Bias and Diversity I: Designing Against Discrimination in Online Markets

Professor Solon Barocas
Cornell University
Department of Information Science

10/2/17

These notes are based off a presentation by Professor Solon Barocas (Cornell, Department of Information Science), with accompaniment from Professor Karen Levy (Cornell, Department of Information Science), for the section on "Bias and Diversity" in the Mechanism Design for Social Good Reading Group. The notes are taken by members of the reading group with some figures and texts taken from the accompanying paper. Questions and comments from reading group members during the presentation are labeled as such. Please contact the reading group organizers for any questions or comments.

- This presentation is based on an upcoming paper by the same title by Karen Levy and Solon Barocas. This paper is not on mechanism design, but mechanism design researchers can have a lot to say about the topics covered here, and based on some of the things that come up, there might be a lot of interesting future directions to explore.
- This paper is a follow up to [2] that started with some colleagues in Data and Society on how bias in ratings of Uber drivers might affect how the platform might then allocate requests to those drivers. People had previously considered how drivers may have been prejudicial in their assessment of riders, and people in the legal community had been thinking about whether there might be legal protections in this case. There had not, however, been as much work about the case when the platform might itself be relying on ratings given to drivers in such a way that the platform inherits some of the bias ratings exhibited by the riders.
- This is an interesting case since discrimination law should cover the case where the employer is discriminating against the workers. (There is of course the question of how to consider Uber drivers since they are considered contract workers and not employees.) Setting that aside, the interesting question is does this count as a case of employment discrimination despite the fact that Uber might not be aware of some of these things. We might think of the analogue that employers cannot cater to the prejudicial tastes of customers. i.e., a reason to justify discrimination might be that if I don't discriminate, then my customers might refuse to come to my store. This argument was dealt with pretty early on in the history of discrimination law as being an unacceptable answer. Is there an analogy here with Uber? It seemed as though there might be a gap in discrimination law in the sense that it wasn't clear it captured these cases where discrimination might be inherited from user ratings, which might, in turn, be used to differentially treat drivers.

- *Question*: Are there past cases where customer surveys were used to evaluate employee performance, bias present there?

  Solon: We thought about this at the time, but we did not find anything along these lines. It seems like an obvious off-line analogy.

- We then wanted to consider more broadly how on-line platforms in general can be a mechanism to address long standing issues of discrimination, but are also vulnerable to inheriting the dynamics that happen off-line. This paper surveys ways in which platforms have some way of making decisions whether intentionally or not, that might mitigate, or exacerbate these dynamics.

- Past work looks at the issue of what information has been presented to users, especially those that are a signal of protected categories. For instance, withholding name, picture, etc. People advocate in favor of this (e.g., Ben Edelman at Harvard), and this type of work is well established in the economics and sociological literature.

- This work: Looks at other remedies as well, and provides a taxonomy of how other problems arise.

- From a legal perspective, one that has attracted a lot of attention is San Fernando Valley v. Roommates.com. This was early on in the Internet. Roommates.com had drop-down menus to indicate race, ethnicity, sexual orientation, etc. It allowed individuals to indicate their race, and so on, as well as request roommates that are of specific race, ethnicity, and sexual orientation.

  There is a law called CDA 230, which provides immunity for platforms from illegal user activity, but this was ruled not to apply to Roommates because the platform was playing an "active role" in deciding what information was provided rather than being a conduit. It was necessary to provide this information to use the website.

- Much less attention is paid to the many things that platforms can do to make discrimination more or less likely.

- We worked with students here at Cornell to survey online platforms in many domains: short-term rentals, long-term housing, ride-sharing, dating, employment, crowd-funding, etc. We wanted to extract common strategies/techniques these platforms use.

- The paper outlines 10 categories all together, bucketed into 3 larger categories: policy (what policies does the platform establish), structure of interactions (what info is shown, matching process), monitoring/evaluating conduct.

- A key argument in this paper is that: platforms can't avoid making these decisions – every choice has some effects.

