DESIGNING AGAINST DISCRIMINATION IN ONLINE MARKETS

Karen Levy[†] *& Solon Barocas*[††]

ABSTRACT

Platforms that connect users to one another have flourished online in domains as diverse as transportation, employment, dating, and housing. When users interact on these platforms, their behavior may be influenced by preexisting biases, including tendencies to discriminate along the lines of race, gender, and other protected characteristics. In aggregate, such user behavior may result in systematic inequities in the treatment of different groups. While there is uncertainty about whether platforms bear legal liability for the discriminatory conduct of their users, platforms necessarily exercise a great deal of control over how users' encounters are structured—including who is matched with whom for various forms of exchange, what information users have about one another during their interactions, and how indicators of reliability and reputation are made salient, among many other features. Platforms cannot divest themselves of this power; even choices made without explicit regard for discrimination can affect how vulnerable users are to bias. This Article analyzes ten categories of design and policy choices through which platforms may make themselves more or less conducive to discrimination by users. In so doing, it offers a comprehensive account of the complex ways platforms' design and policy choices might perpetuate, exacerbate, or alleviate discrimination in the contemporary economy.

DOI: https://doi.org/10.15779/Z38BV79V7K

† Assistant Professor of Information Science, Cornell University; Associated Faculty, Cornell Law School.

†† Assistant Professor of Information Science, Cornell University. We are immensely grateful to our team of research assistants—Jevan Hutson, Jessie Taft, and Olivia Wherry—for their stellar assistance with both data collection and synthesis. In addition, we thank Matt Cagle, Anna Lauren Hoffmann, Amy Krosch, Nicole Ozer, David Pedulla, and participants in the 21st Annual BCLT/BTLJ Symposium, Platform Law: Public and Private Regulation of Online Platforms, for helpful feedback.

TABLE OF CONTENTS

Web-based platforms frequently make a range of socially salient characteristics available to transacting parties. Names and photos, for example, are a standard feature of user profiles on platforms that aim to connect buyers and sellers, hosts and guests, drivers and riders, and all manner of online daters. Such details have a long-standing place on platforms as mechanisms to establish trust between strangers transacting online.

Design features like these have allowed platforms that function as online marketplaces to flourish. The early web was marked by considerable uncertainty as to the reliability of the person on the other side of some exchange. Today's web is dominated by platforms that employ a diverse set of techniques to relieve users of such anxieties—providing assurances that can go far beyond what people might glean from in-person interactions.

In adopting these techniques, however, platforms have begun to exhibit the sorts of worrisome dynamics that are common in face-to-face encounters. Platforms that highlight users' socially salient characteristics invite users to take these characteristics into account, even when they might not—or should not—be relevant to the interactions facilitated by the platform. Names and photos, for example, can reveal users' gender, race,

ethnicity, national origin, religion, age, or disability, among other details. Such details have allowed users to discriminate against one another—either by conscious choice or unconsciously due to implicit bias. When these details are made more prominent, more readily available, or simply unavoidable, they can affect users' behavior in ways that correspond to established patterns of bias in offline markets: users may refuse to transact, make less attractive offers, or evaluate each other less favorably.

When a customer enters a store, a job applicant submits a resume, a passenger flags down a cab, a potential tenant visits a property, or a person strikes up a conversation at a singles bar, the person cannot help but reveal characteristics that might lead to biased impressions. In contrast, online platforms have far more control over how these encounters are structured.[1] Platforms mediate interactions in ways that can both mitigate and aggravate bias. Design and policy choices make platforms more or less conducive to discrimination in user-to-user interactions. Platforms cannot divest themselves of this power; even choices made without explicit regard for discrimination can affect how vulnerable users are to bias. This is true even when pursuing other laudable goals like attempting to ensure greater trust, smoother interactions, or a more efficient market among users.

At the same time, platforms can *purposefully* attempt to address the role of bias in users' exchanges—for instance, by stripping interactions of obvious visual or verbal cues, allowing users to transact in relative anonymity. Ride-hailing services, like Uber and Lyft, have touted features of their platforms that make it difficult or impossible for drivers to discriminate when choosing whether to make a pick-up. Drivers do not learn the identity or intended destination of riders until drivers accept a request.[2] Platforms can decide what information passes through their channels and thus limit the flow of information upon which discrimination depends.

