Uber's Pax: Hidden Bias in Rating Systems

Alex Rosenblat
December 30, 2015
Working Paper[1] prepared for CSCW Workshop: *Algorithms at Work*

The on-demand economy has coalesced around the momentous rise of app or platform-based companies that employ service-providers and connect them with consumers through a digital matching service. These companies structure distributed employment systems through a range of remote, electronic and semi-automated functions. The automation of many managerial and organizational functions, including worker evaluations, is one significant facet of these systems.[2] Many of these companies, such as Uber, Lyft, Handy, etc. prompt consumers to evaluate their experiences with workers through a rating system. This paper will use the Uber system as a case study: passengers are prompted to rate drivers on a 1 to 5 star scale, and drivers must maintain an overall rating that hovers around 4.6 out of 5 stars or they risk deactivation (temporary suspension or permanently fired) from the system. Their overall rating reflects an average of their last 500 rated trips. In some markets, Uber articulates to drivers that it will not count ratings drivers receive during high price surges (which drivers associate with low ratings) in that tally, but that is the only exception. Drivers in the Rosenblat & Stark study express a lot of frustration and anxiety with regards to their ratings, which inevitably seemed to go down at some point, although drivers were not necessarily able to identify what had changed,[3] if anything, in their own work habits. Some observe that they receive low ratings unfairly in response to a variety of things outside of their control, including: surge pricing; GPS or navigation malfunctions; the passenger's misplacement of their own location for pick-up; holding passengers in compliance with both Uber's rules and local laws, such as not taking more passengers than there are seatbelts in the vehicle, etc.[4]

In this model, consumers are empowered to act, in part, as middle-managers of workers, both through the design of the app and in the evaluation functions they perform.[5]

---

[1] The ideas presented in this paper are part of a larger project in progress by Alex Rosenblat, Karen Levy, Solon Barocas, and Tim Hwang. This project is sponsored by the MacArthur Foundation and the Intelligence & Autonomy Initiative at Data & Society.

[2] Lee, M. K., et al, "Working with Machines: The impact of algorithmic, data-driven management on human workers," (pp. 1603–1612), 2015, Proceedings of the 33rd Annual ACM SIGCHI Conference, Seoul, South Korea. New York, New York, USA: ACM Press. http://doi.org/10.1145/2702123.2702548, p. 1603; Rosenblat, Alex and Stark, Luke "Uber's Drivers: Information Assymetries & Control in Dynamic Work," Oct. 15, 2015, Accessed Dec. 29, 2015, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2686227 , p. 01.

[3] Passengers are not generally educated on what the rating should reflect, such as "safe ride from A to B" or "clean vehicle." Passengers may presume that 4 out of 5 stars is a good rating, whereas it is actually a failing grade for drivers. Some drivers make attempts to educate passengers by placing explanatory fliers in their backseat, and many offer free candy, water, and sometimes, cell phone charging cords to elicit a 5-star rating. The vagaries of what a *good* rating should reflect may support a system where passengers channel their frustrations with the Uber system as a whole into the ratings that primarily impact the driver's employment eligibility. Generally, the Uber system is designed and marketed as a seamless experience; any friction could be cause for lower ratings. Many drivers in the Rosenblat & Stark study expressed that they weren't always sure what they were being rated on, but many tried to compensate for anticipated negative ratings by offering snacks, water, a phone-charger cord, or by offering to adjust music, temperature, and evaluating whether the passenger wanted to engage or disengage from conversation.

[4] Rosenblat and Stark, 2015, p. 12

[5] Ibid., p. 11; Stark, Luke., & Levy, Karen, "The consumer as surveillor," June 2015, Paper presented at the

Uber leverages the rating systems to determine the employability of workers.[6] Ratings, as a reflection of consumer preferences, allow companies to institutionalize consumer preferences if they use them as direct assessments of worker performance. A variety of social science research in online marketplaces indicates we should expect bias to creep into consumer-driven contexts.[7] The rating system thus potentially enables systemic discrimination against minorities and women from consumers. The dynamics of implicit and explicit bias has been addressed in a host of social science research demonstrating evidence of racial bias in performance evaluations by supervisors who render more negative scrutiny in evaluations of workers with protected-class characteristics.[8] In Uber's case, the biases held by passengers may be funneled through the ratings model feedback mechanism[9] and they could have a disproportionate adverse impact on drivers who, for example, are women or people of color. While there isn't sufficient data to demonstrate that riders are likely to be less generous with or more critical of drivers who happen to be members of a protected class, there is no way to evaluate whether these concerns have any merit as a third-party – and that is itself a problem. Through the rating system, consumers can directly assert their preferences and their biases in ways that companies are prohibited from doing.[10] In effect, companies may be able to perpetuate bias without being liable for it.

