Bailey Osweiler, Rachel Goodman, Tory Farmer
DCDS 500 - Project 1
03/05/24

Project 1 - Algorithmic Fairness in College Admissions

Undergraduate applications and admissions are an important milestone in students' lives, and decisions have far-reaching impacts beyond where students live and study for the next four years. College admissions offices want to identify the most qualified students for their college, however, admissions criteria can have unintended consequences due to systemic barriers and discrimination. Not all students have equal access to college preparatory resources, and, of greater concern, students may be qualified for college but unable to construct a competitive college application. Thus, admissions criteria have the potential to exacerbate or alleviate disparities during this important academic hurdle.

Part 1 - Simulating College Admissions

We simulated the college admissions decision process to better understand how colleges can avoid perpetuating disparities, and the potential impacts on admitted student qualifications. We made some assumptions about the applicant pool and considered four potential admissions strategies for comparison.

Assumptions:

- 100 students applied to this university
    - 50 belong to a disadvantaged group, and 50 belong to an advantaged group
- Admissions cares about two academic success indicators: X1 and X2
- Each student has probability of a positive trait for X1 and X2, and traits are independent
    - Advantaged group has $P = 2/3$ for each trait
    - Disadvantaged group has $P = 1/3$ for each trait
- The admissions office has the following information about a given applicant
    - X1 outcome (binary) - a measurable academic success indicator (perhaps satisfactory GPA, standardized test scores, or transcripts)
    - X2 outcome (binary) - a measurable academic success indicator (perhaps satisfactory volunteer hours, part-time jobs, or personal essays)
    - Group membership (binary)
- The admissions office knows the expected value of academic fitness function $f = X1 * X2$ for each group at an aggregated level

Potential Admissions Strategies:

1. Consider both X1 and X2. Admit 28 students with both traits.
2. Consider only X1. Admit 28 students with this trait.
3. Consider X1 and G. Admit the top 28 students with X1, maximizing $E(f \mid G)$.

4. Consider X1 only from the advantaged group, and X1 and X2 from the disadvantaged group. Prioritize the most qualified disadvantaged applicants with both features (X1, X2). Then, admit students with positive X1 traits randomly until 28 total students have been admitted.

Evaluation Criteria:

Two quantities – equity and efficiency – indicate the performance of each potential admissions strategy. **Equity** measures demographic fairness and is defined as the percent of admitted students who belong to the disadvantaged group. **Efficiency** measures admitted students' academic qualifications and is defined as $average(f)$ among admitted students.

Simulation approach:

To simulate the applicant pool, we created a pandas dataframe to store 100 applicants. Each applicant was assigned a group membership. We created random numbers for each applicant and if the random number was below the indicated (X1, X2) probabilities for the applicant's group, they received a positive X1 or X2 trait, otherwise a negative trait. We shuffled the dataframe to randomize tiebreakers.

To simulate admissions approaches, we created functions for each admissions criteria scenario to choose the 28 most qualified applicants using the available information. We repeated each scenario 100 times and averaged the equity and efficiency scores over all repetitions to minimize the impacts of randomness.

Comparison of Scenario Performance: (See Figure 1)

1. Consider X1 and X2
    - Efficiency: 0.93, Equity: 0.22
2. Consider only X1
    - Efficiency: 0.54, Equity: 0.32
3. Consider X1 and G
    - Efficiency: 0.65, Equity: 0.001
4. Consider X1 and X2 for disadvantaged; only X1 for advantaged
    - Efficiency: 0.59, Equity: 0.39

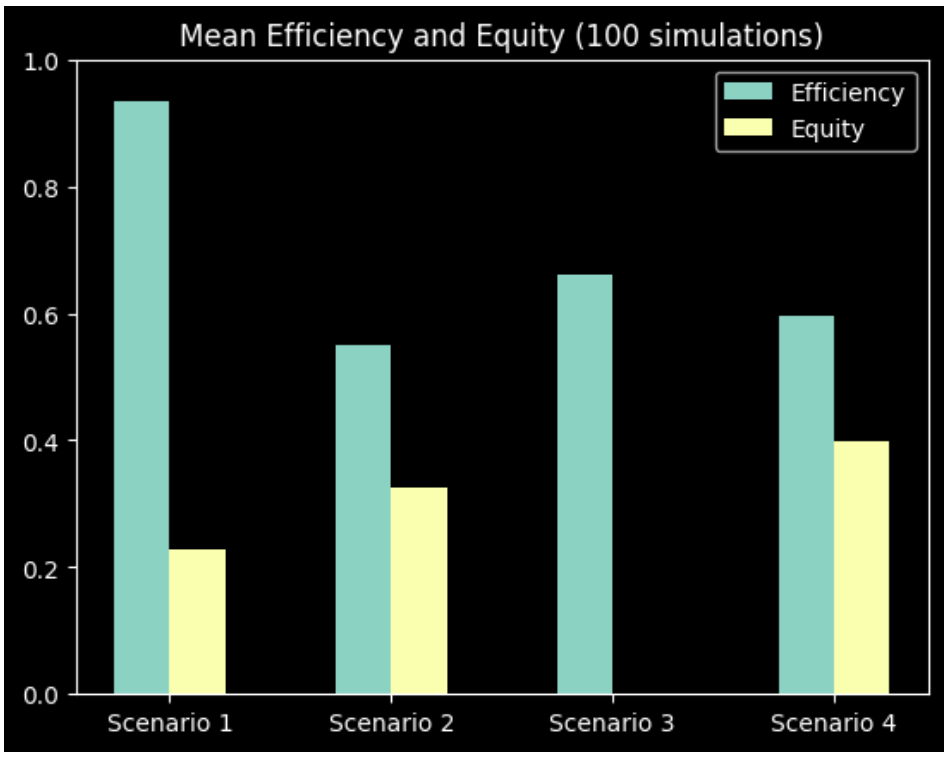


Figure 1

Using the above equity and efficiency calculations, and knowing that the admissions office's goal is to identify a scenario that maximizes efficiency, equity, or both, we can visualize their options using the below table. If the admissions office's **primary goal is to maximize efficiency**, they should look along the bottom row of Table 1; Scenario 1 provides the upper bound of efficiency. In contrast, **if their primary goal is to admit diverse students** and counteract the systemic discrimination against disadvantaged students, they should maximize equity. Their best options are in the far right column of the table - Scenario 2 or 4. To identify a solution which has high results for both equity and efficiency, they may consider Scenario 4 ideal, **but cost may also be a factor**. Scenario 4 incurs higher costs than Scenario 2 because they must collect X2. If these costs are unreasonably burdensome, Scenario 2 provides a good alternative with reasonably high equity, reasonably high efficiency, and low cost.

