Policy Optimization with Advantage Regularization for Long-Term Fairness in Decision Systems

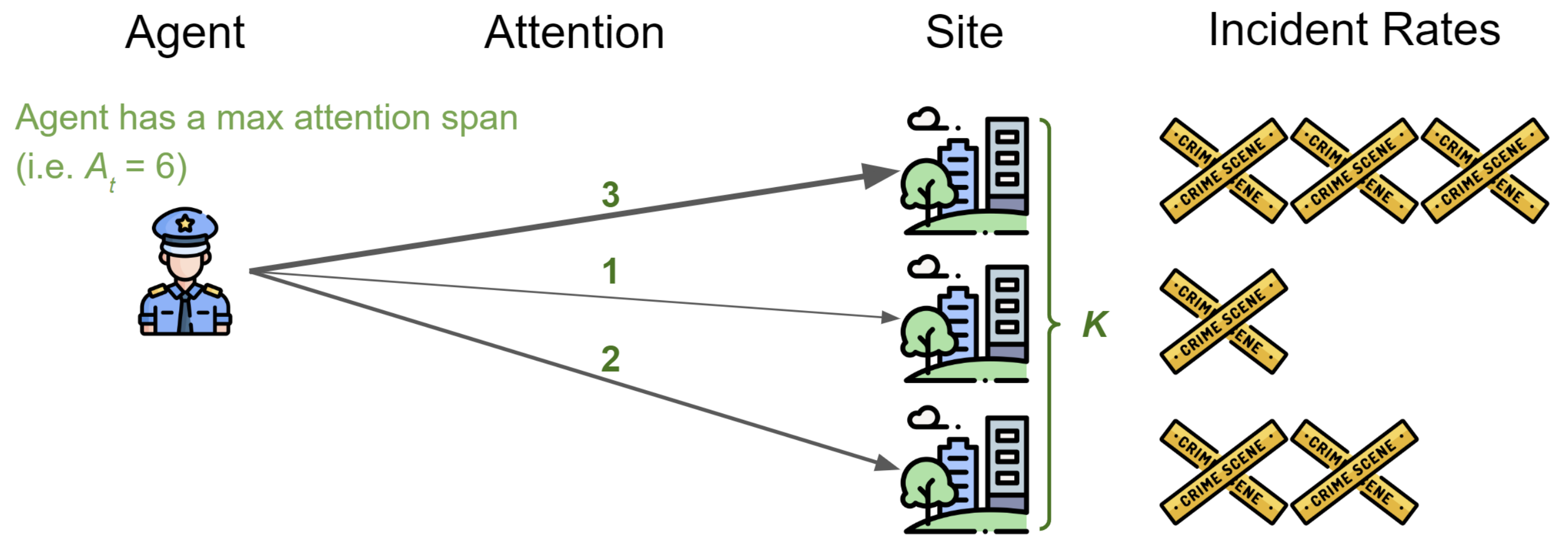
Eric Yang Yu[1], Zhizhen Qin[1], Min Kyung Lee[2], Sicun Gao[1]