2015 Privacy Law Scholars Conference, Berkeley, CA.

[6] See http://sanfrancisco.ubermovement.com/resources/ratings-are-a-two-way-street/ for Uber's explanation of how they value and measure progress according to 3 markers: star rating, acceptance rate, and cancellation rate. The rating drivers are required to maintain varies according to the local market as well as the tier of service (e.g. uberX and uberXL require a minimum of 4.6/5 in San Francisco, but UberBlack and UberSUV require a minimum of 4.7/5. Ratings are not the only metric that can lead to a driver's deactivation: Uber also monitors drivers' ride acceptance rate and ride cancellation rate. Drivers must maintain a high ride acceptance rate, such as 80% or 90%, and a low cancellation rate, such as 5%.These targets vary by market. See Rosenblat & Stark http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2686227 for details on how drivers experience the rating system.

[7] See racial discrimination and E-bay sellers: Nunley, John M., Owens, Mark F., and Stephen Howard, R., "The effects of information and competition on racial discrimination: Evidence from a field experiment," *Journal of Economic Behavior & Organization*, Vol. 80, Issue 3, Dec. 2011, p. 670-679, Accessed Dec. 29, 2015, http://www.sciencedirect.com/science/article/pii/S0167268111001739 ; See discrimination in sales of iPods in online classifieds advertisements: Doleac, Jennifer L. and Stein, Luke C.D., "The Visible Hand: Race and Online Market Outcomes," May 01, 2010, Accessed Dec. 29, 2015, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1615149 ; See an evaluation of mechanisms to prohibit fraudulent or discriminatory behavior in reputation reporting systems: Dellarocas, Chrysanthos, "Immunizing Online Reputation Reporting Systems Against Unfair Ratings and Discriminatory Behavior," *Proceedings of the 2nd ACM Conference on Electronic commerce*, 2000, Accessed Dec. 29, 2015, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.63.6968&rep=rep1&type=pdf ; See racial discrimination on Airbnb: Todisco, Michael, "Share and Share Alike? Considering Racial Discrimination in the Nascent Room-Sharing Economy," Mar. 14, 2015, 67 *Stanford Law Review Online 121*, Accessed Dec. 29, 2015, http://www.stanfordlawreview.org/online/share-and-share-alike

[8] Joseph M Stauffer & M Ronald Buckley, The Existence and Nature of Racial Bias in Supervisory Ratings, 90 J. OF APPLIED PSYCHOLOGY 586 (2005).

[9] Rogers, Brishen, "The Social Costs of Uber," *The University of Chicago Law Review Dialogue 82:85,* May 20, 2015*,* Accessed Dec. 29, 2015*,* *https://lawreview.uchicago.edu/sites/lawreview.uchicago.edu/files/uploads/Dialogue/Rogers_Dialogue.pdf,* p. 97-98. Brishen observes, "Minority drivers, to retain high ratings, may need to over-come white passengers' preconceptions, which can involve "identity work," or a conscious effort to track with white, middle class norms."

[10] Thanks to Dr. Ben Edelman for an illuminating discussion on this topic.

In a legal context, courts have long rejected the argument that companies put forward in the 1970s-80s asserting that they did not act in prejudicial ways when they hired or fired workers on the basis of protected-class characteristics, such as gender or age: rather, they were simply implementing consumer preferences.[11] The bulk of jurisprudence on the consumer-preferences arguments falls under the Bone Fide Occupational Qualification[12] exception to Title VII of the Civil Rights Act, which "…was carefully drafted to prevent employers from being able to discriminate against a group based solely on the preferences of customers."[13] Companies cannot justify disparate treatment by saying that they're simply catering to their prejudiced customers. The fact that customers are racists, for example, does not license a company to consciously or even implicitly consider race in its hiring decisions.[14] The problem here is that Uber can cater to racists, for example, without ever having to consider race, and so never engage in behavior that amounts to disparate treatment. Choosing to act on passengers' ratings means that Uber inherits passengers' prejudices and biases, but Uber is not engaged in anything that would even need a BFOQ carve-out. The question is whether Uber should be liable for letting riders' disparate treatment of drivers affect the company's employment decisions.[15] The question raised then is whether or not the defense of "business necessity" of the rating system would be sufficient in this case to justify any potential uneven effects to a rating system that is facially neutral.[16] The anticipated debate that follows from examining this with a "disparate impact" lens is that Uber doesn't get to avoid liability simply because the rating game is set up to be facially neutral – there's a debate over whether not it's "necessary" for the purposes of operating scalable two-sided transportation networks.