Table 1	**Lowest Equity**	→	**Highest Equity**
Lowest Efficiency			
↓			Scenario 2
	Scenario 3		Scenario 4*
Highest Efficiency		Scenario 1	

* increased cost

Definitions of Fairness: Theoretical backings in algorithmic fairness may also help determine the ideal approach for college admissions. Mehrabi et al.[1] summarize nine definitions of algorithmic fairness in their paper, *A Survey on Bias and Fairness in Machine Learning.* Considering a few of these theoretical definitions in addition to empirical simulation results may illuminate the best admissions approach.

Fairness through Unawareness considers an algorithm fair if it does not explicitly use protected information in decision making. In college admissions, any approach which does not consider students' advantaged or disadvantaged group is fair. Scenarios 1 and 2 meet this criteria of fairness, but because 3 and 4 use group membership as an input, they are not "fair." Additionally, the correlation between protected traits (group status) and negative admissions criteria (X1, X2) brings into question whether this definition is sufficient.

Counterfactual Fairness defines an algorithm as fair if the classification decision would be the same if the individual had the same feature values, but their protected identity feature changed. We can imagine a counterfactual world where a student has the same X1 and X2 traits but belongs to the opposite group, and their admissions outcome should be the same. Figure 2 summarizes results from a simulation that switched each applicant's group and then evaluated the admissions decision in both "worlds". The counterfactual fairness definition is not met for Scenarios 3 and 4 - approximately 25% of admissions decisions in each scenario do, in fact, change when the group changes.

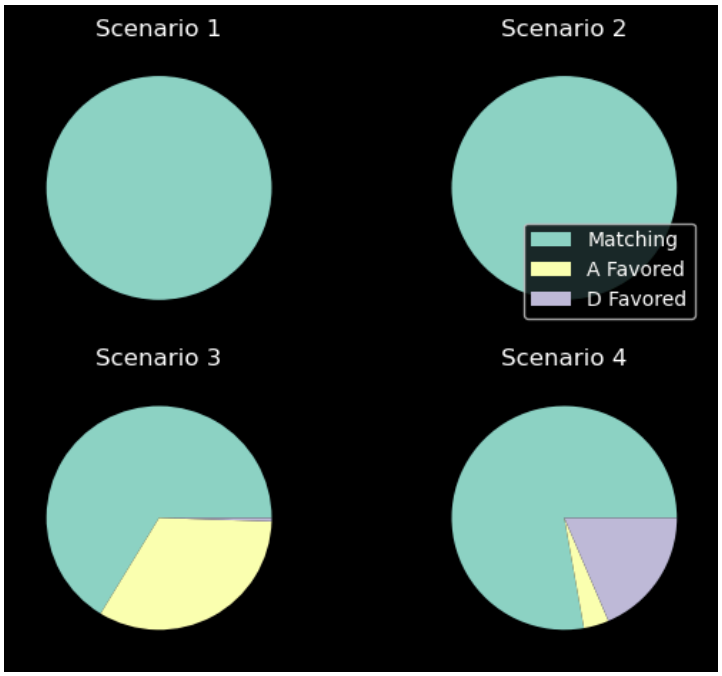


Figure 2

Demographic Parity means that students of under-resourced or underrepresented groups should have the same probability of a positive outcome as the advantaged group. Our equity metric is an exact measure of demographic parity, where 50% would be fair. Because the highest equity measure is Scenario 4 where equity = 0.39, none of these scenarios are considered fair by demographic parity.

Equal Opportunity means that the probability of a positive outcome should be the same for people with positive traits, regardless of their protected characteristics. "Positive traits" could mean either X1 or (X1 and X2), and fair under this definition means that

$$P(Accepted \mid (X1 \ \& \ G \ == \ Advantaged)) \ = \ P(Accepted \mid (X1 \ \& \ G \ == \ Disadvantaged))$$

Equation 1

Or, under a more strict definition of "positive trait",

$$P(Accepted \mid (X1 \ \& \ X2 \ \& \ G \ == \ Advantaged)) \ = \ P(Accepted \mid (X1 \ \& \ X2 \ \& \ G \ == \ Disadvantaged))$$

Equation 2

In Figure 3 below, we plot the probability of being accepted in the advantaged group minus the probability of being accepted in the disadvantaged group. In a perfectly fair scenario, the boxplot would be centered on the blue line, where Equations 1 or 2 would hold true. In the first definition of "positive" trait, we see that only scenario 2 is fair. For the second definition, both Scenarios 1 and 2 are fair.

