaluated over 5 random seeds, each running for 50 rounds, using the `profit` reward mode. We recorded per-round metrics for every seed, and we report mean $\pm$ one standard deviation across seeds.

PPO training. Each PPO agent is trained for 200,000 environment timesteps using Stable-Baselines with the default hyperparameters (learning rate 3×10^{-4}, rollout buffer of 2,048 steps, 10 epochs per update, clip range 0.2, GAE $\lambda = 0.95$, discount $\gamma = 0.99$). We the ran a full factorial sweep of the 3 reward functions $\times$ 5 values of $\lambda \in \{10, 100, 1000, 10000, 100000\} \times$ 5 random seeds, for 75 total training runs. The R_{profit} runs use $\lambda = 0$ and are replicated across the same 5 seeds.

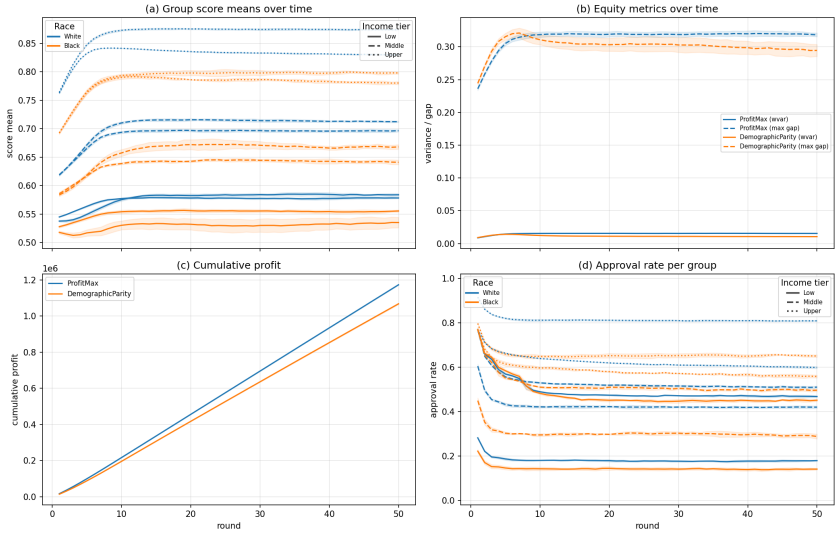

Figure 1: Visualization of baseline results.

Evaluation protocol. After training, each PPO policy is evaluated by rolling out 5 test episodes (disjoint seeds from training) of 50 rounds each, with the policy set to deterministic mode. We report three primary metrics, all averaged over test seeds: (1) *cumulative profit* — total lender profit summed over all 50 rounds; (2) *final weighted variance* — $\mathrm{WVar}(\mu_g, N_g)$ computed from the post-update populations at round 50; and (3) *max pairwise gap* — the maximum absolute difference in group credit score means at round 50. Metrics 2 and 3 are different ways to measure fairness, where lower numbers are better. To facilitate comparison, all policies (baselines and PPO) are evaluated on identical initial populations drawn from the same seeds.

5 Results

We evaluate all policies on three primary metrics over 50 simulated rounds across 5 test seeds: cumulative profit, final population-weighted variance in group credit score means ($\mathrm{WVar}_{\mathrm{final}}$), and maximum pairwise credit score gap at round 50. Baseline policies are compared against PPO agents trained under different reward functions.

5.1 Quantitative Evaluation

Table 4 gives a summary of the quantitative results across all policies. Both non-RL baselines accumulated substantial profit but failed to reduce inter-group disparity over the 50-round horizon. **ProfitMax** achieves the highest cumulative profit (1,173,066 $\pm$ std) but also the worst equity outcome, ending with a population-weighted variance of 0.0154 and a maximum pairwise gap of approximately 0.30. **DemographicParity** reduces the final weighted variance by 31% relative to ProfitMax (0.0106 vs. 0.0154) and narrows the maximum pairwise gap to approximately 0.26, but at a 9% reduction in cumulative profit. Crucially, the score gap still grows monotonically under both baselines, confirming that one-step fairness constraints are insufficient to close the inter-group gap over time.

