I never expected the phrase "perception is reality" to have meaning within the scope of computer science. I saw computers to have a sense of universalism, thinking of them solely as calculators built upon axiomatic truth. But, my viewpoint was incorrect. At the beginning of freshman year, I read a study from ProPublica about the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm--a computational tool that assesses the risk of recidivism for people with criminal records. In their study, ProPublica found COMPAS misclassified black men as risky at twice the rate to that of white men. It is scary to think that with COMPAS, I, as a black male, am more likely to be profiled as risky simply because of the color of my skin. For judges who use machine learning (ML) algorithms such as COMPAS to help determine penalties, they may unknowingly create realities which can unfairly affect the lives for certain demographic groups. By pursuing a PhD in computer science, I'll systematically investigate algorithmic bias and develop novel systems to detect and mitigate disparate impact.

After my freshman year, I spent the summer at Rutgers University identifying and understanding quantitative notions for fairness. Under the supervision of Prof. Anand Sarwate, we realized most existing notions of fairness depended heavily upon the outcomes of the trained models; fairness was defined in terms of the ground truth. Since the ground truth can be difficult to quantify within many domains, we considered how to evaluate the fairness of a model via its decision-making process. As a result, I developed a metric which compares how features are weighted across varying demographic groups. I used home loan data and trained logistic regression models, where each model is trained upon a specific racial group and compared how the weights for each model varied. As the preliminary results pointed to decision processes between groups varying, being a potential indicator of bias, I became interested in understanding if the relationships between features could elucidate the presence of fairness. Identifying whether causal modeling would be a sufficient solution became an interest of mine.

Wanting to know more about how causal models can serve to identify biased features in models, I spent the next summer learning about causality and causal inference. I was at the University of California, Berkeley being supervised by Prof. Moritz Hardt, researching how causal inference methods could be verified. Our project involved utilizing fully controllable, simulated environments to benchmark causal inference methods by running various causal experiments and comparing their outcomes to the intended results. I was responsible for integrating the Dynamic Integrated Climate-Economic (DICE) model to be a simulator to run benchmarks. As a result, I began developing causal experiments, centered around casual interventions, to determine if inference methods could accurately identify outcomes when interventions are imposed on a model. Using causal modeling to answer questions relating to fairness still has its own pitfalls because the models must accurately represent the underlying decision process of the ML model. Limited to only observed data, it can be difficult to represent ML models. But, understanding how combinations of features affect performance can narrow the scope of biased features.

The following fall, I returned to the University of Maryland, Baltimore-County interested in figuring out how the combination of features used for a model affect its performance and fairness. Under supervision of Prof. James Foulds, I conducted an independent research project focusing on how features can be selected for models that would optimize for performance and fairness. I developed a feature selection algorithm that evaluates the efficacy of features based on a linear combination of accuracy and differential fairness—a metric that evaluates the fairness of a model for intersectional demographic groups. With this tool, I was able to see how varying the importance of fairness affected the features selected. From our experiments on home loan data, the most interesting finding was that an outline of a Pareto frontier existed when we plotted accuracy versus differential fairness; this result suggests that there is a tradeoff between accuracy and fairness. But with careful selection, it is possible to find features that satisfy fairness constraints and still perform well. During this project, I realized that ML practitioners also have to seriously think about fairness since feature selection is pivotal in the ML development pipeline. More broadly, fairness is a topic that spans beyond theory and all stakeholders interacting with ML affect the state of fairness. After this experience, I found it important to think more broadly about the mechanisms that affect fairness.

Last summer, I started to think more broadly about how different players contribute to the issues of algorithmic fairness. Unaware users even have an impact on the bias of models. Interning at Microsoft Research in New York, I worked on a project with Miroslav Dudik, Solon Barocas, and Hal Daume researching strategic behavior in the ML setting. When ML models are instituted, agents may be incentivized to artificially alter their features to increase their chance of positive classification by models. While prior work focused on developing mechanisms which minimize the ability for agents to "game" models, we aimed to identify when this occurrence can be detected. Specifically, we used the contextual bandits framework to characterize when a proposed policy suggests a distribution is gaming or improving their features. Our framework allows us to consider how well some groups can change their features compared to others; fairness issues arise when those with more resources can more easily change their features in contrast with those who have little to no resources. In the long-run, it may be possible that the dynamics of the underlying population change as a result of unfair allocations of resources.

My understanding of algorithmic fairness has vastly expanded since the start of my freshman year. Any ML model has the ability to change how people see the world; biased models can further misrepresent the world we live in. During my PhD, I want to understand the broader effects of ML on human perception. Unfair models can negatively affect public policy, markets, and the individual worldviews of many. Identifying indicators which determine the impact of models on populations will make it possible to evaluate the efficacy of developed technologies. It is my dream to use this insight to develop systems that incentivize the diversification, infusion, and representation of many ideas that exist. After completing my PhD, I see myself

collaborating in spaces between industry, government, and academia; all sectors must be aware of the long-term effects that ML can have on humanity.

Pursuing a PhD in Stanford's computer science department would assist in my journey because the insights the faculty could provide would be worthwhile. I see Prof. Leskovec's research related to network analysis as a support for understanding the characteristics of network propagation. His recent work with graph neural networks could be used to model which social structures allow for influence to be more easily propagated. Furthermore, Prof. Ermon's research focusing on creating fair datasets could serve as the hypothetical training data for ideal models; we could investigate whether having these fair models affects how users are influenced by ML models. Additionally, Prof. Bernstein's research with online governance can serve as a basis for representing the desires of individual users in an online platform. An agent-based model representing the individual users interacting with each other and the online moderators can further explain the mechanisms of how users are influenced; we could create a system based off PolicyKit that prevents negative influence from spreading across the network. And having an environment that intrinsically has interdisciplinary, collaborative, and thought-provoking people is what will help find an answer to addressing unfair models.