Extended Abstract: Student Thesis Summary

Lily Hu [1]
[1]Harvard University, Cambridge, MA, 02138, USA
{lilyhu}@g.harvard.edu

In our field of Artificial Intelligence, the machine-behavioral realization of models based in neoclassical economics and utilitarian calculus represents not only a successful test-of-concept, but we may view such principles as in fact, one of the first concrete and systematic operationalizations of an ethical theory in the "real world." Here, a notable paradox stands: While the ethics of the *field* of AI at-large is under continual debate and contention, the conversation about the ethical grounding of *individual* AIs in particular is rather uncontroversial—AI tools, whether it be a reinforcement learning agent or a machine learning-based classifier, are for the most part, act utilitarians—that is, the right action is the one that maximizes (expected) utility.

When the reach of AI was largely limited to solving technical tasks, researchers could view the dictates of utility maximization as a purely procedural approach to determining agent action. But the entry of AIs into the realm of the social has forced a shift in our evaluation of the *rightness* of utility-based models, resurfacing the fundamentally ethical nature of agent decision-making. While philosophers have known for quite a while that utilitarianism alone as an ethical basis of action can provide an impoverished account of justice or fairness, the realization has been a recent rude awakening for computer scientists working in Artificial Intelligence and Machine Learning who are now newly grappling with issues of discrimination that arise when automated utility-maximizing tools are at the helm of social decision-making processes. Cathy O'Neil's *Weapons of Math Destruction* (O'Neil 2017) and ProPublica's audit of the COMPAS recidivism tool (Angwin et al. 2016) are just two of a series of high-publicity investigations that have uncovered the surprising fact that seemingly objective and purely optimization-driven devices can exhibit human-like biases in socially-oriented tasks, in many cases producing behaviors that are considered abhorrent and even unlawful.

But within the utility-maximizing model of AIs, encouraging behavior and outcomes that align with social values tends to be of secondary interest. My thesis examines the extent to which the governing ethos of utility-based rationality built into AI systems is compatible with larger societal interests and norms of fairness. In particular, when AI techniques are employed as resource allocation mechanisms—whether it be sifting through job candidate résumés to offer interview slots or defendant data to produce risk scores bearing on parole decisions—unconstrained maximization of predictive accuracy as utility has been shown to reinforce and deepen racial and gender inequalities in societal outcomes. As such, the demands of fairness must coexist alongside or be built into the existing utilitarian framework of AI systems. My thesis asks: *How can the variable demands of justice as fairness be computationalized, so as to fit within a utility-based AI system, in a way that approximates the dynamic environment of the social world?*

Research in the growing literature of algorithmic fairness has studied similar questions by beginning with a domain-general definitional account of fairness and then constraining the behavior of particular algorithms so to align with the fairness notion presented. However, the problem of generating general principles of fairness is not only a notoriously difficult task in itself, but such an approach is not sufficiently contextual to handle the particular trespasses of justice at stake in domains with distinct histories, patterns of inequality, and moral obligations. I claim that we cannot adequately evaluate the social and ethical impact of an algorithm's behavior without examining deeply the particular system within which it is embedded. Since AIs rarely fully control a process of resource distribution, my work models and analyzes the dynamics of an algorithm's whole system of use in order to determine what type of intervention would be appropriate to achieve an outcome that can be ethically argued as just for a particular system.

In work presented this summer at a talk at FAT/ML (Fairness, Accountability, and Transparency in Machine Learning), I tackle the problem of algorithmic reinforcement of disparate group outcomes in the labor market (Hu and Chen 2017) and argue that relying on leading notions of algorithmic fairness to constrain hiring practices are insufficient to overcome the steeped inequalities that characterize every cut of the employment cycle. I prove that when the group-memberships of job candidates are observable, such as race and gender, and decision-makers are equipped with standard *homo economicus* capabilities such as Bayesian reasoning, conceptions of individual and meritocratic fairness, which constrain algorithms to treat similarly qualified people similarly (Dwork et al. 2012; Kearns, Roth, and Wu 2017), continue to foreground short-term utility maximization, justifying disparate outcomes in a vicious cycle that fails to achieve long-term societal goals of ensuring equality of opportunity.