- The category summaries are:

  1. Setting policies
     (a) Company-level diversity and anti-bias strategies: e.g., increasing diversity within the company workforce, educating employees about bias, engaging underrepresented groups in the design process.

(b) Community composition: restricting community through norms, rules, and structures.

(c) Community policies and messaging: community guidelines, required training on community norms, pledges, language and imagery on and off site.

2. Structuring interactions

(a) Prompting and priming: prompting users to reflect on their behavior at specific decision points.

(b) How users learn about one another: matching users, searching, filtering.

(c) What users learn about one another: encouraging or requiring disclosure of user information, withholding user information, structuring the presentation of user information.

(d) Reputation, reliability, and ratings: testimonials, references, reviews, badges, ratings.

3. Monitoring and evaluating

(a) Reporting and sanctioning: creating mechanisms for user to report biased behavior, sanctioning users who discriminate.

(b) Data quality and validation: requiring more granular information, adjusting ratings, delisting reviews, requiring validation from external data.

(c) Measurement and detection: collecting demographic data to measure disparities in outcome by protected characteristics, experimenting with design to assess effects on bias, opening data to outside scrutiny.

- *Question:* For housing, racial preferences are illegal, but for online dating websites, people can express their racial preferences, and that is seen as being socially and legally acceptable. Are there cases in between and where do you draw the line?

  Solon: For public accommodation (employment, credit, going into a store), things are more straightforward. Less obvious are cases like dating, where preferences are viewed as part of someone's autonomy to make decisions. We are still in the process of figuring out how to think systematically about these things, but it is part of our ongoing work.

- *Question:* We could imagine that for various minorities, there could be sites helping them find roommates within their minority group (e.g., Muslim individuals looking for a roommate in a neighborhood that is known for not being Muslim-friendly or a website for home-schooled children who grew up in fundamentalist Christian households moving to a big city and wanting to be roommates with someone that has the same background.) You can see all of these pulling in opposite directions, and some of these can be seen as understandable. You can sort of think Roommates.com as being a one-stop shop where instead of using several websites for different demographic groups, you can have one website that everyone uses and can specify that roommates they are looking for. If we consider the other websites to be OK, is it the bundling all of these together that makes what Roommates.com was doing problematic?

  Karen: It is unclear whether the individual sites would be considered legal. And the bundling is also an issue.

Comment: Under Seattle law, there's a difference in looking for roommates vs. advertising housing, i.e. you can't only advertise housing to women, but you can look for a female roommate.

Karen: Each state can have different laws, some more stringent. Generally speaking, there's more leeway when you're looking for a roommate/sharing your own home (known as the Ms. Murphy exception), but this is still not sufficient in Roommates case, since it was structuring transactions in a way that might be violating the law. As to the general issue of whether you can structure a site for a particular group of people, it's not clear that this would be legal.

Question: So the issue with Roommates.com wasn't fair housing, but instead that it provided information about protected attributes?

Karen: The issue wasn't just the fact that information was present – the issue was that Roommates.com solicited and conditioned participation on provision of this information. The platform could allow you to provide that content, but structuring the site in a way that required it made CDA 230 not apply. If information is obviously voluntarily provided by user, the site isn't liable. (e.g., if there was a free text box and I disclosed in that my religion and stated that I was looking for roommates that shared that religion, Roommates would not be on the hook for that.)

- *Question:* What if the site provided a free text box, but users come up with some convention for what to write and how to search if you were a particular minority looking for someone else in that minority?

  Karen: That's probably ok. For instance, Craigslist has no structure and is generally protected under CDA 230. If the website came up with the back-door instruction, then that would again not necessarily be covered under CDA 230.

  Solon: It could still be a violation of Fair Housing Act by the users. There are two separate questions: does CDA 230 protect the site? And is the content in violation of the Fair Housing Act?

- *Question:* Housing is particularly protected by Fair Housing Act, but dating is obviously not covered by this. Is there some more general anti-discrimination ordinance that it might be protected by?

