UC San Diego
JACOBS SCHOOL OF ENGINEERING
Computer Science and Engineering

Assessing Fairness through Standardized Environments

Cassandra Ponce Maldonado, Josue Martinez, Suprith Krishnakumar, Xicotencatl Reyes
Advisors: Zhizhen Qin, Eric Yu, Sicun Gao

ERSP
UC San Diego
Early Research Scholars Program

Background

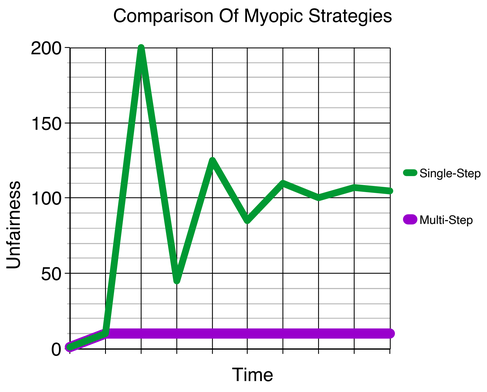
- Machine Learning algorithms are being deployed in high-stake decision-making systems:

Health Loan-lending Crime

- There is a concern regarding the fairness of these algorithms to make sure they make optimal decisions while being fair to different populations.
  - Eg. income disparity from unfair loan lending

Motivation

- Fairness is mostly studied in single-step contexts in recent literature.
- There are **harmful long-term implications** of strategies that **maximize immediate fairness**[1].



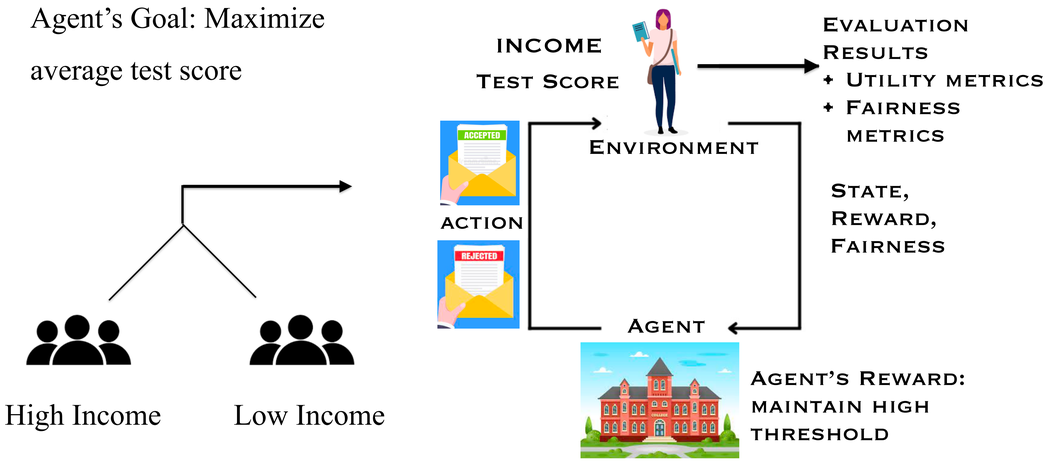
- **Lack of standardized testing** to assess long-term fairness of recent ML algorithms.
- Prior work has established baselines for testing that simulate loan lending, vaccine distribution and attention allocation.

Problem Statement

- We seek to add to the set **standardized** environments that provide concrete **fairness metrics** and plots to compare and assess performance of different algorithms.

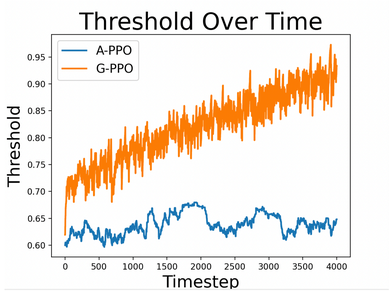
Our Approach

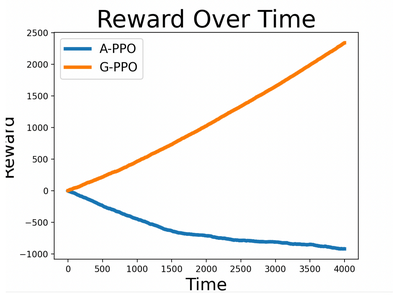
- Simulate the role that income and economic burden has in in the college admission process.
- Agent's Goal: Maximize average test score

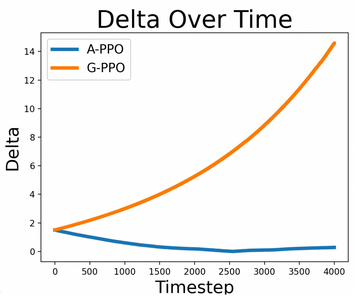


- Fairness, in this scenario, is measured as the income gap between these two groups over time

Results







- G-PPO: maximizes reward, a function of the average score of its attendees.
- A-PPO: maximizes reward whilst minimizing a fairness constraint: the income disparity between our economically advantaged and disadvantaged groups.

Conclusion

- **Our environment successfully tested known algorithms, producing expected behavior.**

- Thus, by increasing the number of baselines environments within the ML community, we allow for the long-term fairness of many algorithms to be tested in a standardized dynamic setting.
- This type of testing is necessary before ML algorithms can be deployed in very complex, real-life settings.

Future Work

- Our baseline can apply fairness evaluations to other sequential decision-making algorithms
- Further improve our current environment to better represent the social science between college admission and expected income
- Continue to build sets of baseline environments for the Machine Learning community that access long-term fairness in this dynamic way

Acknowledgments

- We'd like to thank Professor Sicun Gao, Zhizhen Qin, and Eric Yu as well as Professor Mai ElSherief and Vaidehi Gupta.