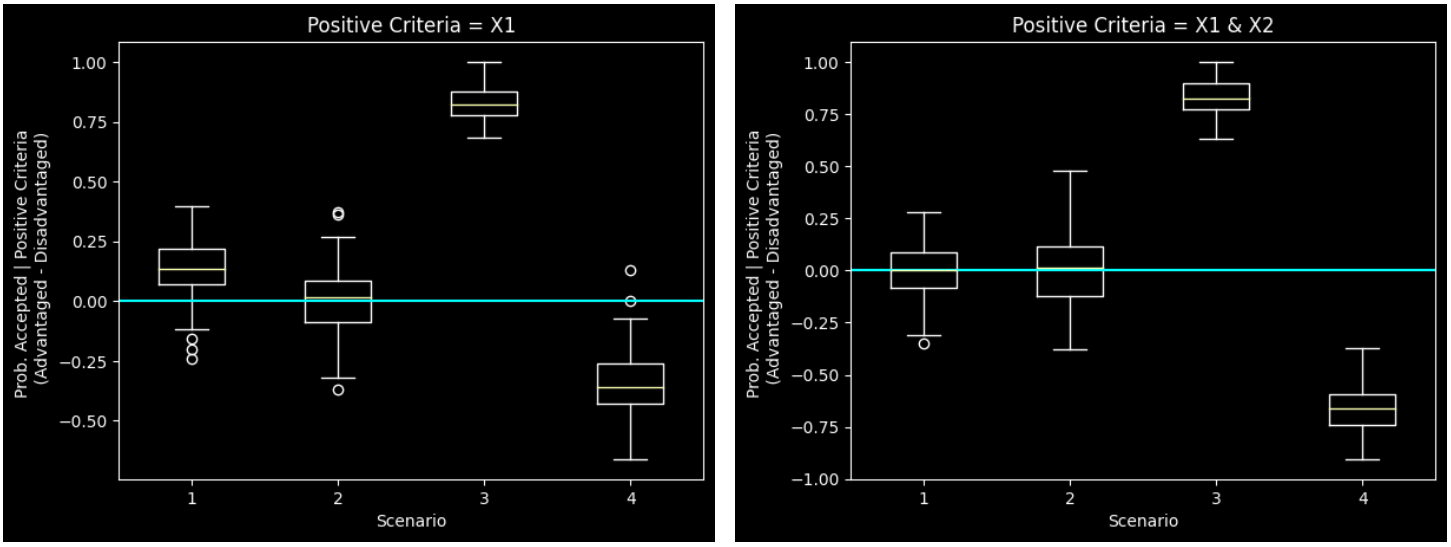


Figure 3

Different definitions of fairness are appropriate for different situations and have unique benefits, therefore we consider all four definitions together to identify which admissions scenario is most fair. The summary of these evaluations is in Table 2, where we see that Scenario 2 is the most fair by these four definitions; Scenarios 3 and 4 do not meet any of the fairness criteria.

Table 2		**Scenario 1**	**Scenario 2**	**Scenario 3**	**Scenario 4**
Fairness Through Unawareness		✓	✓	X	X
Counterfactual Fairness		✓	✓	X (favors A)	X (favors D)
Demographic Parity		X	X	X	X
Equal Opportunity	**X1**	X	✓	X	X
	X1 & X2	✓	✓	X	X

Preferable Admissions Scenario - Considering All Metrics

With all of these results combined – equity and efficiency performance, cost concerns, and theoretical definitions of fairness – we can make an informed decision about the ideal admissions criteria which preserves the importance of accepting high-quality students while ensuring fairness for all applicants. Scenario 2 has high equity, moderate efficiency, low costs, and meets most definitions of fairness. Therefore, Scenario 2 is the ideal approach to optimize all relevant criteria for the admissions team.

Part 2 - Impact over Generations

Experiment Design & Model Parameters

While looking at admissions equity and efficiency in a singular generation can reveal biases and interesting phenomena, this process does not tell the complete story. Namely, this approach overlooks how biases can propagate and evolve over time, leading to changing cohort dynamics. Therefore, we seek to analyze the impact of college admissions practices on social mobility across generations. To do this, we presume that the next generation of students are placed into groups A and D with some probability. The probabilities that each student in the subsequent generation will move from D to A, or vice versa, are summarized in Table 3. These probabilities depend on three factors:

- Admissions status of the parent
- Advantage status of the parent

- Effectiveness of education

The first two factors are relatively straightforward, as they were covered in Part 1. The "effectiveness of education" variable refers to the impact of education on advantage status within a single generation. We decided to make this a variable due to its high impact on the dynamics of the population; demonstrating multiple settings allows us to partially distinguish what parts of our results are attributable to admissions technique and what results are due to the effectiveness of education. For purposes of simplicity, we treat this effectiveness as a binary variable. A value of one indicates that college education has a large impact on advantage status in the successive generation. A value of zero indicates that college education has a minimal impact: the offspring of an advantaged individual that was rejected will still be advantaged with probability 0.9, and the offspring of a disadvantaged individual that was admitted will only become advantaged with probability 0.1.

Table 3. Probability of Next Generation Moving To Opposite Group

	Advantaged (probabilities indicate chance of next generation becoming *disadvantaged*)		**Disadvantaged** (probabilities indicate chance of next generation becoming *advantaged*)	
	Admitted	**Rejected**	**Admitted**	**Rejected**
High Effectiveness	0.0	0.25	0.5	0.0
Low Effectiveness	0.0	0.1	0.1	0.0

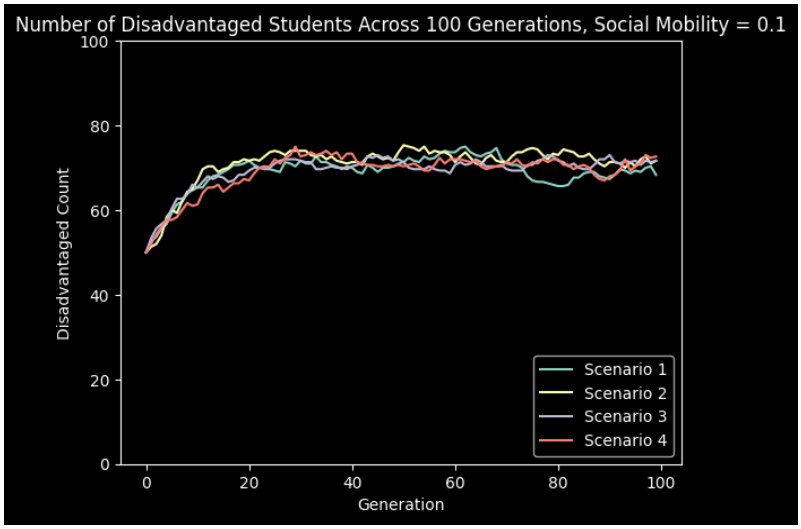
As we did in part 1, we began by generating an applicant pool of 100 individuals. The split between advantaged and disadvantaged began at exactly 50-50 but changed over time. We implemented each of the four admissions criteria described in part 1 and simulated a total of 100 generations for each. This process was initially done using the low-effectiveness environment, after which we tested the high-effectiveness version. Finally, we repeated every simulation three times and calculated the average efficiency, equity, and group size, as this reduced the variance significantly.