---

1. Ray Fisman & Michael Luca, *Fixing Discrimination in Online Marketplaces*, HARV. BUS. REV., Dec. 2016, at 88.

2. Johana Bhuiyan, *Uber and Lyft Position Themselves as Relief from Discrimination*, BUZZFEED (Oct. 7, 2014, 11:05 AM), https://www.buzzfeed.com/johanabhuiyan/app-based-car-services-may-reduce-discrimination-issues-tk. Note that Lyft shows drivers the names and photos of passengers once drivers accept riders' request, which allows Lyft drivers to then cancel the rides once they have learned about passengers. However, Lyft passengers are not required to upload photos, so this dynamic only applies when passengers have volunteered photos of themselves. In contrast, Uber never shows photos to drivers. Eric Newcomer, *Study Finds Racial Discrimination by Uber and Lyft Drivers*, BLOOMBERG (Oct. 31, 2016, 10:51 AM), https://www.bloomberg.com/news/articles/2016-10-31/study-finds-racial-discrimination-by-uber-and-lyft-drivers.

But platforms' role in modulating the extent to which users might discriminate against one another extends well beyond deciding what users can learn about each other. Platforms also decide *how* users learn about each other: users might receive recommendations, perform searches, or filter according to fixed criteria. Some platforms decide *who* can join these online communities, conditioning entry on various facets of users' offline identities. And many decide to adopt mechanisms that allow users to rate and comment upon one another for others to see—the basis of reputation systems that are ubiquitous online. These choices structure users' encounters and interactions in particular ways, even when platforms see themselves as nothing more than passive conduits through which users engage with one another.[3] Platforms that mediate between users necessarily moderate how users behave on these platforms, including how easily users can fall victim to their implicit biases or how effectively they can impose their prejudicial beliefs on others. Platforms are thus in a privileged and difficult position. The *ability* to mediate interactions between users may create a perceived *responsibility* to do so—even if a platform does not bear legal liability for users' biased behavior.

Platforms that recognize the influence they wield over their users—including platforms that have been pressured to acknowledge such influence—tend to rely on a diverse set of mechanisms to minimize the degree to which users can engage in discriminatory conduct: company initiatives that aim to increase sensitivity to issues of discrimination by cultivating greater workplace diversity and fostering inclusion; the development, adoption, and championing of community policies that forbid or repudiate any discrimination on the part of users; direct attempts to intervene in the process by which users' prejudices or implicit biases enter into their decision-making, involving prompts and priming; the use of additional sources of data to validate users' claims; or introducing systems for users to report instances of discrimination and impose corresponding sanctions. Some have even begun to track disparities in users' experiences on the platform, according to their race, gender, or other protected characteristics, and to identify specific cases of prejudicial or biased decision-making.

Platforms' role in mediating users' discriminatory behavior is complicated by the currently unresolved application of law in this area. Numerous user-to-user platforms operate in domains traditionally within

3. *See* Tarleton Gillespie, *The Politics of 'Platforms'*, 12 NEW MEDIA & SOC'Y 347, 352–353 (2010).

the reach of antidiscrimination law, like housing and employment.[4] Within these domains, however, the applicability of anti-discrimination law to platforms—and specifically to users' interactions among themselves *as mediated by* platforms—is unsettled. Despite their immense power to shape a wide variety of interactions integral to social and economic life, platforms' business models have enabled them to largely sidestep the traditional regimes that protect against discrimination and other harms in those interactions.[5] Platforms routinely disclaim legal responsibility for all kinds of harms propagated by their users against one another,[6] and have largely been successful in so doing.

Despite this, recent calls aim to extend liability to platforms for underlying user conduct within subject domains where existing civil rights law prohibits discrimination. Notably, Belzer and Leong argue that public accommodation laws (including Title II of the Civil Rights Act of 1964 and the Fair Housing Act) must be newly interpreted to remedy discrimination that occurs *between users* on "sharing economy" platforms, considering these platforms' functional equivalence to the types of establishments to which those laws have traditionally applied (hotels, taxi services, and the like).[7]

Even if antidiscrimination law can be viably extended to platforms for users' discriminatory conduct, platforms often assert immunity based specifically on their status as platforms. Section 230 of the Communications Decency Act immunizes a provider of "an interactive computer service" from being treated as the publisher of information provided by its users[8]—

4. Other platforms operate in domains to which federal antidiscrimination law is less obviously applicable. On these platforms, where law may not provide a ready remedy for users' behaviors that might systematically disadvantage certain groups, design interventions may be even more important tools for the mitigation of bias.