PPO agents are able to significantly increase profit while improving fairness compared to ProfitMax, yielding a useful result for lending companies. We also find our reward function R_{approval}, which utilizes the variance group means as a penalty, is not as accurate of a reward function in this setting, as the agents trained with it improved less on fairness. Utilizing the variance in group means (R_{mean}) is a slightly more informative reward, as it was able to achieve better equity and higher cumulative profit. The best λ values for R_{mean} is $\lambda = 10$ and for R_{approval} is $\lambda = 1000$, where a higher λ indicates more penalty for inequality. Overall, R_{mean} with $\lambda = 1000$ achieves the best equity-profit tradeoff; compared to ProfitMax, it is able to reduce final WVar by about 11% while improving on profit by 14%.

Another notable result is that the The R_{profit} PPO agent improves on profit compared to ProfitMax while keeping inequality about the same. This shows that even without a fairness signal, reinforcement learning methods are able to learn a greedy policy that achieves more profit than one-step methods. Our results illustrates that PPO discovers strategies that meaningfully beat one-step baselines on profit, but more work with reward shaping is needed to reduce inequality. Table 4 shows our quantitative results with the best values of λ. Figure 2 shows comparisons of the baselines to all PPO agents.

Table 4: Performance comparison across all policies (mean results with 5 test seeds). λ^* is the best λ value for each method. $\uparrow$ means higher is better; $\downarrow$ means lower is better.

Method	Cum. Profit $\uparrow$	$\text{WVar}_{\text{final}}$ $\downarrow$	Max Gap (r50) $\downarrow$
ProfitMax	1,173,066	0.0154	0.318
DemographicParity	1,067,788	0.0106	0.294
PPO (R_{profit})	1,285,990	0.0156	0.357
PPO (R_{approval}, λ^*=1000)	1,332,690	0.0140	0.304
PPO (R_{mean}, λ^*=10)	**1,334,158**	**0.0135**	**0.298**

5.2 Qualitative Analysis

Figure 1 shows per-round trajectories of group score means, equity metrics, cumulative profit, and approval rates for the two baseline policies. Several dynamics emerge clearly from the trajectories. First, the inter-group score gap grows rapidly in the first 10–15 rounds under both policies before plateauing, suggesting that early-round lending decisions have an outsized influence on long-run outcomes. This implies that an RL agent that reasons over the full 50-round horizon may benefit most by intervening in early rounds when group distributions are still close together. Second, income tier creates substantial within-race stratification: White_Upper converges toward a score mean of approximately 0.85 by round 50, while Black_Low barely exceeds 0.55, producing a within-policy gap of roughly 0.30. The DemographicParity policy closes the *approval-rate* gap between White and Black groups within each income tier, but the *score* gap continues to grow because Black applicants default at higher rates under the same threshold, so repayment-driven score gains accumulate more slowly for disadvantaged groups. This decoupling between approval-rate parity and score parity is a central motivation for the R_{mean} reward, which targets the score gap directly rather than the decision gap.

Regarding the PPO results, the R_{approval} agent equalises approval rates in early rounds, but the score gap grows throughout the episode because equal approval rates leave the default dynamics the same. By round 50 it reduces WVar by 9% compared to ProfitMax, which is a modest improvement. The R_{mean} agent addresses the score gap directly and achieves a larger improvement in fairness. Interestingly, there is no tradeoff with profit, as R_{mean} also achieves the highest cumulative profit of any policy. These results indicate that penalizing the distributional gap directly, rather than the results of unequal approvals, yields better long-horizon equity outcomes and better efficiency, which is consistent with our hypothesis.

6 Discussion

Our results showed that, although PPO agents are able to consistently achieve higher profit than the baseline policies, no policy closed the inter-group credit score. This suggests that the fairness penalties we tested failed to give a strong enough signal to redirect the agents towards more equitable outcomes. However, PPO agents outperform the baselines, suggesting that RL is a viable approach for long-horizon lending policy optimization and addressing algorithmic bias.