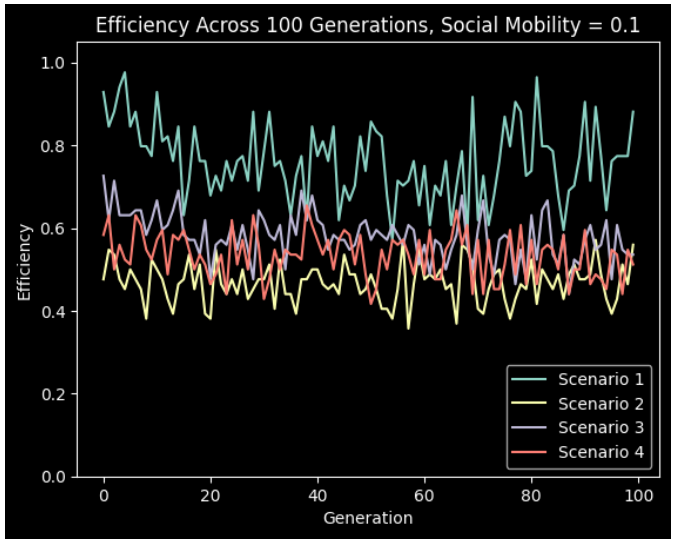
Results

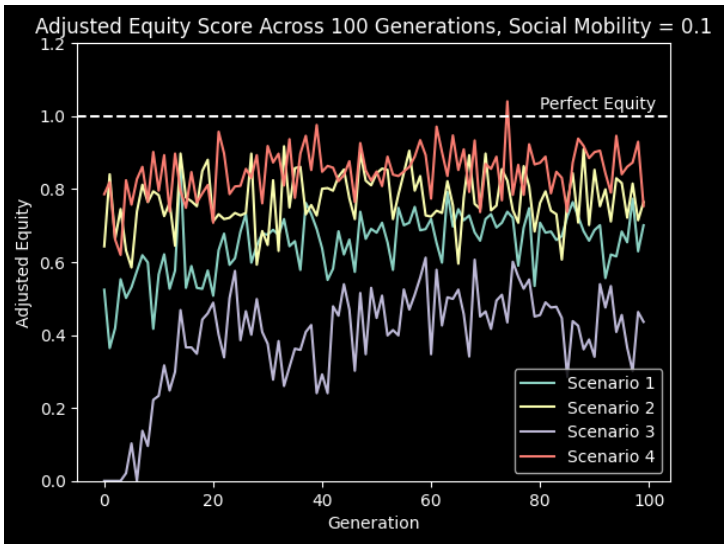
Let's begin with analyzing the low-effectiveness environment. Recall the four scenarios of admissions considered from Part 1:

- Scenario 1: x1 and x2 are available.
- Scenario 2: only x1 is available.
- Scenario 3: x1 and G are available.

- Scenario 4: Collect data about x2 from disadvantaged applicants, note when x1 = x2 = 1 and g = D, otherwise only record x1.







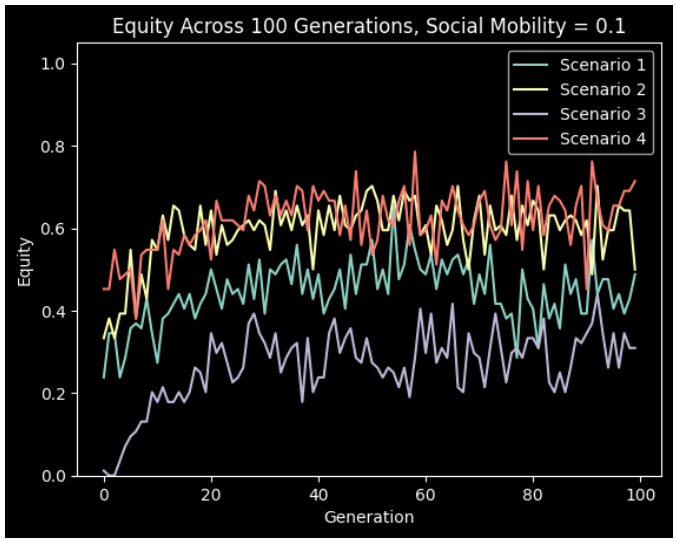
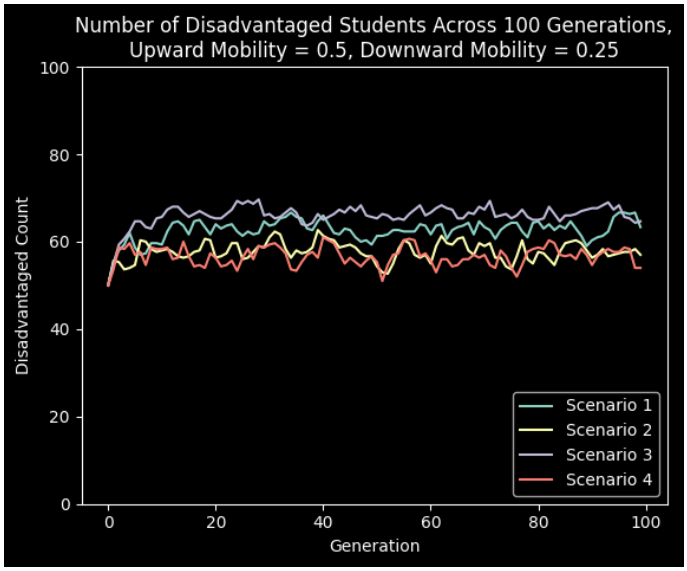


