Extended Abstract

Motivation Predictive policing systems increasingly use reinforcement learning (RL) to optimize patrol deployment based on historical crime data. Yet this data is systematically biased—reflecting patterns of over-policing in minority communities rather than actual crime rates. RL agents trained on such signals risk amplifying these biases through self-reinforcing feedback loops, raising urgent concerns about fairness and long-term harm. We address this challenge by investigating whether RL agents can be explicitly designed to detect crime effectively while avoiding the reinforcement of structural disparities.

Method We model patrol allocation as a Markov Decision Process (MDP), where agents distribute limited patrol units across zones to maximize detected crime. We compare three policy-gradient agents: (1) vanilla Proximal Policy Optimization (PPO), (2) PPO augmented with action normalization and entropy regularization to promote exploration and equitable allocations, and (3) Constrained Policy Optimization (CPO), which enforces explicit fairness constraints during learning. Fairness is evaluated using two novel metrics: outcome DMS, which measures allocation-crime mismatch across an episode, and procedural DMS, which captures unfairness over time with early-time weighting. These metrics also motivate the cost in the implementation of CPO.

Implementation We construct a custom simulation environment using two datasets: a synthetically skewed crime distribution and real arrest data from Oakland, CA. True crime rates are uniform and sampled from a Poisson distribution; observed crimes depend on patrol deployment and detection probability. Each agent is trained and evaluated over 365-day episodes. We report cumulative reward (crime detected) as the utility metric and use DMS scores to assess fairness. Baseline policies include a random patrol strategy and a greedy allocation based on prior observations.

Results In the biased synthetic setting, fairness-aware agents (Tuned PPO and CPO) achieve near-optimal rewards ($\approx$1528.6) while dramatically reducing disparities (Outcome DMS < 0.031), in contrast to the greedy and vanilla PPO agents, which show high disparity (Outcome DMS > 0.70) and poor performance. The random baseline performs well in fairness but lacks adaptive capacity. These patterns hold in the real-world environment, where fairness-aware agents remain competitive in reward and significantly outperform others on fairness metrics, demonstrating generalizability beyond synthetic conditions.

Discussion Our results show that RL agents trained on biased data without intervention tend to entrench disparities, while fairness-aware methods can proactively correct them—without sacrificing performance. Simple mechanisms like entropy regularization and normalization meaningfully improve fairness by encouraging exploration and breaking feedback cycles. Despite embedding fairness directly into the optimization process, CPO provides negligible gains and performance is equivalent to tuned PPO. These findings demonstrate that fairness is not a tradeoff but an achievable design goal in real-world RL applications.

Conclusion We provide empirical evidence that fairness-aware reinforcement learning is essential for responsible predictive policing. By equipping agents with mechanisms to unlearn historical bias, we enable systems that are both effective and equitable. Our results argue for integrating fairness objectives directly into RL design, especially in high-stakes, feedback-driven domains where algorithmic decisions shape future data.

Rapid Feedback Loop Mitigation for Fair Policing

Vyoma Raman
Department of Computer Science
Stanford University
vyoma@stanford.edu

Abstract

Reinforcement learning (RL) is increasingly used in predictive policing to optimize patrol deployment based on historical crime data. However, such data often reflects patterns of over-policing rather than true crime rates, creating feedback loops that can amplify bias. We model patrol allocation as a Markov Decision Process and evaluate three RL agents: standard PPO, PPO with entropy regularization and normalization, and Constrained Policy Optimization (CPO) with fairness constraints. To assess fairness, we introduce two Disparate Mistreatment Scores (DMS) that quantify allocation-crime mismatch over time. Using both synthetic and real-world data from Oakland, CA, we find that fairness-aware agents achieve near-optimal crime detection while substantially reducing disparities (Outcome DMS < 0.031). In contrast, standard PPO and greedy policies reinforce bias and underperform. Our results show that fairness can be effectively integrated into RL systems without sacrificing utility, offering a viable path toward more equitable decision-making in public safety applications.