[1]UC San Diego [2]UT Austin

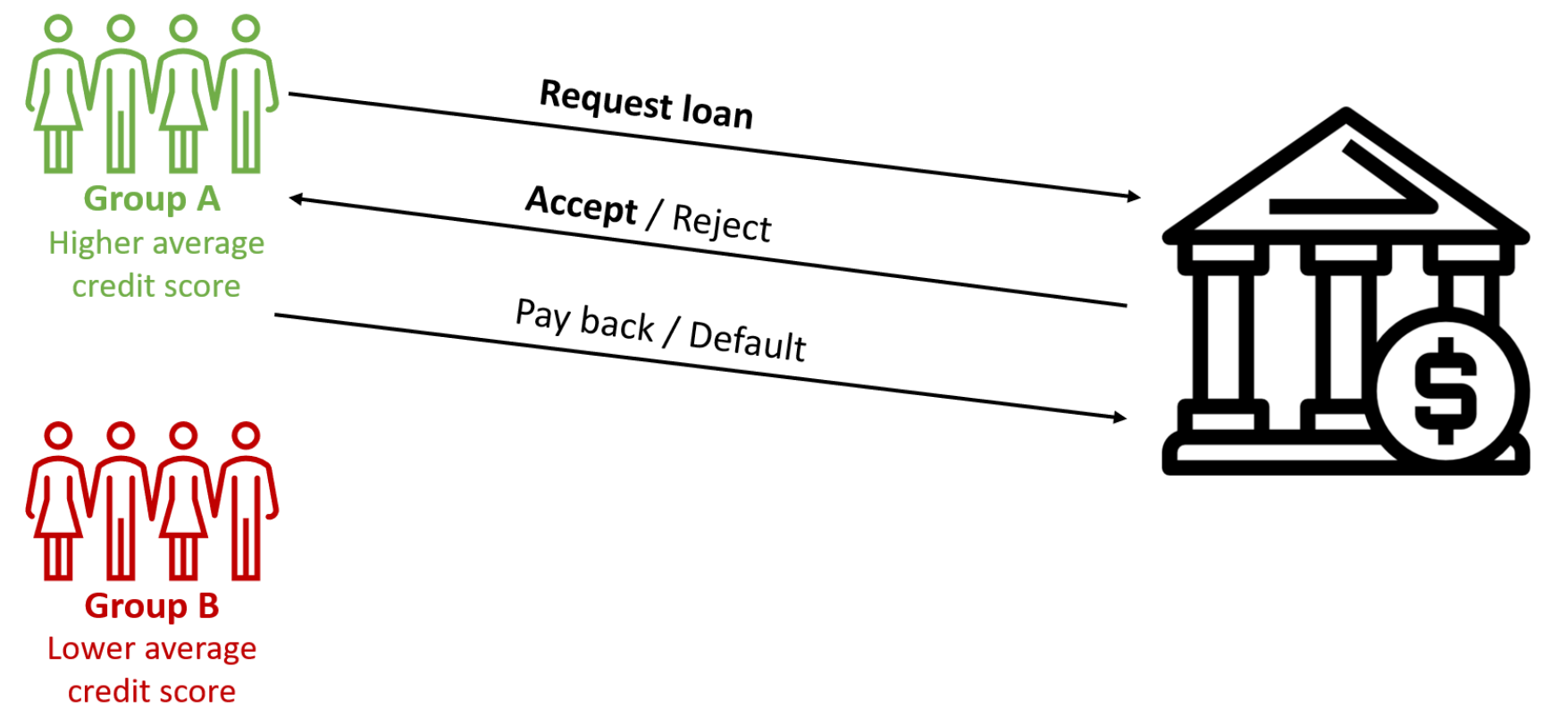
NEURAL INFORMATION PROCESSING SYSTEMS

Background and Motivation

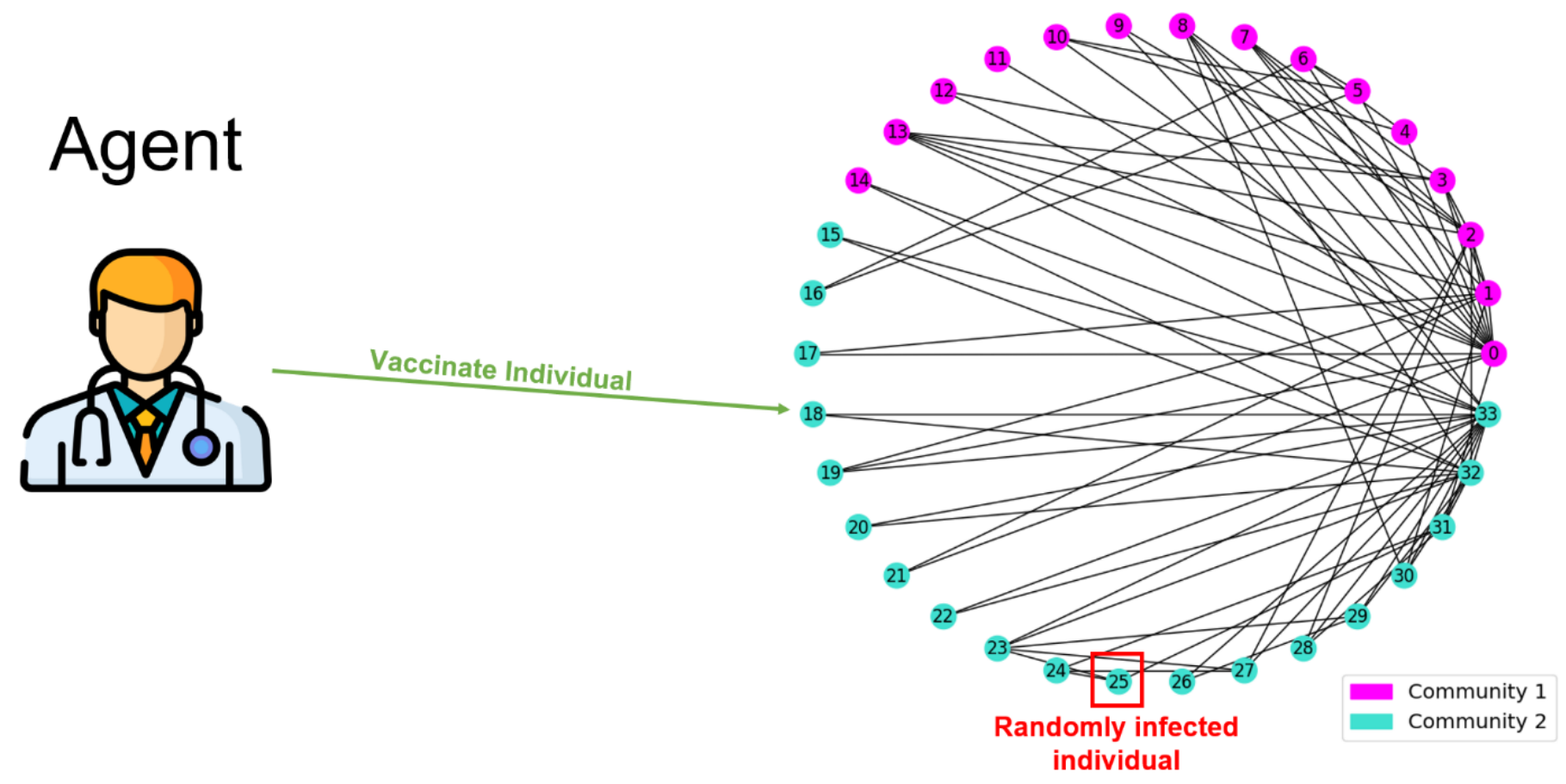
- Most of the fairness literature studies fairness in a ***one-shot context***
- However, maximizing immediate fairness may be detrimental for ***long-term fairness***
- To study this phenomenon, D'Amour et. al. 2020[1], is the first to formulate long-term fairness simulations as MDPs, in the following case studies:
- Attention Allocation



- Lending



- Vaccine Distribution[2]



- We can use constrained reinforcement learning (CRL) to solve these environments, but current CRL techniques can be computationally expensive and unreliable
- We want a **cheaper** and **more reliable** constrained RL approach for addressing long-term fairness

Intuitive but Naïve Constrained RL Approach

- Intuitively, we can model a fairness requirement as a penalty term in the objective function:
- $J^{fair}(s_t) = \omega_0 J(s_t) - \omega_1 \Delta(s_t)$, where Δ measures "unfairness" and ω balances the objectives
- However, this introduces 2 problems:

1. Requires reward engineering (hard to justify trade-offs between objectives)
2. May incentivize the agent to perform reward hacking

Advantage Regularization in Policy Optimization

Let $\Delta : S \rightarrow \mathbb{R}^{\geq}$ be a measure of fairness constraint violation. Then:

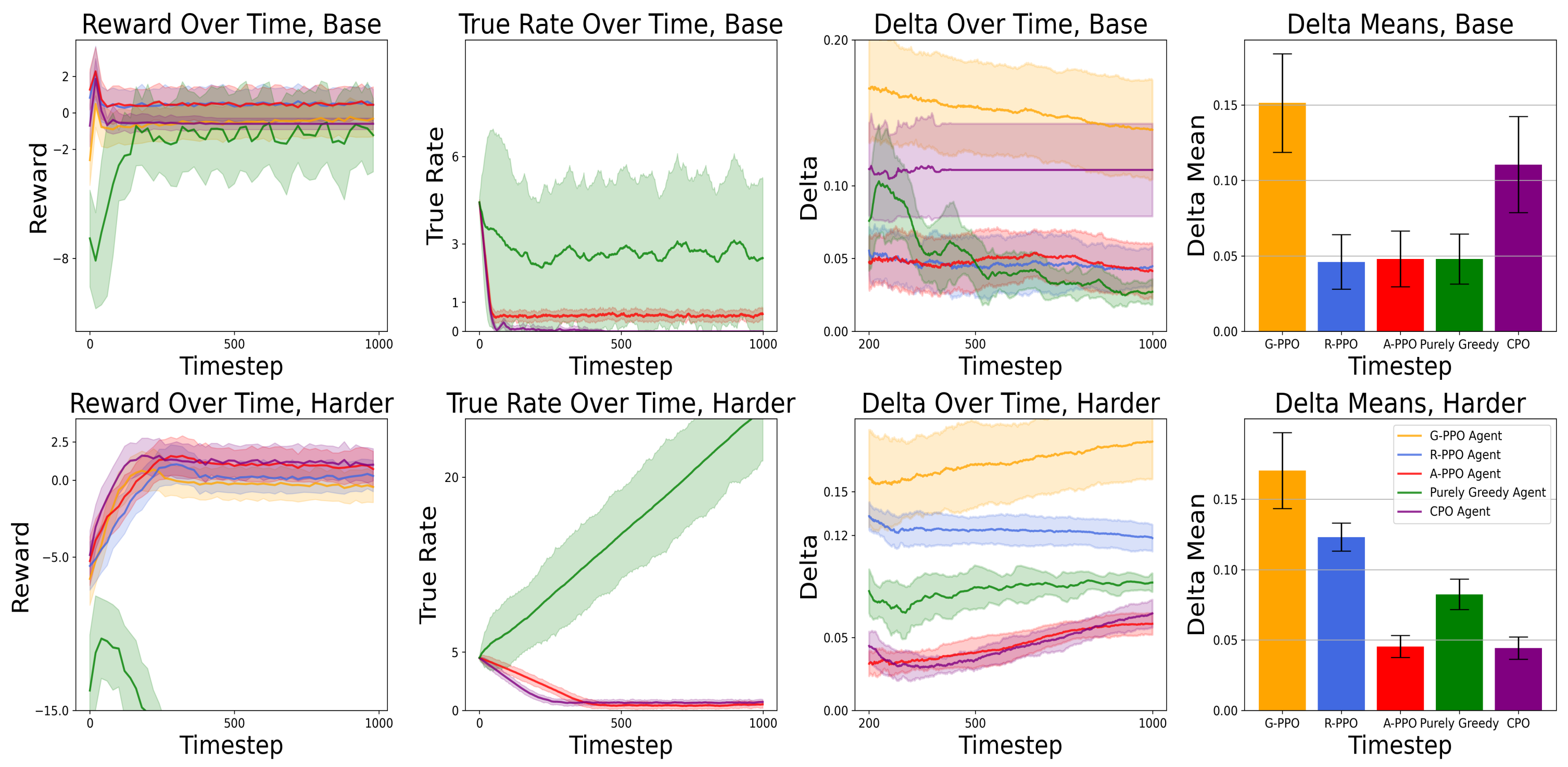
$$\hat{A}_\beta(s_t, a_t) = \beta_0 \hat{A}(s_t, a_t) + \underbrace{\beta_1 \min(0, -\Delta(s_t) + \omega)}_{\text{Value-thresholding term}} + \underbrace{\beta_2 \begin{cases} \min\big(0, \Delta(s_t) - \Delta(s_{t+1})\big) & \text{if } \Delta(s_t) > \omega \\ 0 & \text{otherwise} \end{cases}}_{\text{Decrease-in-violation term}}$$

Our Contributions

1. We show that RL approaches are effective for designing policies that can achieve long-term fairness, whereas existing heuristic and rule-based approaches do not perform well
2. We propose novel methods for imposing fairness requirements by regularizing the advantage evaluation in policy optimization

- Specifically, our method modifies the "how-to-optimize" aspect of RL (i.e. the algorithm), rather than the "what-to-optimize" aspect (i.e. the objective)

Results



In the Attention Allocation environment, our method (**A-PPO**) trains ~2-3x faster than our closest constrained RL baseline (**CPO**)

Future Work

- Extend this framework to other policy optimization methods
- Investigate a harder-constrained RL method with a reasonable computational complexity trade-off (i.e. perhaps modeling the environment as a Constrained MDP (CMDP), taking a stricter "stability control-theoretic" approach to the problem, etc.)
- Publish a set of baseline environments for evaluating long-term fairness

References

1. Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. Fairness is not static: deeper understanding of long term fairness via simulation studies. In Mireille Hildebrandt, Carlos Castillo, L. Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna, editors, FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020, pages 525–534. ACM, 2020.
2. James Atwood, Hansa Srinivasan, Yoni Halpern, and David Sculley. Fair treatment allocations in social networks. CoRR, abs/1911.05489, 2019.