  Solon: Discrimination is legislated on a sector-by-sector basis. Housing, employment, credit, insurance, public accommodation, etc. all have their own laws, but no analogous law exists for dating. For most platforms considered in this piece apart from dating, there's a particular anti-discrimination law that should apply.

  Question: So there is no wide-reaching non-discrimination law that applies to everything?

  Solon: Yes, and existing laws also explicitly depend on which protected attributes we consider. Note these features have been growing over time to reflect what we consider to be attributes that should be protected under the law.

- *Comment:* CS people often try to generalize the spirit of some anti-discrimination law to another domain, so it's useful to appreciate that they're actually separate and don't apply to across domains.

Solon: True, though our way of reasoning about discrimination is informed by existing laws, and it would be hard to imagine them formalized in another way. Also, rules like the 4/5 rule from employment have been generalized to provide guidance in other areas like housing, so they are tied together.

- Going back to the categories, one thing to consider is how users are matched to one another. Different ways of doing this can create more/fewer opportunities for bias. If users can select among each other, this is different from the platform doing the matching automatically. Example: Airbnb, to combat against discrimination for black users, used the option of "Instant Book." Instant book listings are those that do not require approval from the host before they can be booked. Airbnb incentivizes hosts to make their listings instant books. This is not quite automatic matching, since person chooses where they stay, but this is now more comparable to booking a hotel room.

- Gig/employment platforms: Sometimes employers choose who they want, sometimes workers choose which job they want to take, and this impacts potential for bias. If employers choose workers to contact, for instance, they're in a position to only contact certain people and discriminate against others. If people apply to generally announced jobs, this doesn't eliminate the possibility of bias, but at least people get the chance to make contact with one another.

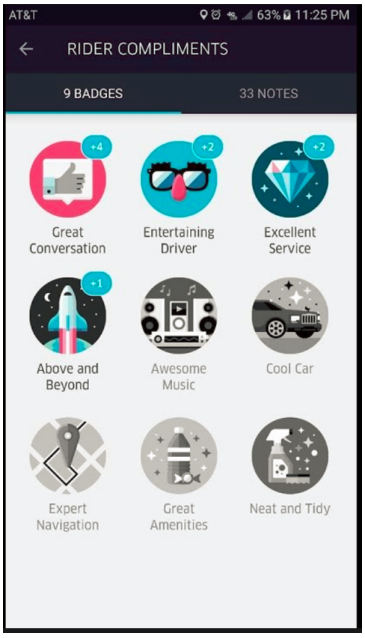
- Karen: One big category that has garnered the most attention is changing the information structure: what information gets made salient? Photos? Names? Adding personal info is supposed to build trust, but can also reveal protected attributes that people can use to discriminate. The knee-jerk reaction is to remove those pieces of information or add them at a later stage of the transaction (e.g., after matches have been accepted.) However, this can backfire – when you withhold information, people might instead try to use proxies to discriminate instead. For instance, Ban the Box [1]: An effort to remove questions about criminal records from hiring forms. Counterintuitively, this has lead to worse outcomes for people of color because people use race as a proxy for whether people have a criminal record. So, statistical discrimination can have more of an impact in this context.

- We want to disentangle what are instances of taste-based and statistical discrimination. The argument has been that for statistical discrimination, we should provide the information that people are trying to infer via proxies. In the above case, that would mean providing information about criminal records so that people don't use race a as a proxy for it. Another way to handle this is for Airbnb to offer insurance to cut down on statistical discrimination.