5. *See generally* Orly Lobel, *The Law of the Platform*, 101 MINN. L. REV. 87 (2016). Some of this is due to platforms' poor fit with traditional models of employment—for instance, their tendency to take steps to ensure that service providers are not characterized as employees of the platform company. *See* Alex Rosenblat et al., *Discriminating Tastes: Uber's Customer Ratings as Vehicles for Workplace Discrimination*, 9 POL'Y & INTERNET 256, 266-67 (2017).

6. *See, e.g.,* Talia G. Loucks, *Travelers Beware: Tort Liability in the Sharing Economy*, 10 WASH. J.L. TECH. & ARTS 329, 335, 338 (2015).

7. *See* Aaron Belzer & Nancy Leong, *The New Public Accommodations*, 105 GEO. L. J. 1271, 1271 (2017). *See also* Michael Todisco, *Share and Share Alike? Considering Racial Discrimination in the Nascent Room-Sharing Economy*, 67 STAN. L. REV. ONLINE 121, 128–29 (2015); Katharine T. Bartlett & Mitu Gulati, *Discrimination by Customers*, 102 IOWA L. REV. 223, 249–50 (2017).

8. 47 U.S.C. § 230(c)(1).

including, often, information and conduct that exhibits bias.[9] CDA 230 has proved the most important tool upon which platforms currently rely to avoid liability for their users' conduct. Though the applicability of the statute has not yet been thoroughly tested with respect to the platform economy,[10] courts have sometimes seen fit not to apply immunity when platforms' actions "help[] to *develop* [users'] unlawful content"[11] through how such information is solicited or structured on the site. In the most prominent case in which a court so found, *Fair Housing Council of San Fernando Valley v. Roommates.com*,[12] the Ninth Circuit held that Roommates.com (a listing service for prospective roommates seeking housing, and vice versa) was not entitled to CDA 230 immunity because its site had featured dropdown menus through which users were required to provide information about their gender and sexual orientation, as well as their preferences about the corresponding characteristics desired in a roommate. In structuring users' interactions in such a way, the Court found that the platform had "[made] answering the discriminatory questions a condition of doing business."[13]

Roommates brought platform design to the fore as a potentially determinative factor in resolving whether platforms are immune from discrimination claims based on user conduct. The case has generated a significant amount of legal scholarship, focused primarily on issues related to how the case augurs for the future contours of the CDA's immunity protection.[14] But less attention has been paid to the first-order question of what sorts of design decisions platforms make that might mitigate or

9. *See, e.g.*, Chi. Lawyers' Comm. for Civil Rights Under Law, Inc. v. Craigslist, Inc., 519 F.3d 666 (7th Cir. 2008). In *Craigslist*, the Seventh Circuit ruled that Craigslist was entitled to CDA 230 immunity for user posts containing discriminatory housing limitations (such as "NO MINORITIES"). In so ruling, the court noted that Craigslist had merely "provid[ed] a place where people can post" housing ads, and that "[n]othing in the service [C]raigslist offers induces anyone to post any particular listing or express a preference for discrimination." *Id.* at 671.

10. Belzer & Leong, *supra* note 7, at 1320–21.

11. *Fair Housing Council of San Fernando Valley v. Roommates.com*, 521 F.3d 1157, 1168 (9th Cir. 2008) (emphasis added).

12. *Id.*

13. *Id.* at 1181. Similarly, the Court also found that the platform's search and filter system was not entitled to CDA immunity, as the platform "designed its search system so it would steer users based on the preferences and personal characteristics that Roommate itself forces subscribers to disclose. … [Roommates.com] designed its system to use allegedly unlawful criteria so as to limit the results of each search, and to force users to participate in its discriminatory process." *Id.* at 1167.

