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

Alex Rosenblat, Karen Levy, Solon Barocas, and Tim Hwang. 2017. "Discriminating Tastes: Customer Ratings as Vehicles for Bias." *Policy & Internet* 9(3): 256–279.

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

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

Karen Levy and Solon Barocas. 2017. “Designing Against Discrimination in Online Markets.” *Berkeley Technology Law Journal* 32(3): 1183–1237.

Bias on intimate platforms

Intimate exchanges are markets too!

Individual decisions aggregate into systematic sorting and segregation

Could (and should) platforms mitigate intimate biases?

Jevan Hutson, Jessie Taft, Solon Barocas, and Karen Levy. 2018. “Debiasing Desire: Addressing Bias and Discrimination on Intimate Platforms.” Proceedings of the ACM Conference on Computer-Supported Cooperative Work (CSCW) 2(1): Article 73.

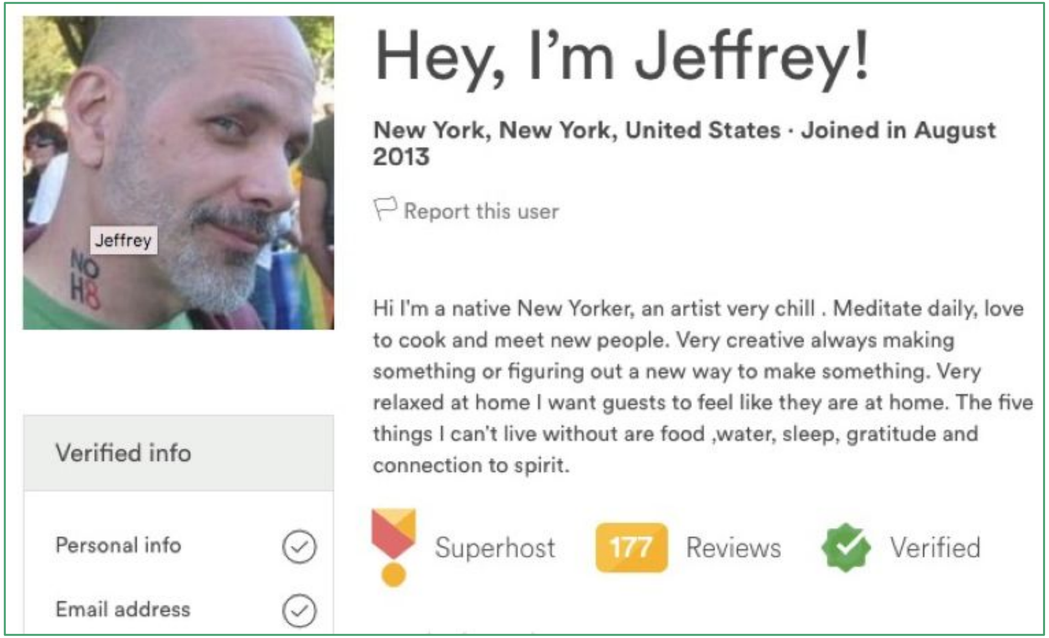
Tension #1:
more information vs.
less information

More information

“A whole person”

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

Counterstereotypical information as disarming mechanism (Steele, *Whistling Vivaldi*)



Hey, I’m Jeffrey!

New York, New York, United States · Joined in August 2013

Report this user

Hi I'm a native New Yorker, an artist very chill . Meditate daily, love to cook and meet new people. Very creative always making something or figuring out a new way to make something. Very relaxed at home I want guests to feel like they are at home. The five things I can't live without are food ,water, sleep, gratitude and connection to spirit.

Superhost 177 Reviews ✓ Verified

Verified info

Personal info ✓

Email address ✓

More information: reliability, reviews, ratings

Authenticatable information (verified users)