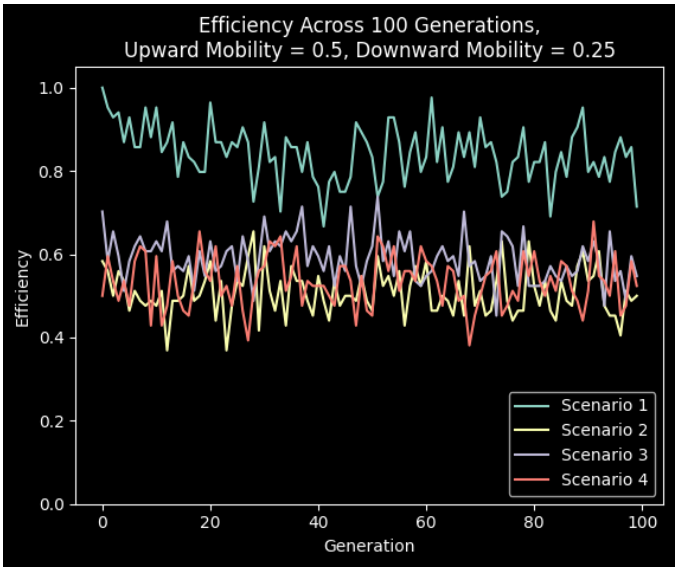
Figure 4

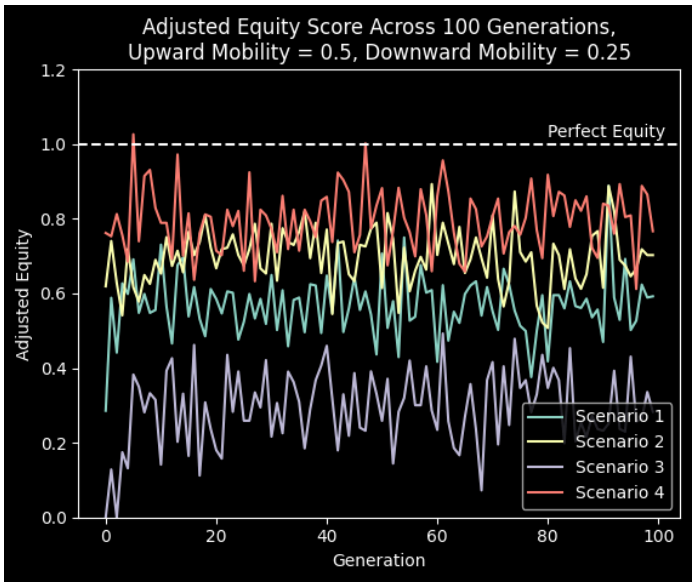
After running the cross-generational simulations, one trend became immediately clear: the number of disadvantaged students grew quickly from generations 1 through 20, after which it plateaus, with minimal difference across scenarios. This conforms with our expectations, as we reject roughly two and a half times the students we admit, and the probability of moving between groups is at most 10 percent in the low-effectiveness scenario. As a result, many families move into the disadvantaged group and stay there for a long period of time.

More interestingly, we see clear discrepancies between scenarios 1 through 4 in both equity and efficiency. Scenarios 2 and 4 are the most equitable regardless of generation, but they are the least efficient, with scenario 4 slightly outperforming 2 in efficiency. Scenario 1 performs far and away the best in efficiency, but lags behind in equity. Scenario 3 is reasonably efficient, but not at all equitable. Overall, we see similar trends to what we saw in Part 1. However, in this simulation, overall equity rises over time, with scenarios 2 and 4 approaching perfect equity. This is indicated by the "adjusted equity" graph, which is normalized to a value of 1 being perfectly proportional representation between the admitted and overall populations. Meanwhile, overall efficiency decreases across the first 20 generations, mostly as a result of more students moving from A to D.

These results showcase the effects of various aspects of our model, but they likely underestimate the real-world odds of movement between advantaged and disadvantaged groups as a result of admission. Therefore, we ran our models again, this time using distinct probabilities of social mobility to simulate a "high effectiveness" environment. The results of these simulations are displayed in Figure 5.







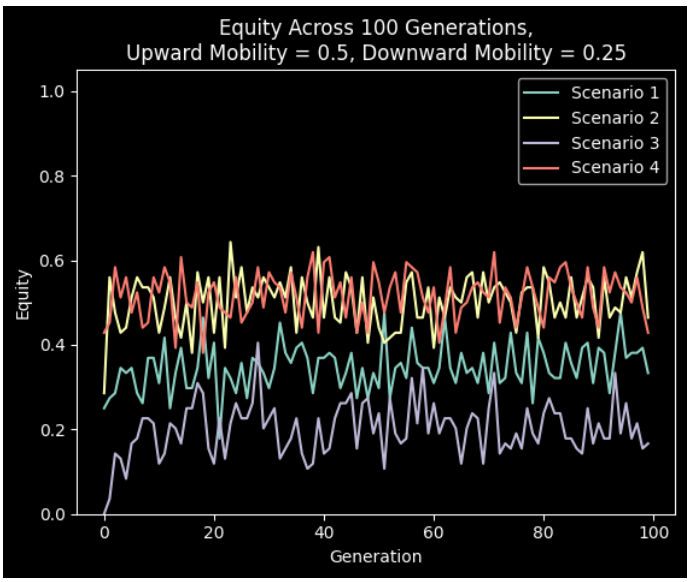


Figure 5

Immediately, we noticed a substantial jump in the number of advantaged students compared to the previous environment. This makes sense, as the odds of moving from A to D when rejected are half of the odds of moving from D to A when accepted. The increased movement from A to D and vice versa reveal differences in advantage breakdown between scenarios. Scenarios 2 and 4 reach an equilibrium with substantially fewer disadvantaged students than in scenario 1 and especially scenario 3. As one would suspect, the ratio of students in A or D in equilibrium is directly related to the equity of the admissions practice, as the odds of moving from A to D when rejected are half of the odds of moving from D to A when accepted (meaning more students have a 50% chance of upward mobility in exchange for more students having a 25% chance of down mobility, a net gain for the number of advantaged students). The trends for equity and efficiency remain the same as the low-effectiveness environment, suggesting that the environment has a minimal impact on how scenarios 1 through 4 compare with one another. It is important to emphasize that the effectiveness of education is an assumption we made about the environment; we didn't consider how it may evolve over time or how other features of our model could impact it. However, our model is robust to this variable; in other words, we can compare admissions scenarios and have the same directionality of impact regardless of how effective education is at inducing social mobility.

These results are generally consistent with real-world observations, at least directionally: more equitable admissions practices generally are less efficient but can have very large impacts on lower-resource communities. However, discriminating against certain groups by inferring unknown things about them using a label (scenario 3) incurs many of the costs of equity without the associated benefits of efficiency.

Part 3 - Employment Trends and Bias Mitigation

Employment Distribution

Expanding our model to include a multi-generational simulation highlights the long-term trends and social outcomes of each admissions scenario. To further explore the effects of admissions criteria, we incorporate a new outcome variable into our model: employment. We utilize a predetermined set of employment probabilities based on individuals' attributes (x1 and x2) and group membership (advantaged versus disadvantaged). These probabilities, which specifically refer to the likelihood that an individual will find gainful employment within ten years of *applying* for college, are displayed in Table 4.