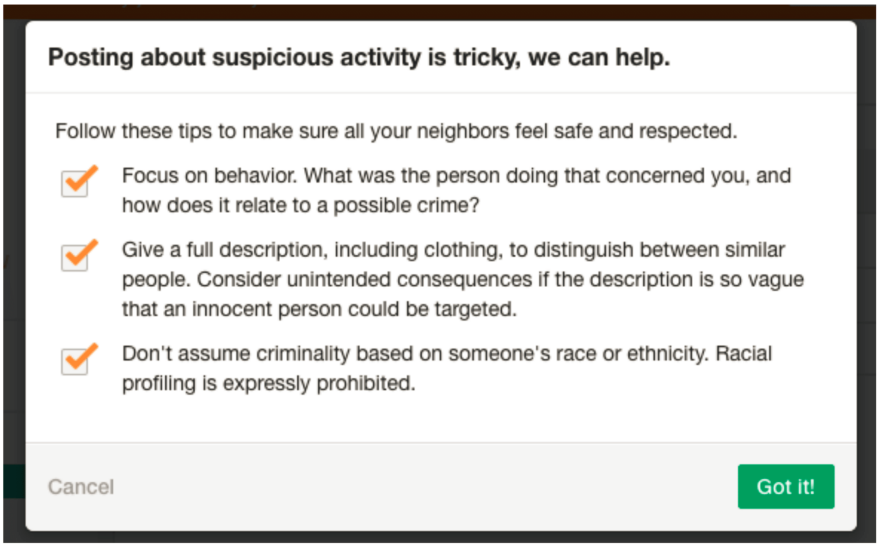
- *Comment:* There are three variables here: criminal record, race, thing that I'm using as a proxy for both race and criminal record, and I need to reason about all three of these.

- Another issue is reputation systems: these have been generally successful in enabling trust between two parties who don't know one another, but they could also be a vehicle for prejudice. How do you compensate for that? Sometimes, platforms solicit feedback in quick and easy way (i.e. star ratings). Usually this isn't very high-quality. For star ratings, we see lots of 1, 4, 5 but not a lot of 2, 3 (J-shaped distribution, where there is a tendency to the extremes, and especially to positive extremes). Coarse reviews don't convey the actual reason for the assessment. There is tension between soliciting more information and getting

people to actually use the system at all. Uber, for instance, asks for optional additional feedback on top of a 5-star review in the form of several categories, and does the same for negative reviews.



Nextdoor.com had a problem with its "report suspicious activity" feature, which just ended up being use to report non-white people in the neighborhood. They used a coarse mechanism that didn't ask for what was actually suspicious. The platform changed to ask for more explicit feedback (e.g. describe the person, describe what was suspicious, etc.).



- *Question:* Do driver ratings actually have a major impact on Uber? How are these ratings

used, i.e., does the impact come from Uber not matching me with a lower-rated driver, or am I more likely to reject that match?

Solon: A number of different things can happen: a lower rating means you're matched less frequently (by Uber), and if your rating drops too low, you get kicked off the platform or have to go on some sort of remedial training. It depends on the current demand/supply, but Uber has a preference for matching higher-rated drivers.

Comment: It is hard to tell such fine-grained differences in rating as a user, and it seems an odd intervention for the platform to treat users 4.6 and 4.7 or 4.2 and 4.6 differently.

Comment: It might not be at the individual level, but in aggregate it could have some large impact on customer retention. (e.g., if you go with a 4.6 versus a 4.7 there is a slightly larger chance that you have a bad experience and stop using the platform.)

- *Question:* Platforms could incentivize feedback (free rides, etc.) to get more fine-grained feedbacks and increase engagement. Have there been instances of this and how well did these work?

  Solon: We have not come across anything that induces feedback.

- *Question:* What is your takeaway from Uber 5-star system – is it too coarse?

  Solon: It is very coarse and therefore more vulnerable to inheriting bias/prejudice. More explicit prompts can be a way to mitigate this. Nextdoor: reduced number of reports, but not clear whether this was because it cut down on reports due to prejudice vs. burden of completing the process increasing.

- *Question:* Does making reporting more burdensome just skew the feedback even more to the extremes?

  Solon: Good question, we have not come across this and it's not clear.

  Comment: Changes to Uber were a response to coarse feedback, and an attempt to get more information.

  Karen: Uber made the system more granular over time, even as we were writing this paper.

- Airbnb commissioned a report on bias in its system, which said that they wanted to use machine learning to deal with this problem. Platforms wanted to be able to model whether its users are biased based off of their decisions. It's unsettling that companies want to be in the business of trying to determine how racist/sexist its users are.