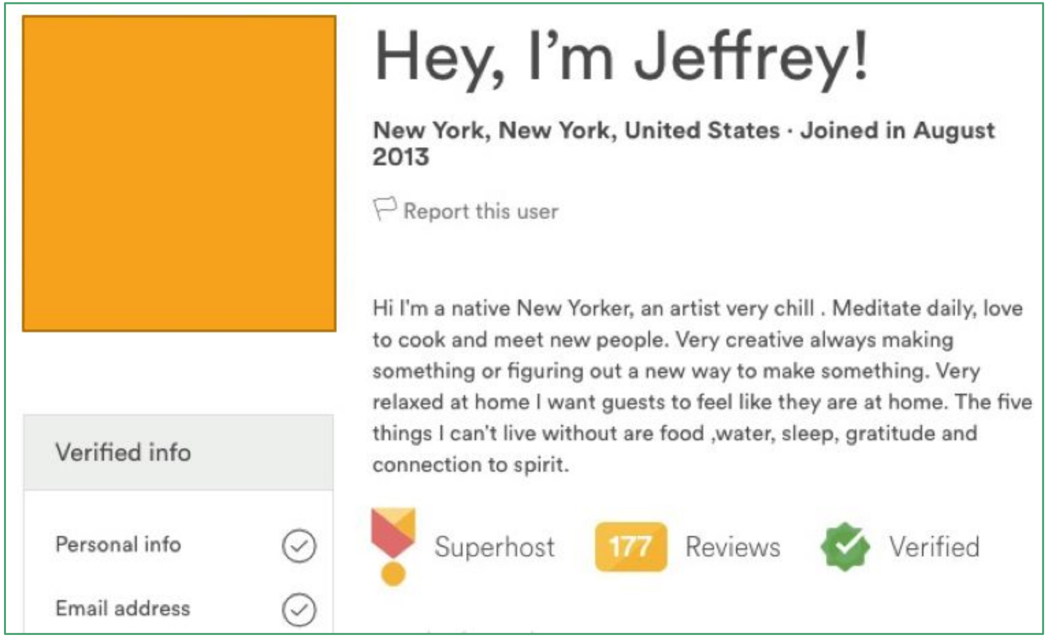
When black and white Airbnb guests each have one positive review, acceptance rates equalize (Cui et al. 2017)

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)



Hey, I'm Jeffrey!

New York, New York, United States · Joined in August 2013

Report this user

Hi I'm a native New Yorker, an artist very chill . Meditate daily, love to cook and meet new people. Very creative always making something or figuring out a new way to make something. Very relaxed at home I want guests to feel like they are at home. The five things I can't live without are food ,water, sleep, gratitude and connection to spirit.

Superhost 177 Reviews ✓ Verified

Verified info

Personal info ✓

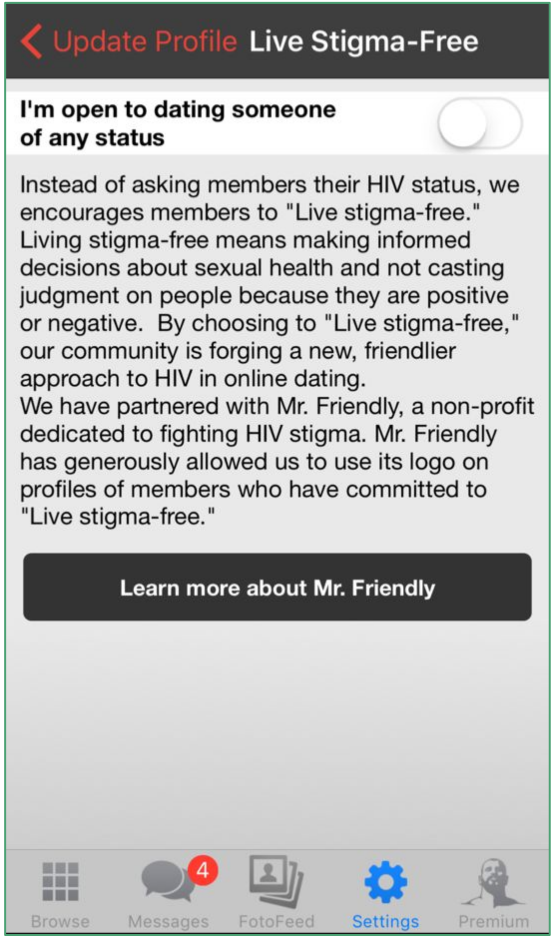
Email address ✓

(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)