Table 4. Employment Probabilities

X1	X2	G	Probability of Employment
1	1	D	0.75
1	1	A	0.99
1	0	D	0.50
1	0	A	0.75
0	1	D	0.25
0	1	A	0.50
0	0	D	0.05
0	0	A	0.25

We incorporated these probabilities into our single-generation admissions model to simulate the resulting employment distribution. After simulating each of the four admissions scenarios 100 times and averaging the results, we found that the employment distribution by group was equal across all four scenarios, with roughly 25% of disadvantaged individuals and 75% of advantaged individuals securing employment. These results, displayed in Figure 6, are consistent across all four admissions scenarios because the employment probabilities are not determined or affected by admissions status. Therefore, the distinct admissions outcomes across the scenarios have no bearing on the ultimate employment distribution.

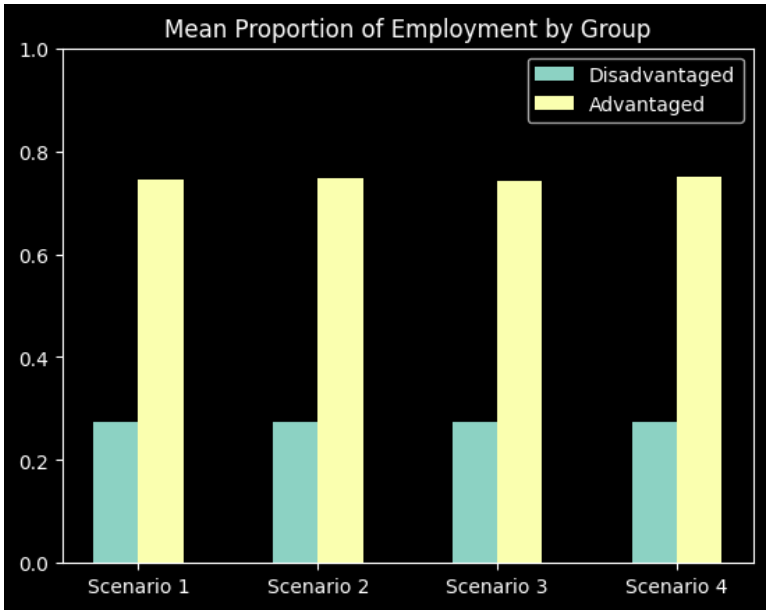


Figure 6

This same employment pattern holds true for each scenario even across 100 generations. When we included employment probabilities in our multi-generational model, the resulting employment rates for each generation hovered near 25% for disadvantaged applicants and 75% for advantaged applicants. (For most generations, the employment rates were slightly higher or lower than these averages due to the random noise in our model.) These cross-generational trends are captured in Figure 7.

Figure 7

Although Figure 7 demonstrates that rates of employment by group remain steady, we found that the total number of individuals employed did change over the course of our simulation. As discussed in Part 2, in every version of our multi-generational simulation, the number of applicants in the disadvantaged group grew over time. Given the lower probability of employment for disadvantaged individuals, we expect that as more people shift from the advantaged to disadvantaged group, total employment in the sample will decrease. Our simulation supported this hypothesis, with the total number of employed individuals falling from 50 to just over 40 in the course of roughly 20 generations. Figure 8 illustrates this trend. However, it is important to note that in the real world, there exists a lower bound on employment rates due to demand for work. In other words, total employment should never drop below some exogenously-determined threshold. Including such a threshold was outside the scope of our simulation, but is an important consideration when thinking about what these results tell us about the real world.