1 Introduction

Predictive policing systems aim to optimize law enforcement deployment by learning patterns from historical crime data. However, these data are often shaped by longstanding racial and spatial biases, leading such systems to replicate and amplify existing disparities rather than mitigate them. Feedback loops—where biased patrol decisions influence future crime observations—compound this problem, locking models into distorted perceptions of crime distribution.

Despite growing critiques of predictive policing, most analyses stop at identifying bias without offering algorithmic corrections. Conversely, fairness in reinforcement learning (RL) has been studied in isolation, typically ignoring the recursive feedback dynamics that are central to policing contexts. This disconnect leaves a gap: existing fairness-aware RL frameworks are not equipped to handle the compounding bias introduced by policy–data interactions.

This work addresses that gap by developing and evaluating fairness-aware RL agents that explicitly mitigate feedback-driven disparities in patrol allocation. We simulate urban crime environments with biased historical data and compare three policy-gradient agents: standard PPO, PPO with normalization and entropy regularization, and Constrained Policy Optimization (CPO) with fairness constraints. To quantify fairness, we introduce two metrics—Outcome and Procedural Disparate Mistreatment Scores (DMS)—which measure spatial and temporal disparities between patrol allocation and true crime distribution.

Our central question is whether reinforcement learning agents can achieve high crime detection while avoiding reinforcement of historical bias. We find that fairness-aware methods significantly outperform standard and greedy approaches in both synthetic and real-world crime settings, demonstrating that equity and utility are not inherently at odds. This work contributes a principled, empirical

framework for evaluating and designing RL systems in socially sensitive, feedback-prone domains like predictive policing.

2 Related Work

Predictive policing systems have been widely deployed to allocate law enforcement resources using historical crime data, yet these systems often reinforce racial and spatial biases rather than mitigate them. Models like PredPol, described by Mohler et al. (2015), learn from arrest data to predict future crime locations and influence patrol deployments. However, such arrest data are deeply shaped by historical over-policing of communities of color, distorting the model's input and reinforcing biased patterns (Gilbertson, 2020). Empirical studies have shown that these systems disproportionately direct police attention toward minority neighborhoods, thereby amplifying disparities in surveillance and enforcement (Lum and Isaac, 2016; Ensign et al., 2018). For instance, by replicating PredPol's algorithm on Oakland drug crime data, Lum and Isaac (2016) expose a racialized discrepancy between predicted patrol intensities and actual drug use prevalence, revealing how feedback loops rooted in biased data can lead to disparate impact. Similarly, Ensign et al. (2018) apply fairness metrics like equal opportunity (Hardt et al., 2016) to demonstrate that even small initial disparities in historical data can cascade into significant predictive error and unfair policing outcomes when feedback effects are ignored.

In the reinforcement learning literature, fairness has received increasing attention, driven by the recognition that RL agents influence the environment through their actions, which in turn shape future observations and rewards. Reuel and Ma (2025) surveys existing approaches to fairness in RL, including multi-objective optimization, welfare-based objectives, action parity via Q-values, and calibration by group outcomes. One of the earliest treatments, Jabbari et al. (2017), defines fairness as taking similar actions for similar expected utility, while other approaches incorporate fairness constraints into policy optimization using techniques like actor-critic models or multi-objective MDPs (Reuel and Ma, 2025). Constrained Policy Optimization (CPO) (Achiam et al., 2017), originally developed for safe exploration, provides a general framework for satisfying fairness or safety constraints during learning and is particularly relevant in high-stakes settings such as predictive policing.

Despite these developments, there remains a critical gap at the intersection of predictive policing and fair reinforcement learning. Critiques of predictive policing systems generally do not incorporate rigorous fairness criteria into their analysis and never attempt to revise or improve the underlying modeling assumptions to mitigate feedback-driven bias. Conversely, most fairness-aware RL research neglects the recursive feedback effects introduced by biased historical data, an essential feature of the predictive policing setting. This disconnect leaves both fields ill-equipped to address the compounding harms that arise when biased decisions shape future data and policy. Closing this gap requires models that explicitly account for these feedback loops and embed tailored fairness constraints in a principled, dynamic framework.