14. *See, e.g.,* Belzer & Leong, *supra* note 7; Varty Defterderian, *Fair Housing Council v. Roommates.com: A New Path for Section 230 Immunity*, 24 BERKELEY TECH. L.J. 563 (2009); Bradley M. Smyer, *Interactive Computer Service Liability for User-Generated Content After Roommates.Com,* 43 U. MICH. J. L. 811 (2010).

exacerbate the role of users' biases. This Article's focus, then, is not to explore the potential applicability of antidiscrimination statutes or CDA immunity to platforms; rather, it is to complement and expand those areas of active scholarly discussion with empirical exploration of platforms' design and policy choices.[15]

This Article provides a conceptual framework for understanding how platforms' design and policy choices introduce opportunities for users' biases to affect how they treat one another. We do so through empirical review of design-oriented interventions used by a range of platforms and the synthesis of this review into a taxonomy of thematic categories. In so doing, we hope to prompt greater reflection on the stakes of such decisions as they are made by platforms already, guide platforms' future decisions, and provide a basis for empirical work measuring the impacts of design decisions on discriminatory outcomes.

We proceed as follows. In Part I, we describe our empirical review of platforms, and the strategies we used to develop our taxonomy. In Part II, we present in detail the ten thematic categories that emerged from this review and describe how platforms' design interventions might mediate or exacerbate users' biased behaviors, drawing from social and psychological research on bias and stereotyping. Further, we discuss the interactions among design features, and the importance of acknowledging these interactions to effectively address bias. Part III describes the ethical dimensions of platforms' design choices—including when platforms might *not* want to attempt to mitigate users' biases—and concludes.

I. METHODS

Our analysis is based on a review of over fifty platforms spanning several domains. We specifically set out to identify platforms in the following seven areas, where online platforms have assumed an important role: consumer-to-consumer sales; transportation; tasks and gigs; hiring; housing; crowdfunding and lending; and dating. We identified the dominant platforms in each area. We also included widely recognized companies in

15. Notably, legal proceedings (or the specter thereof) have had the effect of requiring platform companies to address user-to-user bias through design alterations. As described *infra*, the voluntary agreement between Airbnb and the California Department of Fair Employment and Housing (in resolution of the lawsuit filed by the agency against Airbnb) mandates that Airbnb adopt, or consider adopting, a number of design-relevant strategies in order to reduce discrimination. Hence, platform design choices can intersect intimately with anti-discrimination law, both by moderating avenues to liability and by serving as a means to resolve litigation.

our survey, particularly those that have received attention in existing scholarship and media coverage on the problem of bias in users' interactions. Finally, we attempted to identify platforms that have made public statements committing to address bias.

We then explored the user experience on each of these platforms, systematically documenting design choices that created or limited opportunities for users to purposefully discriminate against one another or for users' implicit biases to influence their behaviors. Our initial efforts were informed by existing research on how design decisions can assist in successfully constructing and regulating online communities,[16] particularly dealing with troubling user behavior (e.g., trolling and harassment).[17] We examined how users experience these platforms when connecting via the web or apps; of the platforms we examined, the majority exist as mobile-only apps. In both cases, we created user accounts on each platform, as most do not allow non-users to access even basic features or functionality. We reviewed how a typical user might interact with the website or app. Our review paid particular attention to features that had been the focus of previous research, notably the contents of user profiles and product listings, all manner of sorting and rating mechanisms, the presence or absence of terms of service, community guidelines, or any explicit statements about bias or discrimination, as well as tools to report unsanctioned user behavior.

For ethical reasons, we limited our exploration to situations where we could observe or make use of a feature without directly interacting with other users. We reasoned that engaging users in transactions that we did not intend to complete, for example, would be dishonest and would waste the time of the other user. Sometimes, we were unable to document features only available to users during or after their interactions with others. We also refrained from taking any action that could potentially cause harm or negative outcomes for other users. For example, in many cases, platforms offer the opportunity to report or flag other users or content (as we describe in Part II.C.1, *infra*). To avoid reporting innocent users, we refrained from clicking on "Report" links unless it was obvious that the link would lead to a list of options (which we would not complete) and would not immediately report the user or content. In cases where some information about the platform was inaccessible, we made attempts to find relevant details by

16. *See, generally,* Robert E. Kraut et al., BUILDING SUCCESSFUL ONLINE COMMUNITIES: EVIDENCE-BASED SOCIAL DESIGN (2012). Kraut et al.'s framework of design strategies for online communities consists of eight design categories that can govern how users are allowed to behave in their interactions.