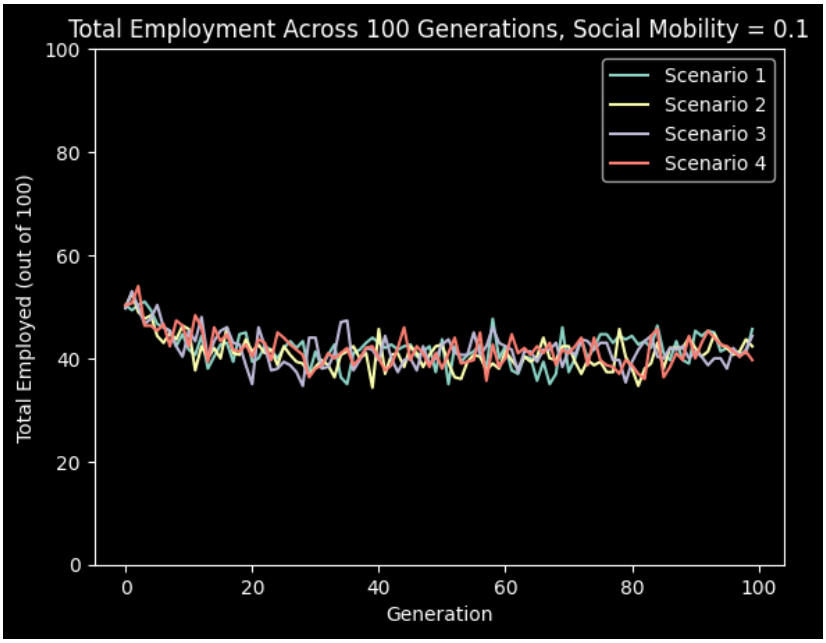


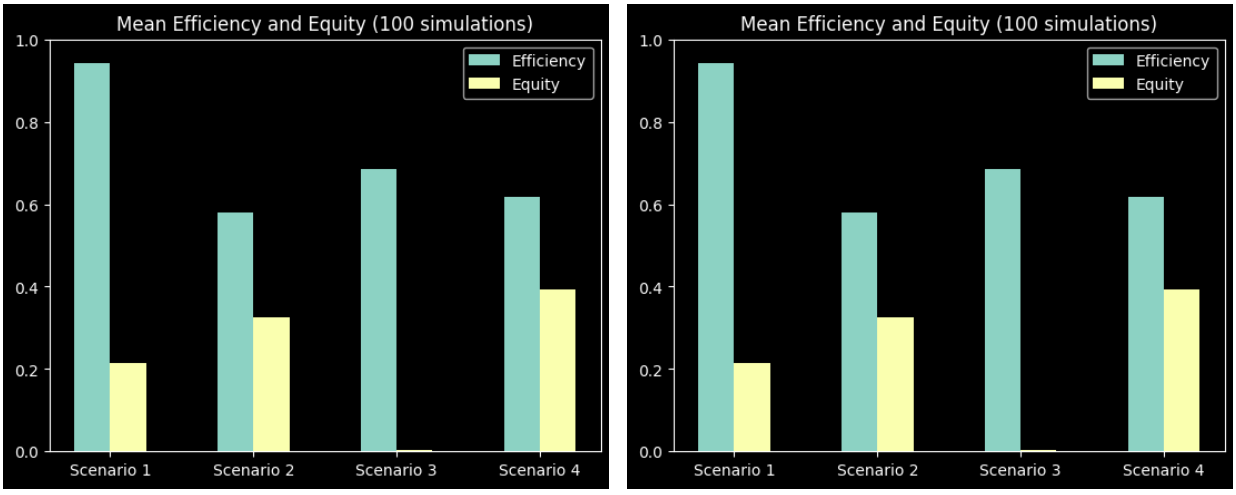
Figure 8

Bias Mitigation

Given that every scenario eventually leads to a suboptimal employment rate and simultaneously fails to meet all of our prioritized definitions of fairness, it is clear that a new approach to college admissions is necessary. The current system leads to biased admission outcomes and does nothing to remedy disparate employment rates among the disadvantaged and advantaged groups. Therefore, we developed a bias mitigation approach to improve our model, with the ultimate goal of reducing disparities in admissions as well as employment outcomes.

To mitigate bias in our model, we first identified the initial source of bias: the dataset itself. More specifically, advantaged applicants have higher probabilities of displaying positive attributes X1 and X2, relative to disadvantaged applicants. The effects of this reality are lower admission rates and lower employment rates for those in the disadvantaged group. Therefore, we decided that the most direct approach to mitigating bias would be to increase the probability that disadvantaged applicants exhibit positive attributes X1 and X2. To increase these probabilities, we added a new element to our hypothetical admissions system: trainings for disadvantaged applicants. These trainings are incorporated into our model as increased probabilities of positive attributes for the disadvantaged group. In particular, we increased the probability of having each attribute from the original ⅓ to ½. The probabilities for the advantaged group remain the same (⅔).

After selecting our bias mitigation approach, we ran all of our simulations again using the updated probabilities. We then compared these results to the results from our original simulations. In our single-generation model, our bias mitigation approach increased equity across most scenarios, and efficiency increased for scenarios one and four. Importantly, we did not see a significant decrease in efficiency in any scenario after implementing bias mitigation. The efficiency and equity results are displayed in Figure 9. In addition, our bias mitigation increased the rate of employment among the disadvantaged group from 25% to nearly 40%.

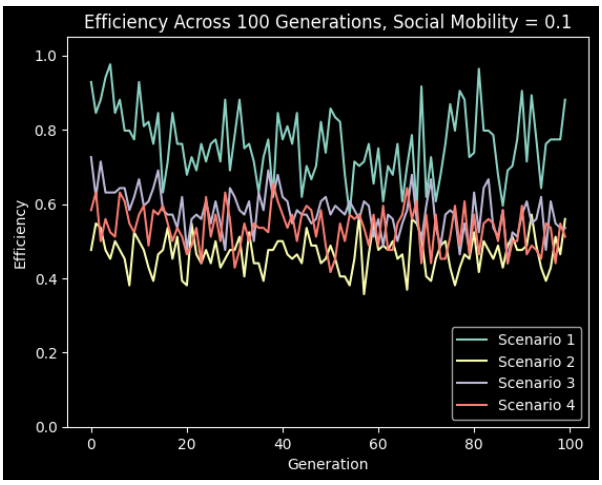


Original Simulation

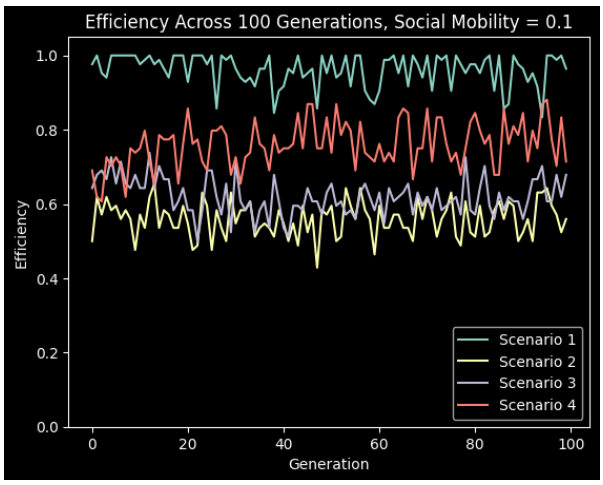
Bias Mitigation Simulation

Figure 9

Next, we compared the results of our bias mitigation approach across 100 generations with the results from our initial multi-generational simulation. When compared to our previous simulation, we found that bias mitigation increased overall efficiency in scenarios one and four but had little effect on efficiency in scenarios two and three. Equity also increased across the board, with the exception of scenario three. Most encouragingly, we found an overall increase in the total number of individuals employed over time, with the model reaching equilibrium around a 50% employment rate. This represents a nearly 20% increase in total employment compared to the simulation without bias mitigation. These efficiency, equity, and employment trends are displayed in more detail in Figures 10 through 13.