3 Method

We formulate the patrol allocation task as a Markov Decision Process (MDP) defined by state space $\mathcal{S}$, action space $\mathcal{A}$, transition dynamics P, and reward function R. At each timestep t, the agent observes the state s_t, representing historical and current crime statistics across N zones, and selects an action a_t, allocating a fixed budget of patrol units across these zones. The environment then generates the true number of crimes in each zone by sampling from a Poisson distribution with a uniform mean λ, reflecting the underlying (unbiased) crime rate.

We compare three reinforcement learning agents, each built upon a policy-gradient foundation but differing in key regularization and constraint mechanisms. All agents employ neural network policy parameterizations and are trained using trajectories sampled from the environment as described above. We select on-policy approaches that learn after each timestep of interacting with the environment because that best reflects real-world conditions around predictive policing feedback loops.

3.1 Vanilla Proximal Policy Optimization (PPO)

The baseline agent uses the standard Proximal Policy Optimization (PPO) algorithm. The objective is to maximize the clipped surrogate objective:

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t \left[\min \left(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t \right) \right],$$

where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$, $\hat{A}_t$ is the estimated advantage, and ϵ is the clipping parameter. PPO is selected for its sample efficiency, robustness to noisy gradients, and reliable convergence properties in environments with partial observability and feedback loops.

3.2 PPO with Normalization and Tuned Entropy Bonus

The second agent extends vanilla PPO by incorporating output normalization and entropy regularization. The action vector, representing patrol allocations, is normalized to form a valid probability distribution over zones, ensuring that total allocations are feasible and discouraging degenerate solutions with highly concentrated patrols. The objective becomes:

$$L(\theta) = L^{\text{CLIP}}(\theta) + \beta\, \mathbb{E}_t \left[\mathcal{H}(\pi_\theta(\cdot|s_t)) \right],$$

where $\mathcal{H}$ is the entropy of the action distribution and β is an entropy coefficient. Entropy regularization encourages exploration, which is particularly important given the biased nature of the historical observations; it helps the agent discover and correct for spurious patterns caused by the initial data skew. See Appendix A for tuning details. Normalization ensures fair and interpretable allocation decisions throughout learning.

3.3 Constrained Policy Optimization (CPO)

The third agent employs Constrained Policy Optimization (CPO), which extends PPO with a formal constraint on expected policy-induced disparities. CPO solves:

$$\begin{aligned} &\max_\theta \ \mathbb{E}_{\pi_\theta} \left[\sum_{t=0}^{T} r_t \right] \\ &\text{subject to} \quad \mathbb{E}_{\pi_\theta} \left[\sum_{t=0}^{T} c_t \right] \leq \delta, \end{aligned}$$

where c_t represents the instantaneous fairness or disparity cost at time t, and δ is a developer-specified threshold. Policy updates are performed using a trust-region method to guarantee monotonic improvement in reward while maintaining constraint satisfaction. CPO is initialized with the same tuned entropy and normalization settings as the second agent, layering constraint-driven optimization on top of robust exploration. This approach provides explicit control over fairness-utility tradeoffs. We then tuned the threshold as shown in Appendix A. However, implementing cost based on the fairness metric requires insight into actual crime rate, which is not realistic in practice.

3.4 Design Rationales

The three agent variants are selected to systematically test the effect of increasingly sophisticated regularization and constraint mechanisms: vanilla PPO serves as a strong policy-gradient baseline, entropy-regularized PPO explicitly promotes exploration and distributional fairness, and CPO enforces provable fairness guarantees through constrained optimization. This suite of models enables a comprehensive analysis of learning dynamics and the interplay between reward maximization and fairness constraints in predictive policing contexts.