17. J. Nathan Matias et al., *Online Harassment Resource Guide*, WIKIMEDIA (July 3, 2015), https://meta.wikimedia.org/wiki/Research:Online_harassment_resource_guide.

other means. This often meant locating the platform's how-to guides, which walk new users through the process of interacting or transacting with others and often include images of the platform's interface at each stage in the process.

In reviewing the features of each platform, we found commonalities across various design and policy choices and grouped these into ten categories described below and summarized in the following table. These categories cluster into three general groups: setting platform- and community-wide policies, structuring users' encounters and experiences on the platform, and monitoring and evaluating platform activity to root out bias. Our taxonomy of design choices is not intended as an endorsement of any particular strategy for dealing with bias, nor as an empirical assessment of the efficacy of these interventions. Instead, this Article aims to provide the first comprehensive and coherent account of the many ways that platforms can—and often already—attempt to limit discrimination between users.

Table 1: Overview of platform policy and design strategies

Strategy		Examples
Setting policies	Company-level diversity and anti-bias strategies	Increasing diversity within the company workforce; educating employees about bias; engaging underrepresented groups in the design process
	Community composition	Restricting community through norms, rules, and structures
	Community policies and messaging	Community guidelines; required training on community norms; pledges; language and imagery on- and off-site
Structuring interactions	Prompting and priming	Prompting user to reflect on their behavior at specific decision points
	How users learn about one another	Matching users; searching; filtering
	What users learn about one another	Encouraging or requiring disclosure of user information; withholding user information; structuring the presentation of user information
	Reputation, reliability, ratings	Testimonials; references; reviews; badges; ratings
Monitoring and evaluating	Reporting and sanctioning	Creating mechanisms for user to report biased behavior; sanctioning users who discriminate
	Data quality and validation	Requiring more granular information; adjusting ratings; delisting reviews; requiring validation from external data
	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

II. HOW PLATFORMS MEDIATE BIAS IN USER-TO-USER INTERACTIONS

A. SETTING POLICIES

Platforms' corporate and community-wide policies may affect the degree to which discrimination occurs among users. A focus on diversity and inclusion in corporate teams and the design process can sensitize the platform to bias-related issues. Defining and limiting the community along specific lines can help to establish greater affinity among users, thereby reducing the likelihood that users will rely on biased heuristics when engaging with others. Finally, creating and communicating rules and norms for users' conduct can be a way to discourage and sanction biased behavior.

1. Company-level diversity and anti-bias strategies

To address issues of systemic bias on platforms, some companies have adopted strategies that aim to address the problem by reforming the company *itself*. Companies might seek to address platforms' role in mitigating bias *from within*, by altering their own organizational makeup, internal policies, and design processes. For example, platform companies may seek to increase the representation of underrepresented groups within their engineering teams through targeted hiring initiatives or strategies aimed at mitigating bias and discrimination in internal hiring processes. Companies may offer specialized training to their engineering teams and other members of their workforce about implicit bias and its effects, or firms may devote particular personnel and other internal resources to the project of bias elimination. Finally, firms may explicitly integrate engagement with underrepresented groups in their design processes.

Airbnb employed each of these approaches as part of the reforms associated with its 2016 nondiscrimination review, conducted in response to findings of systemically worse outcomes for black users seeking short-term housing on the site.[18] It sought to increase diversity in its workforce by implementing a "Diversity Rule" requiring that all candidate pools for senior positions include underrepresented minorities and women,[19] premised on the idea that the company "may have been slow to address concerns about discrimination because [its] employees are not sufficiently

18. Laura W. Murphy, Airbnb's Work to Fight Discrimination and Build Inclusion 22, 24 (Sep. 8, 2016), http://blog.atairbnb.com/wp-content/uploads/2016/09/REPORT_Airbnbs-Work-to-Fight-Discrimination-and-Build-Inclusion.pdf.