Generally, the popularity of rating systems for keeping people and companies accountable for providing good, reputable services has developed across long spectrum of public business-consumer relations, prompting inquiries into rates of fraud and fairness in rating and algorithmic ranking systems.[17] Uber and other companies operative in the on-demand economy that relegate the role of evaluations to consumers through rating systems represents a broadening of platform-consumer relations that hinges on how well the workers in the loop deliver an advertised service. Critics of this mode of accountability generally

---

[11] See e.g., Diaz v. Pan Am. World Airways, Inc. 442 F2d 385 (5th Cir 1971). Reference and quotation found in P. 507 of *Market Definition Analysis and BFOQs.* *http://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=1275&context=uclf*

[12] BFOQ permits intentional discrimination by employers if certain protected-class characteristics are reasonably necessary to the normal operation of a business. E.g. airlines can mandate that pilots retire at age 60 for safety reasons.

[13] Cantor, Rachel L. "Consumer Preferences for Sex and Title VII: Employing Market Definition Analysis For Evaluating BFOQ Defenses", *University of Chicago Legal Forum*, Vol. 1999, Issue 01, Article 12, Accessed Dec. 29, 2015, http://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=1275&context=uclf , p. 507

[14] The BFOQ exception carves out a very limited space for how disparate treatment under Title VII can be evaluated, but the bulk of the consumer-preference jurisprudence falls under BFOQ.

[15] In the Uber system, drivers risk deactivation for low ratings, and some are recommended to take a driver class before they can apply for reactivation. They can also be deactivated based on other "quality control" metrics, such as low ride acceptance rates or high cancellation rates, or behaving in any way that Uber prescribes against in its Terms of Service.

[16] For a relevant discussion of the "business necessity" defense, see: Grover, Susan S., "The Business Necessity Defense In Disparate Impact Discrimination Cases," *Faculty Publications*, Paper 19, 1999, Accessed Dec. 29, 2015, http://scholarship.law.wm.edu/cgi/viewcontent.cgi?article=1038&context=facpubs

[17] Luca, Michael and Zervas, Georgios, "Fake It Till You Make It: Reputation, Competition and Yelp Review Fraud," *Nov. 08, 2013, Accessed Dec. 29, 2013, https://consumermediallc.files.wordpress.com/2014/09/ssrn-id2293164.pdf*

focus on unfair ratings inflation or deflation: the possibility that passengers give positive ratings for bad service,[18] or the grievances drivers raise about receiving negative ratings for factors beyond their control.[19] Indeed, distinctions in interactive service work between the worker and the system are difficult to make,[20] and can impact how drivers are rated. As a managerial strategy, the rating system serves to automate and alert Uber to drivers who are under-performing. Uber, for example, provides drivers with advice on how to improve their behaviors so that passengers give them better ratings.[21] Arguably, since each singular rated trip constitutes a small percentage of a driver's evaluation – and a percentage that decreases in relation to the increase in trips a driver completes, such that the most active drivers are the least impacted by a singular bad rating – one biased passenger does not have a significant impact on drivers' evaluations. It would be easy to dismiss these issues as unimportant because they are aberrations, but *specific* drivers might be subject to systematic bias.[22]

Automated or semi-automated recruitment mechanisms may facilitate reduced discrimination in employment. Uber may be an example of this because it processes initial driver registration through online processes, and is only checking basic qualifications, such as an eligible vehicle and vehicle registration documents. Drivers proceed to an Activation Center after the initial recruitment process has determined their eligibility. Uber's system may thus overcome some of the long-standing problems with discrimination in hiring mangers or amongst dispatcher discrimination, particularly since the dispatch process is automated. However, the way Uber evaluates drivers may introduce a new way for prejudice and bias to affect drivers' employability.