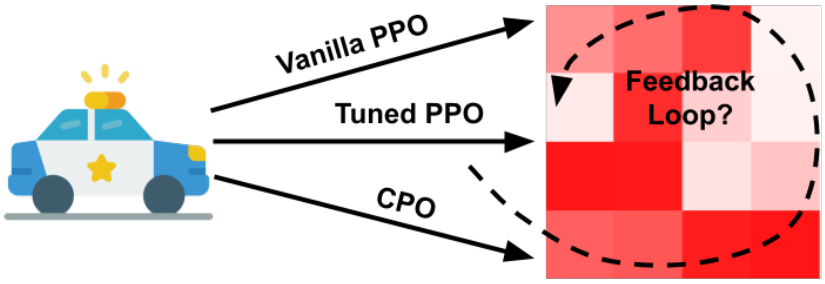


Figure 1: Patrol decisions based on observed crime data can create feedback loops, reinforcing or amplifying initial data biases. We attempt to build agents that unlearn this.

4 Experimental Setup

To evaluate the fairness and effectiveness of patrol allocation strategies produced by different agents, we developed a simulated predictive policing environment (Figure 1). This simulation models a simplified urban area divided into multiple zones, each subject to fluctuating crime activity over time.

4.1 Simulated Environment

The simulation environment initializes crime dynamics using two biased datasets of historical crime that overrepresent certain zones to reflect real-world disparities in arrest data resulting from over-policing. The number of zones in the environment corresponds to the length of the datasets. The first synthetic dataset contains 10 zones where the first three are highly skewed to contain more historical crime rate. The second dataset contains 59 zones with crime statistics from the past five years in Oakland, California.

At each timestep, the true number of crimes in each zone is generated independently by sampling from a Poisson distribution. The mean of this distribution is uniform across all zones and derived from the average crime rate in the historical data. This uniform true crime rate serves as a baseline to evaluate how effectively agents can detect crime while avoiding reinforcement of historical bias.

Agents interact with the environment by allocating a fixed number of patrol units across the zones. The probability of detecting at least one crime in a zone increases with the number of patrols assigned, following this formula:

$$P(\text{detect any crime}) = 1 - (1 - p)^n$$

where $p = 0.3$ is the base detection probability and n is the number of patrols in the zone. Detected crimes are drawn from a binomial distribution based on this detection probability and the true crime count.

The agent receives a reward at each timestep equal to the total number of crimes it observes, incentivizing policies that increase crime detection. This reward fits the context because actual crime is always unknown. Episodes run for 365 timesteps, simulating one year.

4.2 Fairness Metrics

To assess fairness, we introduce the Disparate Mistreatment Score (DMS), which measures the difference between the distribution of patrols and actual crime. DMS comes in two versions: outcome DMS and procedural DMS, which correspond to outcome fairness and procedural fairness.

Outcome DMS captures cumulative fairness by comparing the total fraction of patrols assigned to each zone with the actual fraction of crimes that occurred there across the final episode. The measure is computed as the total variation distance:

$$\text{DMS}_{\text{outcome}} = \frac{1}{2} \sum_{z} |P_{\text{alloc}}(z) - P_{\text{crime}}(z)|$$

A lower score indicates a more equitable match between patrol resources and real crime needs.

Procedural DMS evaluates fairness over time, incorporating a temporal discount factor λ (defaulting to 0.6) to prioritize early correction of unfairness. At each timestep, a variation distance is computed as in the outcome metric, but discounted to emphasize early-time behavior. The procedural DMS is aggregated over time and averaged across episodes:

$$\text{DMS}_{\text{procedural}} = \sum_{t=0}^{T} \gamma^t \cdot \left(\frac{1}{2} \sum_{z} |P_{\text{alloc}_t}(z) - P_{\text{crime}_t}(z)| \right)$$

This metric allows us to distinguish between agents that gradually adjust their behavior and those that enforce fairness from the outset.