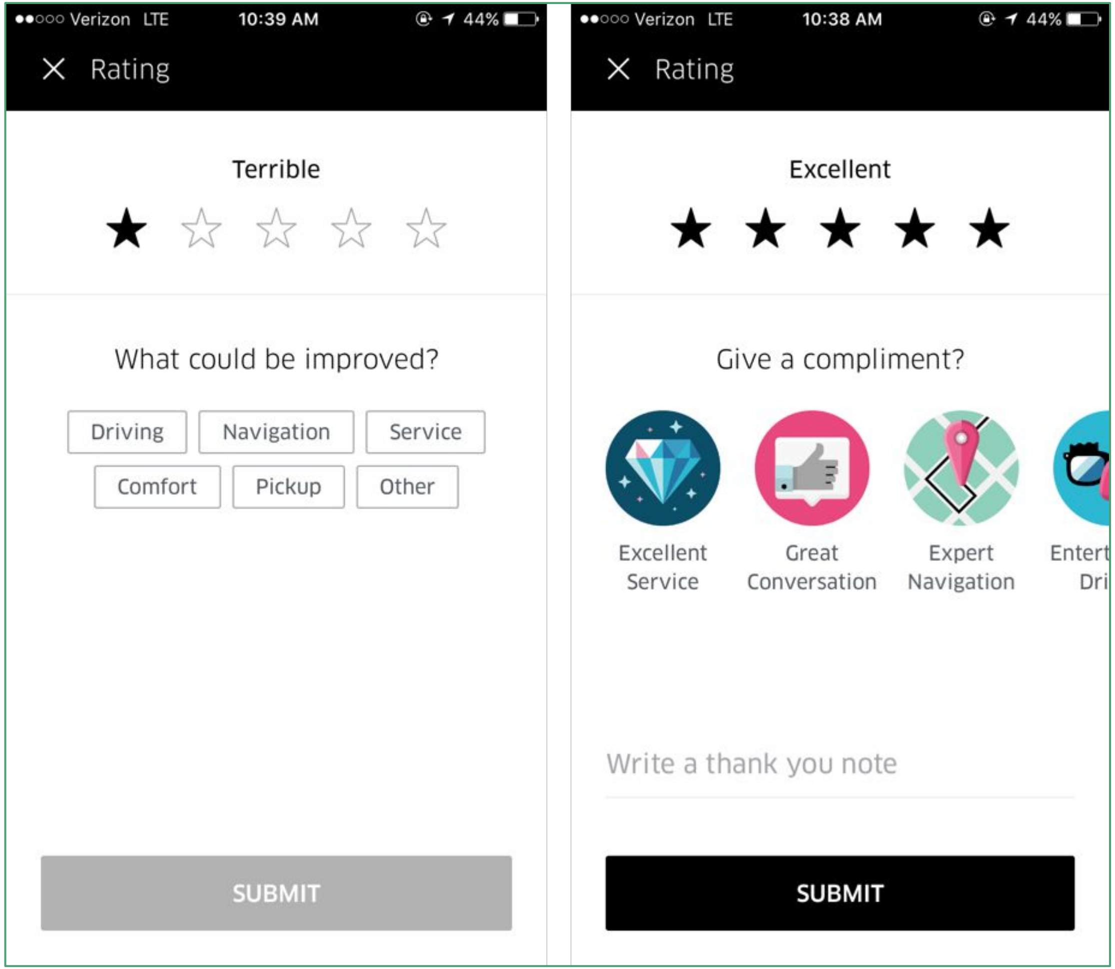
I'm open to dating someone of any status

Instead of asking members their HIV status, we encourages members to "Live stigma-free." Living stigma-free means making informed decisions about sexual health and not casting judgment on people because they are positive or negative. By choosing to "Live stigma-free," our community is forging a new, friendlier approach to HIV in online dating.

We have partnered with Mr. Friendly, a non-profit dedicated to fighting HIV stigma. Mr. Friendly has generously allowed us to use its logo on profiles of members who have committed to "Live stigma-free."

Tension #2:
granular information vs. user burden

More explicit deliberation → less reliance on crass heuristics/implicit bias



●●○○○ Verizon LTE 10:39 AM 44%

× Rating

Terrible

★ ☆ ☆ ☆ ☆

What could be improved?

Driving | Navigation | Service | Comfort | Pickup | Other

SUBMIT

●●○○○ Verizon LTE 10:38 AM 44%

× Rating

Excellent

★ ★ ★ ★ ★

Give a compliment?

