Tensions and Trade-offs in Designing Against Discrimination

Karen Levy
Cornell University
Roadmap

1. (quickly) the law and platform-mediated discrimination
2. designing against discrimination
3. tensions and trade-offs
User-to-user discrimination on platforms

Rideshare

Ge et al. 2016

Markets for goods

Doleac & Stein 2013, Ayres et al. 2015, Kricheli-Katz & Regev 2016

Short-term rental

Edelman et al. 2017; Wang et al. 2015

Peer-to-peer lending

Pope & Sydnor 2011

Dating markets

Mendelsohn et al. 2014, Rudder 2014

(and probably others!)
Law isn’t particularly useful here

Not all domains covered by federal discrimination statutes (though some states are broader)

Platforms *generally* immune from liability under CDA 230

By deferring decisions to users, companies may avoid disparate impact liability
Discriminating tastes

Rideshare firms make employment decisions based on ⭐⭐⭐⭐⭐ ratings

⭐⭐⭐⭐⭐ ratings very likely to exhibit bias in aggregate

Distributed ratings may provide new avenue for bias to creep into employment decisions

So we might look to design

A first-order question: what do platforms do? (descriptive, not evaluative)

“Design” interpreted broadly:

UI elements

market mechanisms

policies and practices
10 strategies for designing against discrimination

<table>
<thead>
<tr>
<th>Setting policies</th>
<th>Company-level diversity and anti-bias strategies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Community composition</td>
</tr>
<tr>
<td></td>
<td>Community policies and messaging</td>
</tr>
<tr>
<td>Structuring interactions</td>
<td>Prompting and priming</td>
</tr>
<tr>
<td></td>
<td>How users learn about one another</td>
</tr>
<tr>
<td></td>
<td>What users learn about one another</td>
</tr>
<tr>
<td></td>
<td>Reputation, reliability, ratings</td>
</tr>
<tr>
<td>Monitoring and evaluating</td>
<td>Reporting and sanctioning</td>
</tr>
<tr>
<td></td>
<td>Data quality and validation</td>
</tr>
<tr>
<td></td>
<td>Measurement and detection</td>
</tr>
</tbody>
</table>

Bias on intimate platforms

Intimate exchanges are markets too!

Individual decisions aggregate into systematic sorting and segregation

Could (and should) platforms mitigate intimate biases?
Tension #1: more information vs. less information
More information

“A whole person”

More disclosure → more trust (Ma et al. 2017)

Counterstereotypical information as disarming mechanism (Steele, *Whistling Vivaldi*)
More information: reliability, reviews, ratings

Authenticatable information (verified users)

When black and white Airbnb guests each have one positive review, acceptance rates equalize (Cui et al. 2017)

But reviews and ratings can also be inflected by bias
Less information

Purposeful withholding, e.g. photos and names (Edelman et al. 2016; Goldin & Rouse 2000)

But statistical discrimination may persist via fall-back on available data—e.g. ban-the-box (Doleac 2016), eBay (no name, photo, or gender, but still women do worse; Kricheli-Katz & Regev 2016)
(Sort of) less information

Daddyhunt stigma-free pledge:
Sends message about community norms
Allows users to learn something about one another, but not so much as to be stigmatizing (plausible deniability)
Tension #2: granular information vs. user burden
More explicit deliberation → less reliance on crass heuristics/implicit bias
Nextdoor: if race is used in report of suspicious activity, users prompted to fill in additional fields
25% reduction in reports
Tension #3: validation data vs. invasive surveillance
Measure behavior directly (sensors, cameras, etc.)

Tie rewards to specific performance criteria, reducing reliance on user-provided data

Corroborate/adjust user-provided data in cases of complaint
But...

Can fix one problem while creating another

Surveillance will almost always be of less powerful party, used for discipline as well as anti-bias

Security risks; consent problems
Tension #4: 
stated preferences vs. revealed preferences
How do platforms decide whom to match?

Should platforms privilege behavioral data or stated intention? (Ekstrand & Willemsen 2016; Yang et al. 2019)

Should platforms privilege the user who exists, the user she aspires to be... or the user the platform thinks she should be?
Is “no preference” a preference?

“The Dating App That Knows You Secretly Aren’t Into Guys From Other Races

Even if you say “no preference” for ethnicity, the dating app tends to show you people of your own race.

Katie Notopoulos
BuzzFeed News Reporter

Posted on January 14, 2016, at 11:44 a.m. ET

“Our data shows even though users may say they have no preference, they still (subconsciously or otherwise) prefer folks who match their own ethnicity. It does not compute “no ethnic preference” as wanting a diverse preference.”
Tension #5: user agency vs. paternalism
Platforms may want to maximize user autonomy and avoid intervention...

... but they have no choice but to choose (Gillespie 2010)

What does it mean to debias an ambiguous, subjective rating?

Domains where it’s more or less appropriate to intervene? Categories?