19. *See id*. at 12.

diverse."[20] In addition, Airbnb increased recruitment efforts for underrepresented groups, and included a diversity measure in the assessment of hiring managers.[21] It expanded and required anti-bias training for all its employees, with specialized training for customer service representatives who interact directly with hosts,[22] and created a new team of engineers, designers, data scientists, and researchers devoted to anti-bias projects.[23] Airbnb's review was developed in consultation with a range of end users, as well as representatives from a number of civil rights and advocacy organizations.[24]

Some strategies in this category assume a sort of "trickle-down" approach to bias mitigation. The presence of more diverse company personnel (perhaps especially engineers) and explicit training about the problem of implicit bias may attune companies to the potentially disparate effects of their design decisions that might not have come to the fore otherwise.[25] Internal strategies such as targeted hiring or company-wide

20. *Id.* at 17. This policy—commonly known as the "Rooney Rule," after Pittsburgh Steelers chairman Dan Rooney, who implemented a similar approach in the National Football League—has recently gained traction in Silicon Valley in response to calls for greater diversity in tech employment. Davey Alba, *The NFL is Showing Silicon Valley How to Be More Diverse*, WIRED (Oct. 26, 2015), https://www.wired.com/2015/10/tech-silicon-valley-nfl-rooney-rule-diversity/ (describing tech companies' increasing adoption of the Rooney Rule to increase diversity in management roles).

21. Murphy, *supra* note 18, at 24 (noting that Airbnb planned to increase the proportion of underrepresented minorities in its U.S. workforce from 9.64% in September 2016 to 11% by the end of 2017).
Id. at 22.

23. *Id.* at 24.

24. *Id.* at 15; *see* Jessi Hempel, *For Nextdoor, Eliminating Racism is No Quick Fix*, BACKCHANNEL (Feb. 16, 2017, 12:00 AM), https://backchannel.com/for-nextdoor-eliminating-racism-is-no-quick-fix-9305744f9c6 (describing Nextdoor's consultation with racial justice groups and government officials in their efforts to mitigate user bias on the platform, in which they "began holding regular working groups in which they included these people in the product development process"). These approaches relate to participatory design processes, through which designers consult with and integrate stakeholders throughout the design process in order to address needs and concerns of users that might not otherwise be apparent. *See* Christopher A. Le Dantec & Carl DiSalvo, *Infrastructuring and the Formation of Publics in Participatory Design*, 43 SOC. STUD. SCI. 241 (2013).

25. Similarly, measures to increase within-company diversity have been espoused in response to other instantiations of bias on platforms and in algorithms. *Cf.* Kate Crawford, *Artificial Intelligence's White Guy Problem*, N.Y. TIMES, Jun. 26, 2016, at SR11 (relating lack of corporate inclusivity to bias in artificial intelligence); Charlie Warzel, *"A Honeypot for Assholes": Inside Twitter's 10-Year Failure to Stop Harassment*, BUZZFEED (Aug. 11, 2016), https://www.buzzfeed.com/charliewarzel/a-honeypot-for-assholes-inside-twitters-10-year-failure-to-s (relating Twitter's leadership homogeneity and corporate culture to its difficulties in stemming abuse and harassment for users).

training may also appeal to, and dovetail with, other corporate goals about diversity and inclusion, particularly given recent efforts to address the vast underrepresentation of women and minorities in Silicon Valley and the exclusionary corporate cultures at tech firms.[26]

However, diversity and inclusion measures are unlikely to themselves be a panacea for addressing bias in hiring and corporate culture—not to mention for influencing platform design choices that may facilitate user-to-user bias. The efficacy of unconscious bias and diversity training has not been established empirically.[27] In some cases, such trainings may even *exacerbate* biased behavior by normalizing biased attitudes (e.g., by suggesting that "everyone is biased")[28] or by causing people to retaliate against feelings of pressure, social judgment, and control.[29]

2. Community composition

A second set of strategies involves creating barriers to membership in a platform-mediated community by implementing rules or expectations about the characteristics members must meet in order to participate on the platform. Such restrictions may be based on membership in a demographic category (e.g., JDate,[30] a dating website for Jewish users), geographic proximity (e.g., neighborhood sites on Nextdoor[31]), shared professions or interests (e.g., FarmersOnly,[32] for farmers, and VeggieDate,[33] for vegetarians), or other limitations.