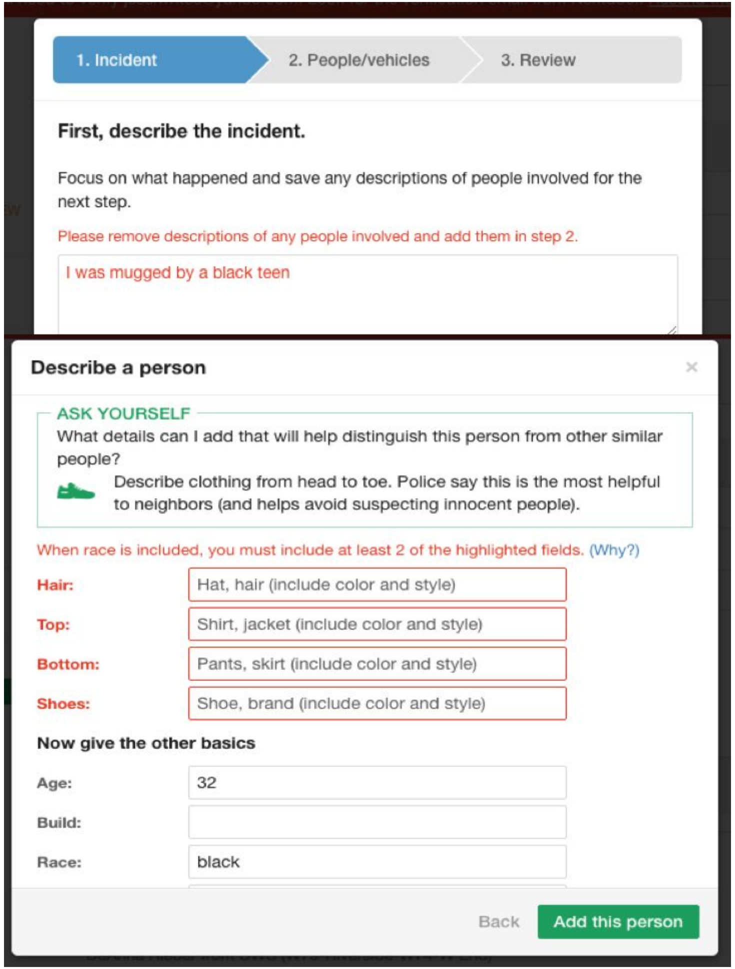
Excellent Service | Great Conversation | Expert Navigation | Entert Dri

Write a thank you note

SUBMIT

Nextdoor: if race is used in report of suspicious activity, users prompted to fill in additional fields

25% reduction in reports



First, describe the incident.

Focus on what happened and save any descriptions of people involved for the next step.

Please remove descriptions of any people involved and add them in step 2.

I was mugged by a black teen

Describe a person

ASK YOURSELF

What details can I add that will help distinguish this person from other similar people?

Describe clothing from head to toe. Police say this is the most helpful to neighbors (and helps avoid suspecting innocent people).

When race is included, you must include at least 2 of the highlighted fields. (Why?)

Hair:	Hat, hair (include color and style)
Top:	Shirt, jacket (include color and style)
Bottom:	Pants, skirt (include color and style)
Shoes:	Shoe, brand (include color and style)