The most significant finding here is that PPO at $\lambda = 1000$ achieves substantially higher profit than `ProfitMax` ($1.33M vs. $1.17M) while improving fairness levels (WVar 0.014 vs. 0.0156). This suggests that the fairness penalty is able to find a middle ground between fairness and profit that improve both over the baseline. More speculatively, if replicated with real-world results, the improvement in profit might offer additional evidence that current lending policies are unnecessarily punitive toward certain groups. If approving more borrowers from disadvantaged groups increases

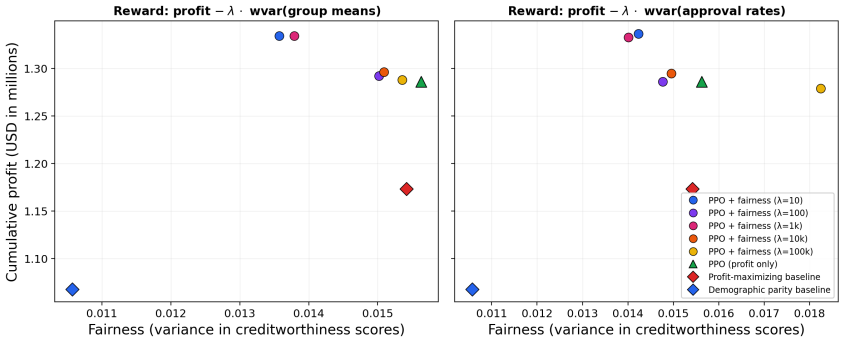

Figure 2: Comparison of profit and fairness results for baseline methods and PPO agents trained with different penalty magnitudes (λ). The diamond points are baselines, the triangle point is the PPO agent trained with R_{profit}, and the circular points are the PPO agents trained with varying values of λ in the reward function that penalize inequality. Since fairness is measured by the variance between groups, lower values on the x-axis are desirable and indicate more fairness.

profit, then it implies that those borrowers were capable of repayment and their prior denial reflected discriminatory biases rather than a genuine financial risk assessment.

There is nonetheless a prevalent tradeoff between fairness and profit, where reward modes that strongly prioritize fairness demonstrate notable drop offs in profit, as seen most clearly in the demographic parity baseline (best PPO WVar: 0.014 vs DemographicParity WVar: 0.106). From this, we can conclude that RL finds a meaningfully better point on the profit-fairness tradeoff curve than policies that only prioritize profit, but they are not optimal when considering fiarness above all.

Our results were impacted by our reward design choices, and might be improved by training longer with better-calibrated λs and further exploration into reward functions that more directly penalize inter-group disparities. A more thorough analysis of this subject would also be expanded to assess disparities across additional demographic groups (other racial categories, for example) rather than just White and Black racial categories.

7 Conclusion

In this work, we investaged whether reinforcement learning is capable of learning long-horizon lending policies that have higher fairness than static, rule-based policies. To test this, built a population-dynamics simulator and used real HMDA mortgage data to train PPO agents under three reward functions. These agents were compared to a profit-maximing and demographic parity baseline over 50 rounds.

We found that the PPO agents are able to consistently outperform both baselines on profit, and that agents trained with fairness penalizing at $\lambda = 1000$ achieve the best middle ground (14% improvement on profit and 11% improvement on reducing inter-group variance when compared to ProfitMax). However, the improvements on fairness are relatively marginal, and no policy is able to close the credit score gap between groups. Additionally, the policies perform notably worst than the DemographicParity baseline on fairness, instead sitting at a mid-point between `ProfitMax` and `DemographicParity` in terms of fairness scoring. The main takeaway from this is that RL is capable of finding a more central point on the tradeoff curve between profit and fairness than simple heuristic polices are, but the reward functions we used here were insufficient in overcoming the underlying gap.