---

26. *See, e.g.,* Liza Mundy, *Why is Silicon Valley So Awful to Women?* THE ATLANTIC (April 2017), https://www.theatlantic.com/magazine/archive/2017/04/why-is-silicon-valley-so-awful-to-women/517788/ (describing how "unconscious-bias training has emerged as a ubiquitous fix for Silicon Valley's diversity deficit").

27. *See generally* Elizabeth Levy Paluck & Donald P. Green, *Prejudice Reduction: What Works? A Review and Assessment of Research and Practice,* 60 ANN. REV. PSYCH. 339 (2009); Frank Dobbin & Alexandra Kalev, *Why Diversity Programs Fail*, HARV. BUS. REV., Jul. 2016 at 52.

28. *See* Michelle M. Duguid & Melissa C. Thomas-Hunt, *Condoning Stereotyping? How Awareness of Stereotyping Prevalence Impacts Expression of Stereotypes*, 100 J. APPLIED PSYCHOL. 343, 354 (2015) (suggesting that increasing awareness of stereotyping can normalize prevalent stereotypes). *See also* Jessica Nordell, *Is This How Discrimination Ends?*, THE ATLANTIC (May 7, 2017), https://www.theatlantic.com/science/archive/2017/05/unconscious-bias-training/525405/ (detailing strengths and weaknesses of the widely used Implicit Association Test ("IAT") and various anti-prejudice interventions).

29. Lisa Legault et al., *Ironic Effects of Antiprejudice Messages: How Motivational Interventions Can Reduce (but Also Increase) Prejudice,* 22 PSYCHOL. SCI. 1472 (2011).

30. JDATE, https://www.jdate.com (last visited July 15, 2017).

31. NEXTDOOR, https://nextdoor.com (last visited July 15, 2017).

32. FARMERSONLY, https://www.farmersonly.com (last visited July 15, 2017).

33. VEGGIEDATE, http://www.veggiedate.org (last visited July 15, 2017).

Community composition can be limited by a platform in a number of ways. The simplest mechanism is for the platform to establish that the community is *for* some users—and, concomitantly, *not* others—through messaging. JDate, for example, does not police whether its members are in fact Jewish, but indicates that the site is intended for Jewish users through its site text, logo, and other elements.[34] The operative assumption seems to be that this norm and purpose will be enforced through self-selection into the community of users.

Other sites enforce community composition norms more structurally—for instance, by requiring users to possess a credential, like an email address from a prescribed domain or set of domains. In its early days, Facebook was restricted only to users with .edu email addresses from select colleges and universities.[35] Other sites' user bases are restricted based on network proximity (for instance, new users must be within *x* degrees of existing users in an articulated social network, like LinkedIn or Facebook), or allow new users in when invited by existing members. Nextdoor, for example, requires prospective users to verify their home addresses before gaining access to the site. It offers several options for doing this, including providing social security or credit card information linked to the address. Alternatively, prospective users can skip the verification process by receiving invitation codes from already-verified neighbors.[36] The civic engagement platform PlaceSpeak (which provides a venue for residents of an area to comment on local issues) allows users to authenticate themselves via several methods, including "linking to social media profiles, verifying IP addresses, and confirming identities over the phone"[37]; further, the more heavily authenticated a user is, the more her input is weighted in the discussion.[38]

34. JDate addresses the possibility that non-Jewish users might sign up for the site. *See* About JDate, Who Uses JDate?, https://www.jdate.com/help/about/(noting "JDate is designed for Jewish singles of all ages looking to connect based on common ground. That being said, JDate is open to all singles 18 years or older. If you're not Jewish and you're still interested in joining JDate, please ensure that the religion section of your profile indicates whether or not you are willing to convert").

35. Sarah Phillips, *A Brief History of Facebook*, THE GUARDIAN (Jul. 25, 2007), https://www.theguardian.com/technology/2007/jul/25/media.newmedia.

36. Getting Started, NEXTDOOR, (Jun. 20, 2017), https://help.nextdoor.com/customer/en/portal/articles/805357-verify-your-address.