Now give the other basics

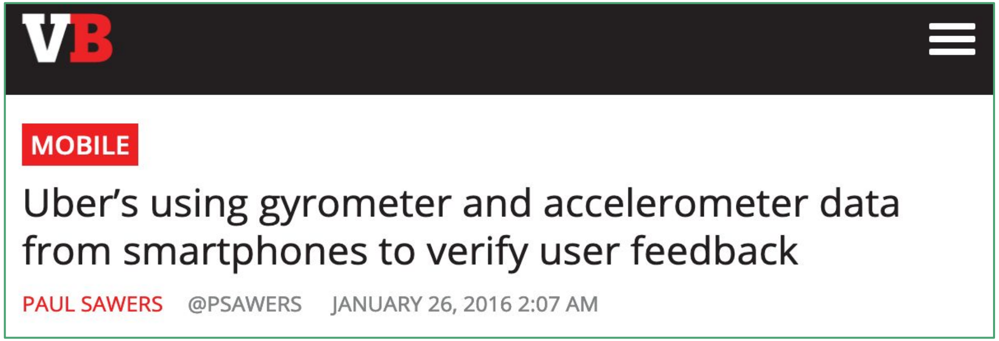
Age:	32
Build:	
Race:	black

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

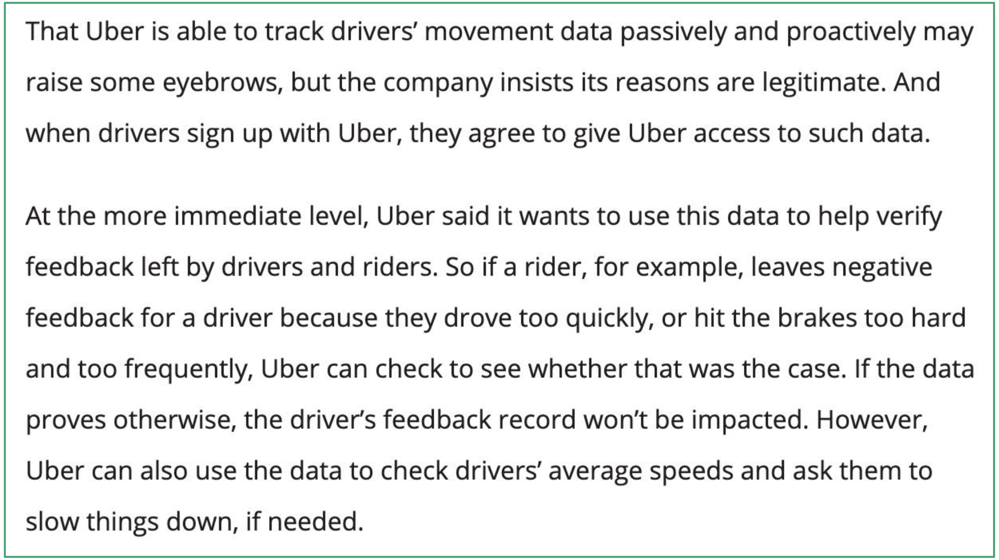


VB

MOBILE

Uber's using gyrometer and accelerometer data from smartphones to verify user feedback

PAUL SAWERS @PSAWERS JANUARY 26, 2016 2:07 AM



That Uber is able to track drivers' movement data passively and proactively may raise some eyebrows, but the company insists its reasons are legitimate. And when drivers sign up with Uber, they agree to give Uber access to such data.

At the more immediate level, Uber said it wants to use this data to help verify feedback left by drivers and riders. So if a rider, for example, leaves negative feedback for a driver because they drove too quickly, or hit the brakes too hard and too frequently, Uber can check to see whether that was the case. If the data proves otherwise, the driver's feedback record won't be impacted. However, Uber can also use the data to check drivers' average speeds and ask them to slow things down, if needed.

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

Home

Case Study: The Gig Economy and Exploitation

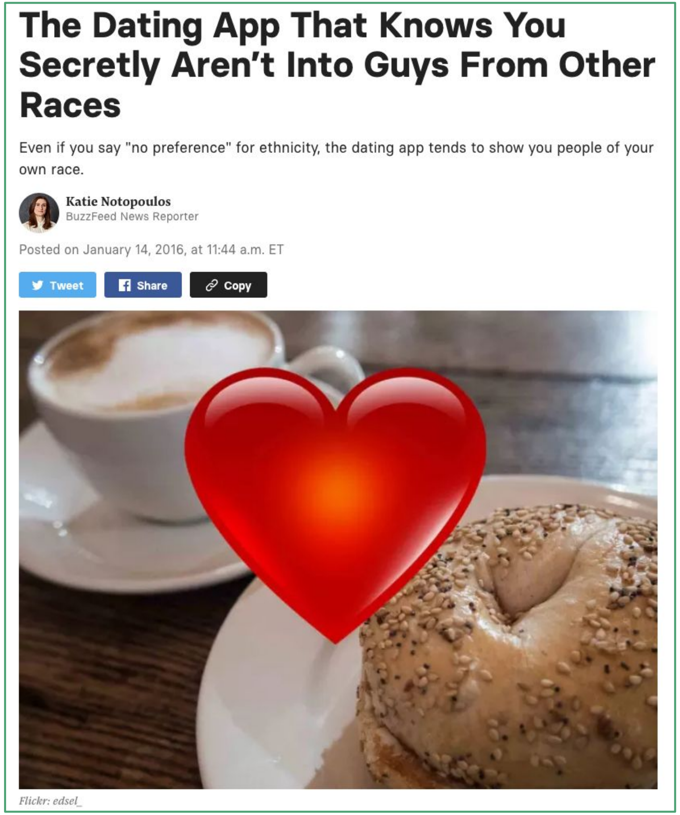
Tension #4:
stated preferences vs.
revealed preferences

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

Flickr: edsel_

“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?

Karen Levy and Solon Barocas. 2017. “Designing Against Discrimination in Online Markets.” *Berkeley Technology Law Journal* 32(3): 1183–1237.

Alex Rosenblat, Karen Levy, Solon Barocas, and Tim Hwang. 2017. “Discriminating Tastes: Customer Ratings as Vehicles for Bias.” *Policy & Internet* 9(3): 256–279.

Jevan Hutson, Jessie Taft, Solon Barocas, and Karen Levy. 2018. “Debiasing Desire: Addressing Bias and Discrimination on Intimate Platforms.” Proceedings of the ACM Conference on Computer-Supported Cooperative Work (CSCW) 2(1): Article 73.