4.3 Baseline Methods

To contextualize the performance of our reinforcement learning agents, we implement two baseline patrol allocation strategies: a random baseline and a greedy baseline.

Random Baseline. The random baseline uniformly allocates patrols across all zones at each timestep, independent of both historical and current crime observations. This approach serves as a fairness-oriented benchmark: because patrols are distributed without regard to observed crime patterns (which may be biased), this policy avoids reinforcing feedback loops and, in expectation, matches the true uniform distribution of crime. However, it does not adapt to any potential spatial or temporal variation in true crime rates.

Greedy Baseline. The greedy baseline allocates patrols in direct proportion to the most recently observed crime counts in each zone. At each timestep, the agent assigns more patrols to zones where more crimes were detected in the previous step. While this approach aims to maximize immediate observed crime detection, it is highly sensitive to initial data bias and can exacerbate disparities. The greedy policy typifies the logic of many real-world predictive policing deployments, which risk amplifying existing feedback loops and producing unfair outcomes.

These baselines provide useful reference points: the random baseline captures the lower bound of intervention, while the greedy baseline demonstrates the pitfalls of naive data-driven allocation without fairness considerations.

5 Results

Overall, we find that randomized and fairness-aware patrol allocation methods (Tuned PPO, CPO) robustly outperform naive and greedy approaches on both fairness and performance, in both synthetic and real environments. By visualizing DMS metrics and reward tradeoffs, we also highlight the necessity and effectiveness of explicit bias mitigation in RL for predictive policing.

5.1 Quantitative Evaluation

We evaluate each model on both a synthetically skewed dataset and real crime data from Oakland. Metrics include average reward, outcome DMS, and procedural DMS (see Tables 1 and 2).

In the skewed environment, the random baseline achieves the highest reward (1530.77) and lowest disparity (Outcome DMS = 0.00061, Procedural DMS = 0.00034). Despite being uninformed, its uniform allocation avoids reinforcing bias. In contrast, the greedy baseline earns a low reward (482.09) and exhibits high disparity (Outcome DMS = 0.700, Procedural DMS = 0.00196), illustrating how naive data-driven methods entrench existing bias. Vanilla PPO performs similarly poorly, with the lowest reward (202.67) and highest Outcome DMS (0.776).

Fairness-aware methods significantly improve performance. Tuned PPO achieves a near-optimal reward (1528.64), while reducing Outcome DMS to 0.0308 and Procedural DMS to 0.00035. CPO

Table 1: Skewed Environment Metrics

Method	Reward	Outcome DMS	Procedural DMS
Random Baseline	1530.77	0.00061	0.00034
Greedy Baseline	482.09	0.700	0.00196
Vanilla PPO	202.67	0.776	0.00243
Tuned PPO	1528.64	0.0308	0.00035
CPO	1528.60	0.0293	0.00035

Table 2: Actual Environment Metrics

Method	Reward	Outcome DMS	Procedural DMS
Random Baseline	1100.41	0.00084	0.00044
Greedy Baseline	1118.29	0.00076	0.00047
Vanilla PPO	466.08	0.49666	0.00198
Tuned PPO	1098.24	0.03303	0.00044
CPO	1097.82	0.03217	0.00044

yields comparable results (Reward = 1528.60, Outcome DMS = 0.0293, Procedural DMS = 0.00035), validating that fairness and utility are not inherently in conflict.

In the real-data environment, reward trends shift due to non-uniform true crime rates. The greedy baseline now performs well in reward (1118.29) but maintains similar disparity (Outcome DMS = 0.00076). Vanilla PPO again underperforms in reward (466.08) and exhibits the highest disparity (Outcome DMS = 0.49666). Tuned PPO (Reward = 1098.24, Outcome DMS = 0.03303) and CPO (Reward = 1097.82, Outcome DMS = 0.03217) remain both fair and effective, with procedural DMS near 0.00044 in both cases. These results demonstrate that fairness-aware methods generalize robustly across environments.