- There is a considerable amount of pressure on platforms to at least consider if there is variance in outcomes by class membership. For instance, do acceptance rates requesting an Uber differ across groups? These types of questions can't establish bias because we don't know the underlying distributions and how they differ across groups – it could be that differences exist because of differences in some relevant characteristics across groups. We need controlled studies to actually establish a link. So, people have tried to do controlled studies with individuals that look essentially identical except for varying by a protected attribute and see whether they are suffering unfair treatment.

  Question: Is there enough data to do studies like this?

  Solon: I don't know; they haven't talked publicly about it yet.

- Conversation right now is a bit confused, and people take any differences in rates at which things happen as being evidence that there is discrimination. An extreme version of this: conduct user studies to measure implicit bias of users, and use that as training data to build a model of bias based on users' features.

- *Comment:* Say some male Uber drivers are sexist, but only women riders experiences this. The model might say that women are biased against men because they give them lower ratings, but it might actually be because men are biased. It is hard to tell which side the bias is coming from by just looking at data.

  Solon: Good point, this wasn't considered in the paper.

  Karen: One thing that is related is: de-biasing ratings is hard when ratings are just about your subjective experience. This is ethically tricky and conceptually slippery. For example, assume that you're rating your Uber experience and maybe certain minority drivers don't have as nice cars in general or foreign drivers don't speak English as well. These can be highly correlated with protected attributes, but it's hard to say these aren't legitimate bases for reviews.

- *Question:* Is it fair to say that there is a three-way trade-off between maximizing trust in the platform, minimizing discrimination, and minimizing user effort? You have some space of information, you can choose a subset of it to show, and you can solicit more information for some cost.

  Solon: Sure, it's not clear how to resolve this or how to formalize this. It would be interesting to empirically study this. Nextdoor would be a good case for this. The paper is not advocating for any particular method, but there are a whole host of things these platforms can do. It would be good to use some of them in some strategic manner. This paper is overall meant to raise concerns.

- *Question:* We might be able to take advantage of the structure of discrimination, where people who experience discrimination may only experience it from a subset of the population, and the same is true for people who discriminate. One would want to try to avoid penalizing non-discriminatory users.

  Karen: Yes, though this might lead to another ethically dubious situation where platforms are deciding who their most discriminatory users are and taking action against them.

- Solon: How have people in mechanism design thought about dealing with objectionable preferences while also satisfying acceptable ones? Or more broadly, how have people thought about discrimination?

  Comment: This has been thought about somewhat in school choice literature, i.e., should we be imposing certain types of preferences on students? There have not been conclusive answers.

  Comment: Inherently, the whole point of mechanism design is that the people in the system want something different from what the designer wants. Discriminatory preferences are just one instantiation of this. Models are rich enough to encompass this, but we might need a more specific understanding of how the preferences are misaligned in order to be able to draw any interesting conclusions. This misalignment is probably different from the models we

typically study, and would raise new questions that we may not already have the answers for, but we might have reasonable intuitions for how to approach these questions. The important question is how discriminatory preferences differ from the selfish agents we typically study.

- *Comment:* You don't necessarily want your mechanism to produce outputs that match how you want your societal distribution to look. For example, with public health funds you may not want to equally distribute money for healthcare even though ultimately you want everyone in your society to receive care. Also, we need to consider the question of moral entitlement, and why we believe that across certain attributes, everyone should have the same moral entitlement to a particular good. Metrics might better serve as diagnostic tools to determine whether discrimination is happening rather than things to optimize.

References

[1] Jennifer L. Doleac and Benjamin Hansen. Does ban the box help or hurt low-skilled workers? statistical discrimination and employment outcomes when criminal histories are hidden. Working Paper 22469, National Bureau of Economic Research, July 2016.

[2] Alex Rosenblat, Karen EC Levy, Solon Barocas, and Tim Hwang. Discriminating tastes: Uber's customer ratings as vehicles for workplace discrimination. *Policy & Internet*, 9(3):256–279, 2017.