These findings leave many paths open for future work, including experimenting with different reward functions that more directly target the gaps between groups rather than aggregate variance and the inclusion of more demographic groups. This work adds broadly to the growing body of evidence that long-horizon sequential learning can have tangible impacts on reducing algorithmic bias, with applications in lending environments.

8 Team Contributions

- **Naomi Boneh:** Naomi worked on implementing the infrastructure for the simulation environment, as well as training the PPO agents and testing lambda scaling in the reward functions. She generated Figure 2, which compares the baselines with the PPO agents. She redesigned the fairness metrics and reward functions to support six demographic groups instead of two, allowing for deeper insights into systemic inequality in lending, and also worked on the Methods and Results sections of the writeup.
- **Christelle Millos-Lopez:** Christelle implemented both non-RL baseline policies and generated Figure 1, showing the results from the baseline policies. She contributed to the theoretical design and project setup of the demographic group division and the tradeoff choices with the PPO agents. She worked on the Methods and Results sections, and wrote the Experimental Setup section.
- **Brydie Sigg:** Brydie contributed heavily to extracting and calibrating data from the HMDA dataset, and created the creditworthiness score estimate that is the foundation of measuring inequalities between demographic groups. She also contributed to the pipeline for lambda scaling. For the writeup, she wrote the Introduction, Related Works, Discussion, and Conclusion.

  All members contributed equally to the writeup and poster.

Changes from Proposal We worked collaboratively and synchronously on much of the project and split work as it came. Any adjustments from the original team contributions breakdown seen here were agreed upon by all group members, and we feel as though work was divided evenly amongst the team.

In terms of content, we have extended the project beyond our original proposal. First, we expanded the population from two racial groups to six groups defined by the intersection of race (non-Hispanic White and non-Hispanic Black) and income tier (Low, Middle, and Upper, based on the ratio of applicant income to FFIEC area median family income). This group discovers substantial inequalities visible in HDMA, where low-income White applicants and upper-income Black applicants experience different lending dynamics.

Moving from two to six groups required defining a different metric for fairness. Originally, we proposed using the intergroup gap, which would have worked for two groups. We changed to use the the population-weighted variance of group-level statistics for both reward functions. R_{approval} penalises the weighted variance of approval rates across the six groups, and R_{mean} penalises the weighted variance of group score means. Weighted variance reduces to the squared absolute gap in the two-group case, so this generalization preserves the original reward design and intent as well. All other elements of the proposal, including PPO with per-group threshold actions, group-level state, 50-round horizons, HMDA-calibrated initial distributions, and the two static baselines (ProfitMax and DemographicParity) are unmodified.

References

Eitan Altman. 1999. *Constrained Markov Decision Processes* (1 ed.). Routledge, Boca Raton. doi:10.1201/9781315140223

Consumer Financial Protection Bureau. 2025. Home Mortgage Disclosure Act (HMDA) Snapshot National Loan-Level Dataset: 2024. https://ffiec.cfpb.gov/data-publication/snapshot-national-loan-level-dataset/2024

Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. 2020. Fairness is Not Static: Deeper Understanding of Long Term Fairness via Simulation Studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, 525–534. doi:10.1145/3351095.3372878

Federal Financial Institutions Examination Council. 2026. Census and Demographic Data. https://www.ffiec.gov/data/census/overview

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2016/file/6a9659feb1216f14f7384ba499518b38-Paper.pdf

Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2017. Fairness in Reinforcement Learning. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*. PMLR, 1617–1626.

Sean Lewis-Faupel and Nicholas Tenev. 2024. Racial and Ethnic Disparities in Mortgage Lending: New Evidence from Expanded HMDA Data. https://arxiv.org/abs/2405.00895

Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed Impact of Fair Machine Learning. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*. PMLR, 3150–3158.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application* 8 (2021), 141–163. doi:10.1146/annurev-statistics-042720-125902

Anka Reuel and Devin Ma. 2024. Fairness in Reinforcement Learning: A Survey. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 1218–1230. doi:10.1609/aies.v7i1.31718

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 54)*. PMLR, 962–970.