5.2 Qualitative Analysis

Figure 2 compares outcome and procedural DMS across models. In both environments, the random, tuned PPO, and CPO agents achieve extremely low DMS values. By contrast, vanilla PPO and greedy baselines show large disparities—particularly in the skewed case, where vanilla PPO exceeds 0.75 in outcome DMS and 0.0024 in procedural DMS. Tuned PPO and CPO maintain procedural DMS near 0.00035 across both settings, indicating early and sustained fairness corrections.

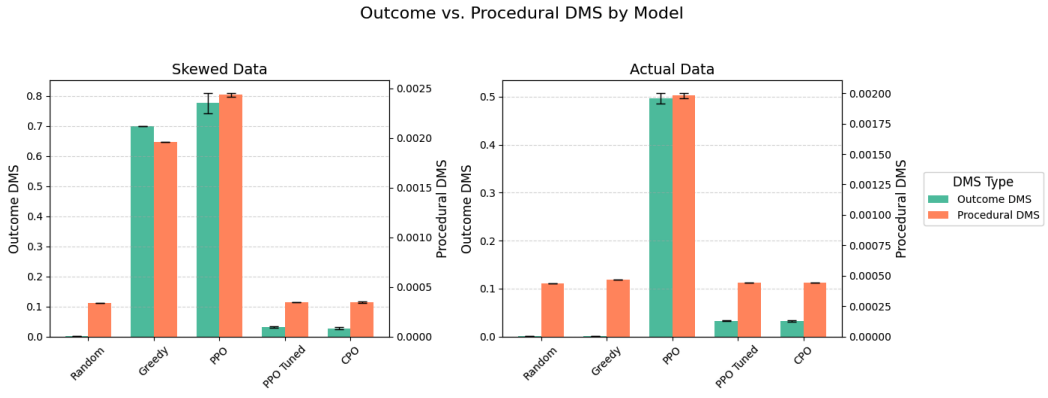


Figure 2: Comparison of Outcome and Procedural DMS across models in both the skewed and actual data environments. Fairness-aware agents (Tuned PPO, CPO) exhibit significantly lower disparity across both metrics, while naive agents (Greedy, Vanilla PPO) amplify bias—especially in the skewed setting.

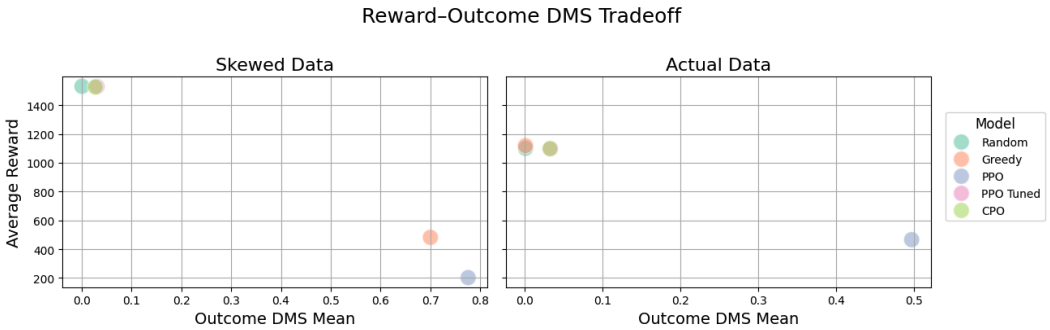


Figure 3: Reward vs. Outcome DMS tradeoff for all models in skewed and actual data environments. Only fairness-aware agents (Tuned PPO, CPO) and the Random baseline achieve both low disparity and high reward. Greedy and Vanilla PPO occupy poor tradeoff regions, highlighting the cost of reinforcing biased observations.