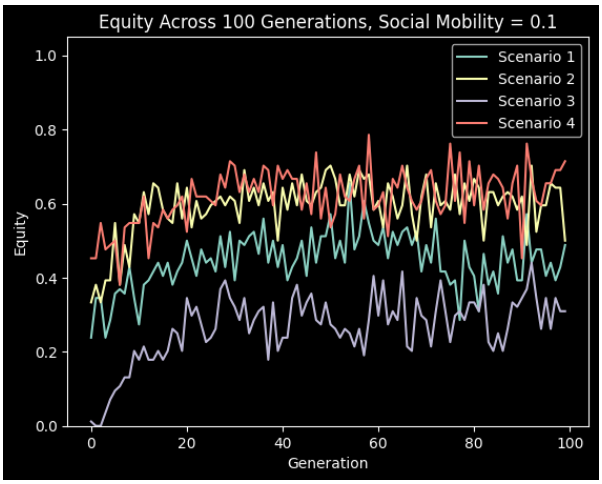


Original Simulation

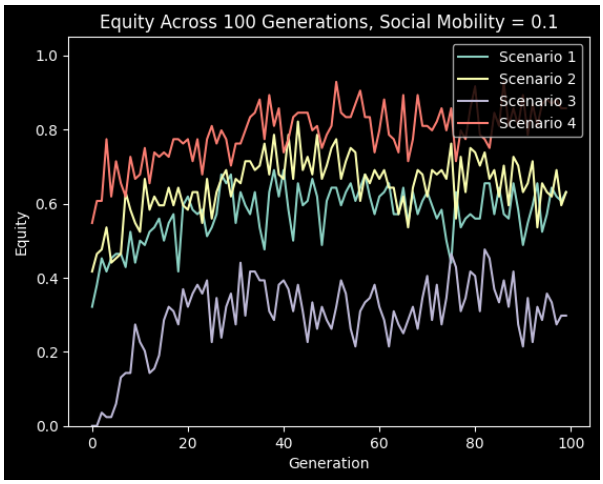


Bias Mitigation Simulation

Figure 10

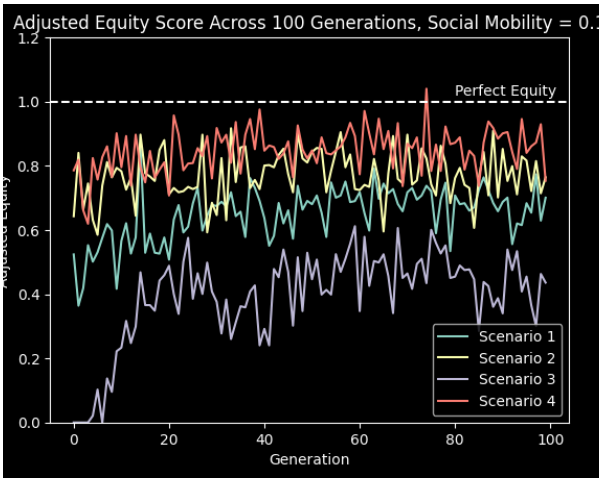


Original Simulation

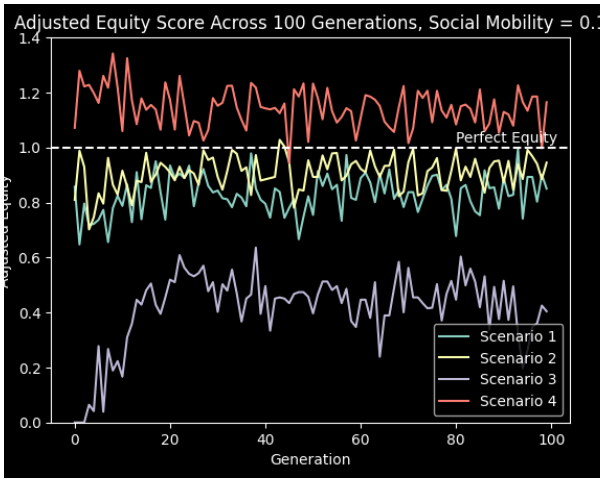


Bias Mitigation Simulation

Figure 11

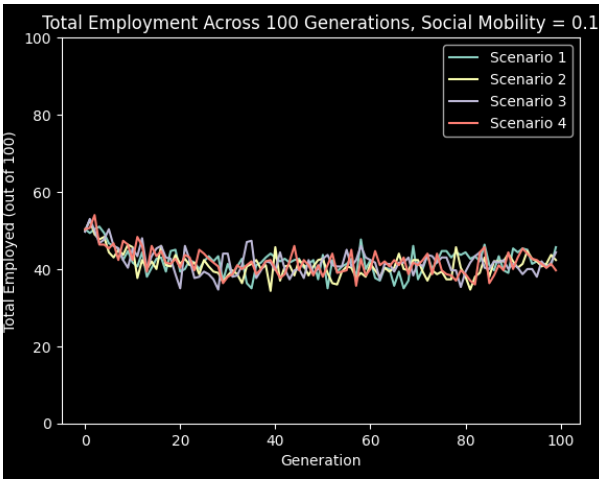


Original Simulation

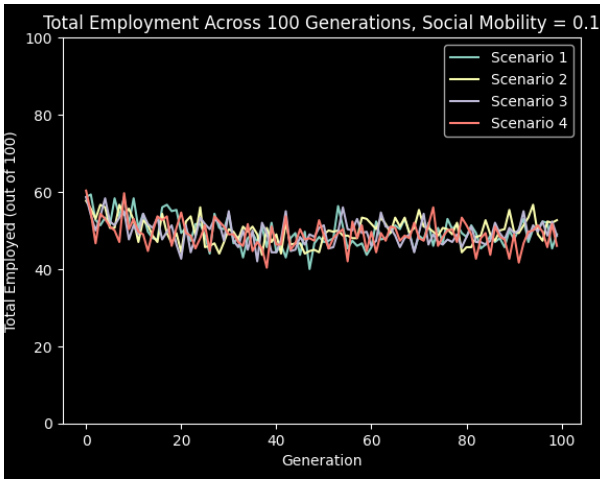


Bias Mitigation Simulation

Figure 12



Original Simulation



Bias Mitigation Simulation

Figure 13

Insights and Implications for AI Systems

Our initial results, as well as our bias mitigation results, highlight the need for a concerted effort to improve fairness within college admissions and other AI systems that can impact people's lives. Although efficiency and accuracy are highly desirable outcomes, they should not come at the expense of equity. Algorithmic bias can be mitigated through a variety of means, including complex statistical approaches and the addition of human-in-the-loop techniques. However, while such approaches can reduce bias within the algorithm and its outputs, they generally cannot solve the underlying cause of inequity. Our bias mitigation approach illustrates that it is sometimes appropriate to search for real-world solutions, as opposed to data manipulation or algorithmic changes, when attempting to improve the fairness of

AI systems. When resources are available, working to mitigate bias within institutions is a worthwhile endeavor, as it can lead to improved efficiency, equity, and other positive outcomes.

Code is available at: https://github.com/baileywellen/AlgorithmicFairness_CollegeAdmissions

References

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