A central argument of my research contends that these *static* utility-based conception of optimal hiring, wherein algorithms predict and hire the "good" workers out of a candidate pool, is ill-suited for understanding the dynamics of complex social processes and as a result, the societal obligations to which AI tools may be bound. Instead, my work widens the view of algorithmic fairness to consider the dynamics of the entire labor market system, from workers' investment opportunities prior to entering the labor market to their tenure within the market as they interact with various firms and cycle through different jobs. In re-embedding algorithms in their social and human contexts, my work preserves aspects of rational choice theory that bear on human behavior while departing from a popular machine learning practice of treating human data as *a priori* parameters of a utility function rather than the products of structurally influenced human actions. My stylized model casts a worker as a rational actor navigating a sequence of stages wherein she has attributes both personal (such as ability level) as well as social (such as group membership), faces individualized education investment costs, and makes employment-related decisions. Labor market interactions between workers and hiring-agents are embedded within a reputational dynamic repeated game where changing group reputations, which approximate societal standing, bear on members' investment costs—for two workers equal in innate ability, a lower reputation group member faces higher costs—as well as a hiring agent's perception of the worker's qualifications.

When initial group reputations are unequal, worker and firm best responses may cause the system to converge to an asymmetric equilibrium with disparate outcomes. Further, I prove that this reputation system has feedback and externality properties such that even when standard algorithmic fairness definitions are in place, the asymmetric equilibrium in which workers of the same ability level but of different groups face disparate wage prospects is maintained.

As evidenced by the methodology of this work, I do not dispose of the concepts of rationality and utility, rather I borrow techniques from both economics and sociology to build a model of the labor market pipeline that is better able to pinpoint an underlying *origin* of algorithmic disparate outcomes, shifting from a data-centric to an action-centric view of the world. For the final upshot of this work, I designed a fairness intervention on hiring practices to address the empirically-validated social phenomenon of development bias, in which members of a disadvantaged group are disproportionately excluded from opportunities required to realize their goals, a leading source of disparate employment outcomes (Loury 2009). I prove that my proposed short-term intervention installs long-term social fairness by converging the system to a group-equitable steady-state, and that moreover, under weak market conditions, the "fair" equilbrium outcome Pareto-dominates the asymmetric steady-state arising under unconstrained or procedurally fair hiring.

My paper on fair hiring in the labor market is one of the first works in the algorithmic fairness community that explicitly models the impact of algorithms *in situ* and makes a comparative statics social welfare argument against existing propositions of fairness . I also developed an argument grounded in legal and philosophical discourse for the ethicality of both the intervention proposed and the final group-egalitarian outcome in the labor market that is not based in utilitarian calculus. My inclusion of such content is rare in the fairness literature and highlights the central role that I believe ethics and justice must play even in computer science and mathematical research on algorithmic biases.

In another paper that is situated at the intersection of AI and ethics, David Gray Grant and I have written a review of the reigning definitional accounts of fairness within the algorithmic fairness community and their relation to substantive, distributive, and procedural claims of justice in philosophical discourse. My current computational research designs a dynamic online task that incorporates common externalities of machine decision-makers' previous choices into future rounds of learning. Within a multi-armed bandit setting, I ask how a decision-maker's policy changes when the underlying probability distribution of arms morph in response to her sequence of actions. Embedded within appropriate social contexts such as assessing the returns to public school funding, such a learning task can be used to conceptualize the demands of distributive and intergenerational justice. Similarly to my previous work on labor market fairness, the algorithmic learning task is designed to reflect the ethical challenges of genuine resource distribution scenarios within which AI agents are currently involved with unique emphasis on the feedback and side-effects of a decision-maker's own actions.

While the constructs of utility and rationality continue to be invaluable for AI systems grappling with societal values, it is also crucial that a conception of justice may exist as independent and distinct from any other utility maximization problem. My thesis sits at this region lying in between, at the intersection of utilitarianism as a framework and methodology of algorithm theory and justice-as-fairness as an ethical and social aspiration, characterizing aspects of this still under-explored landscape.

References

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias: Theres software used across the country to predict future criminals. and its biased against blacks. *ProPublica, May* 23.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226. ACM.

Hu, L., and Chen, Y. 2017. A short-term intervention for long-term fairness. *arXiv preprint arXiv:1712.00064*.

Kearns, M.; Roth, A.; and Wu, Z. S. 2017. Meritocratic fairness for cross-population selection. In *International Conference on Machine Learning*, 1828–1836.

Loury, G. C. 2009. *The anatomy of racial inequality*. Harvard University Press.

O'Neil, C. 2017. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.