Figure 3 visualizes the reward–DMS tradeoff. In the skewed case, only the random, tuned PPO, and CPO agents achieve both high reward and low disparity. The greedy and vanilla PPO models lie far to the right, demonstrating their poor fairness–utility tradeoff. In the actual environment, the separation is more compressed, but fairness-aware agents continue to dominate the top-left of the tradeoff space. These visuals reinforce the quantitative findings: fairness-aware RL agents can effectively balance equity and performance, while naive methods systematically fail to do so.

5.3 Model-Level Takeaways

The random agent achieves strong fairness by avoiding feedback-driven allocation, and performs surprisingly well in the uniform-crime setting. The greedy agent performs poorly on skewed data due to bias lock-in but improves on actual data where observed and true crime distributions partially align, though at the cost of persistent disparities. Vanilla PPO lacks any corrective mechanism for biased rewards and quickly converges to unfair policies, leading to poor performance across metrics. Tuned PPO mitigates this through action normalization and entropy regularization, promoting exploration and reducing over-allocation to biased zones, which results in low DMS and high reward. CPO further improves fairness by imposing explicit disparity constraints, maintaining competitive reward while ensuring robust fairness across settings.

6 Discussion

Our findings highlight the importance and effectiveness of fairness-aware reinforcement learning in predictive policing contexts. Without explicit interventions, even advanced agents like PPO are prone to amplifying feedback loops present in biased observational data. This is particularly evident in the skewed environment, where both vanilla PPO and the greedy baseline reinforce historical disparities and perform poorly in terms of both fairness and utility.

By contrast, we demonstrate that fairness interventions—through action normalization, entropy regularization, and constrained optimization—can break this cycle. Tuned PPO and CPO not only reduce disparate treatment but also maintain near-optimal reward levels, even under challenging conditions. These results underscore that fairness and performance are not inherently in conflict; with proper algorithmic design, both objectives can be achieved.

Our evaluation on actual crime data from Oakland further supports the practical value of these approaches. Despite the shift to non-uniform true crime rates, fairness-aware agents continue to perform competitively, validating their robustness and real-world relevance.

Limitations. This work relies on simulated crime dynamics and fairness proxies, which do not capture the full complexity of real-world policing impacts or community harm. While the procedural DMS metric captures temporal fairness, it remains an aggregate measure and does not account for compounding effects across demographic or geographic groups.

7 Conclusion

This work demonstrates that fairness-aware reinforcement learning strategies can significantly reduce disparities in predictive policing without compromising performance. In both biased synthetic environments and real crime data from Oakland, fairness-driven agents—particularly those using normalization, entropy regularization, or explicit constraints—consistently achieve high crime detection while avoiding reinforcement of data-driven bias. By contrast, naive approaches like greedy allocation or standard PPO amplify existing disparities, especially under skewed data conditions.

Our results support a broader conclusion: fairness should not be treated as an afterthought in policy learning systems, particularly in domains involving historically marginalized communities. Explicit design interventions—such as fairness constraints and exploration incentives—are essential to prevent unintended harms. While limitations remain in terms of simulation realism and metric scope, this work provides strong evidence that fairness-aware RL is both feasible and necessary for responsible AI in public safety applications.

8 Team Contributions

As indicated in the project proposal, the author completed this project independently.

References

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. 2017. Constrained Policy Optimization. In *Proceedings of the 34th International Conference on Machine Learning*. PMLR, 22–31. https://arxiv.org/abs/1705.10528

Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. Runaway Feedback Loops in Predictive Policing. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 160–171. https://proceedings.mlr.press/v81/ensign18a.html

Annie Gilbertson. 2020. Data-Informed Predictive Policing Was Heralded As Less Biased. Is It? *The Markup* (2020). https://themarkup.org/the-breakdown/2020/08/20/does-predictive-police-technology-contribute-to-bias