Both algorithms and automated systems have been discussed as salutary in reducing both favoritism, referencing the role of human dispatch operators in assigning work to for-hire vehicle drivers, and in reducing discrimination. The logic of the latter is that automated systems that sort through credentials of employment applicants are less discriminatory than their human counter-parts. Indeed, this logic may well apply to the ways that Uber drivers are recruited and hired. Most of their on-boarding process can be done digitally, such as by uploading their insurance documents and copies of their driver's license, and they communicate almost exclusively with Uber via email. However, the use of ratings to determine employment eligibility may actually push bias downstream, even if it is somewhat remedied at the initial hiring point.

Going Forward: Tracking Bias in Rating Systems & Legal Liability

1. The first step going forward would be to perform a disparate impact analysis: do members of protected classes receive systematically lower ratings? Do they receive these lower ratings even when they resemble other drivers on all the relevant dimensions? If Uber collects information about – say – race as part of its background checks (which are likely outsourced to a third party), the company could easily do this analysis. At the very least, the company has photos of drivers' faces. It's worth

[18] Kane, Kat, "The Big Hidden Problem Problem With Uber? Insincere 5-Star Ratings," *Wired*, March 19, 2015, Accessed Dec. 29, 2015, http://www.wired.com/2015/03/bogus-uber-reviews/
[19] Rosenblat & Stark, 2015, p. 12
[20] Leidner, "Robin. Emotional Labor in Service Work," *Annals of the American Academy of Political and Social Science*, 1999, p. 81.
[21] Rosenblat & Stark, 2015, p. 13
[22] Barocas, Solon and Selbst, Andrew, "Big Data's Disparate Impact," *California Law Review, Vol. 104, 2016,* Aug. 14, 2015, Accessed Dec. 30, 2015

noting that it would be extremely difficult for outside researchers to conduct a statistically rigorous audit.

2. Uber could also track who is deactivated based on low ratings; who amongst them are recommended to reactivation classes by Uber; who attends them; and how their rating fares afterwards. Additionally, it would be interesting to observe where good and bad ratings fall on a spectrum of drivers who are identified by protected-class characteristics. If there is marked bias in customer ratings, the company could pro-actively adjust ratings to make up for the bias it identifies. However, it is unlikely that the company is obliged to do so. Disparate impact doctrine would suggest that an employer consider an alternative approach that achieves the same business goal but reduces the disparity; it would *not*, in general, suggest that the company simply adjust scores.

3. Another possible solution for the company to reduce its role in hiring and firing decisions is to not deactivate drivers whose ratings fall below an acceptable threshold; instead, it could arrange the system such that passengers set the minimum rating at which they are willing to be paired with a given driver, which would effectively empower the consumer without entirely implicating the company in suspension or firing decisions in relation to ratings.

4. Or, a given driver could be given a more diverse set of passenger-reviewers as a system parameter; the system learns rating biases of certain demographics of passengers, and weights them accordingly; the system parses driver ratings in the vicinity and uses it to weight ratings (ex. accounting for bad ratings related to traffic or surge pricing).

Clearly, there is room for a lot of discussion and debate about whether it's worthwhile for the company to engage in proactively seeking out discrimination in its platform; and what steps might be useful in remedying it. I hope to use the workshop to open discussion on this issue, and to brainstorm at the workshop or in follow-up communications about this portentous change to discrimination issues in the workplace in platform-employment via rating systems.

Potential Broader Implications

Public discourse surrounding the impact of growing sophistication in automation tends focus on the issue of displacement: i.e. the types of job that we anticipate will be replaced by machines in the near future. Less discussed is the impact of a nearer-term/present-day scenario: hybrid organizations that blend automation and human work, particularly in the design of platforms that automate management and coordination of workers. Uber is a template for this type of system: a semi-autonomous system which relies on collected user ratings as signals to replace the work of a middle manager in deciding whether to hire or fire human drivers. To that end, the discrimination issues raised in this provocation piece are broader than just Uber. They are potentially latent in the proliferation of automated systems that employ an ad hoc, distributed labor force regulated largely by consumer feedback. The application of Title VII to these situations is significant in part

because it spurs the discussion of what less discriminatory alternatives could and should be in these business models. Each intervention will distribute costs and benefits across all players in the system: riders, drivers, and the platform itself. As Uber-like models continue to multiply, employment discrimination might become hotly contested political ground, which joins the current employee v. contractor debate. From a policy and law perspective - issues like these may in fact be the most immediate need when considering interventions, which address the impact of automation on work and jobs, rather than the raw question of displacement.