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

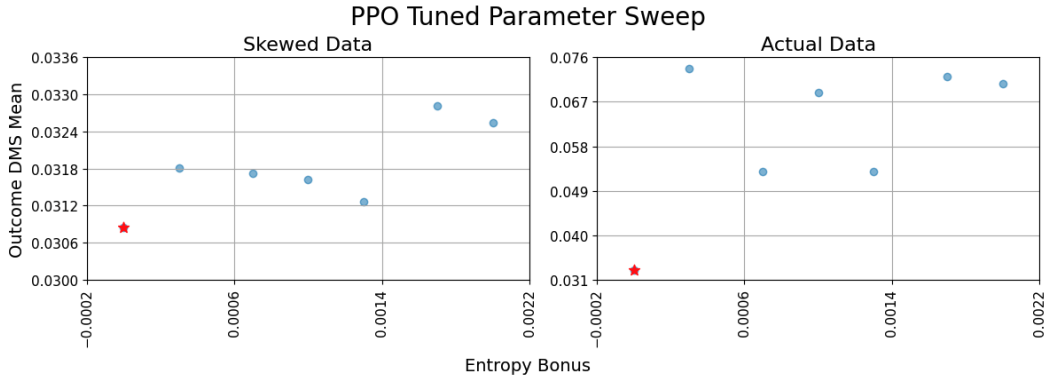
Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2017. Fairness in Reinforcement Learning. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 1617–1626. https://proceedings.mlr.press/v70/jabbari17a.html

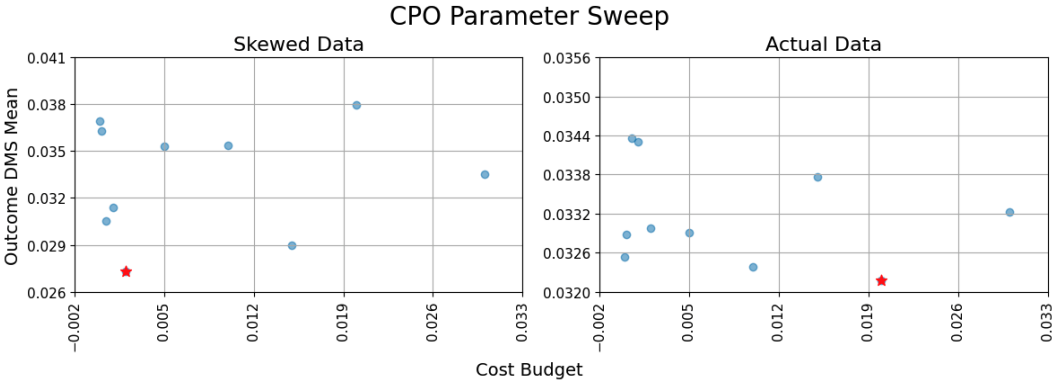
Kristian Lum and William Isaac. 2016. To Predict and Serve? *Significance* 13, 5 (2016), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x arXiv:https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1740-9713.2016.00960.x

George Mohler, M. Short, Sean Malinowski, Mark Johnson, George Tita, Andrea Bertozzi, and P. Brantingham. 2015. Randomized Controlled Field Trials of Predictive Policing. *J. Amer. Statist. Assoc.* 110 (10 2015), 00–00. https://doi.org/10.1080/01621459.2015.1077710

Anka Reuel and Devin Ma. 2025. Fairness in Reinforcement Learning: A Survey. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society* (San Jose, California, USA) *(AIES '24)*. AAAI Press, 1218–1230.

A Hyperparameter Sweep


PPO Tuned Parameter Sweep
Skewed Data
Actual Data
Outcome DMS Mean
0.0336
0.0330
0.0324
0.0318
0.0312
0.0306
0.0300
0.076
0.067
0.058
0.049
0.040
0.031
−0.0002
0.0006
0.0014
0.0022
Entropy Bonus


CPO Parameter Sweep
Skewed Data
Actual Data
Outcome DMS Mean
0.041
0.038
0.035
0.032
0.029
0.026
0.0356
0.0350
0.0344
0.0338
0.0332
0.0326
0.0320
−0.002
0.005
0.012
0.019
0.026
0.033
Cost Budget