37. Zack Quaintance, *New Resident-Facing Platform Seeks Public Input, Minus the Trolls*, GOVTECH (Jun. 30, 2017), http://www.govtech.com/civic/New-Resident-Facing-Platform-Seeks-Public-Input-Minus-Trolls.html.

38. *Id.* In this way, user authentication is treated as a proxy for data quality. *See infra* Part II.C.2.

Finally, platforms may police membership through independent vetting. Uber screens potential drivers by requiring a background check, which it contracts out to a third-party company; that company screens the potential driver for criminal history, suspected terrorist activity, presence on the National Sex Offender Registry, and other information.[39] Airbnb and Uber are reportedly considering using Aadhaar, India's controversial biometric identity database, to validate users' identities.[40] However, logistical concerns—such as the expense of background checks, the time required to complete them, and diverse regulatory environments—may limit the use of such tools on other platforms.[41]

Restricting access to a community—through norms, rules, or structures—may, of course, be a relatively overt way for a platform to *itself* propagate bias through explicit exclusion of particular groups from participation (and, concomitantly, from the social and economic opportunities that such participation might afford). But, in addition, measures that restrict community composition may exacerbate or mitigate bias in interactions *between users* on platforms.

In some cases, restricting access may help to cultivate a baseline level of affinity, homogeneity, connection, or trust among users that may militate against bias in their interactions. In particular, users may refrain from reliance on stereotypes about *secondary* characteristics because negative associations based on those characteristics are neutralized by membership on the platform—and the associated characteristics it imparts. Users who perceive themselves to be similar along some dimension, who are prone to repeated interaction (based, for instance, on geographic proximity), or who are inclined to trust one another because of external vetting or credentialing may be less likely to fall back on biases about other characteristics[42]: for

39. Sarah Kessler, *The Truth About Uber's Background Checks*, FAST COMPANY (Aug. 26, 2015), https://www.fastcompany.com/3050172/the-truth-about-ubers-background-checks. Uber has been criticized for the thoroughness of its background checks; *see* Adrienne LaFrance & Rose Eveleth, *Are Taxis Safer than Uber?*, THE ATLANTIC (Mar. 3, 2015), https://www.theatlantic.com/technology/archive/2015/03/are-taxis-safer-than-uber/386207/.

40. Pranav Dixit, *Airbnb, Uber, and Ola Are Considering Using India's Creepy National ID Database*, BUZZFEED (Jul. 19, 2017), https://www.buzzfeed.com/amphtml/pranavdixit/airbnb-uber-and-ola-may-start-using-aadhaar-indias.

41. Kessler, *supra* note 39.

42. Social psychology research demonstrates that in conditions of high ambiguity, people are more likely to fall back on stereotypes as heuristics that guide their behavior. *See* Samuel L. Gaertner & John F. Dovidio, *Understanding and Addressing Contemporary Racism: From Aversive Racism to the Common Ingroup Identity Model*, 61 J. SOC. ISS. 615, 621 (2005).

instance, someone with strong negative biases against members of a particular ethnicity may find those stereotypes neutralized by cues about that person's membership in a common group—even if those cues are themselves stereotypic in nature.[43] Behaviors and traits that "attenuate perceptions of threat" from a stereotyped group are sometimes called "disarming mechanisms"[44]; membership in a particular group may suggest the presence of such counter-stereotypical traits that may ultimately mitigate the effects of a primary bias. In addition, if restrictive community composition creates the impression that platform users are members of a cohesive "in-group," such identification might lend users a sense of common identity, overriding biases among group members based on other factors.[45]

At the same time, understanding a platform to be restrictive in some way could have the opposite effect on users by *encouraging* biased judgments, potentially by normalizing the idea that the platform is itself "exclusive" or elite, subduing norms of nondiscrimination in members' interactions. Users may thus feel emboldened to rely even more upon stereotypes in their interactions on a platform.

In practice, the effects of community composition interventions seem likely to depend on users' recognition of these constraints (e.g., are members aware of the barriers to entry into the user base?) and the social associations primed by such exclusions and inclusions (e.g., membership in a shared university community may impart strong positive associations). To the extent that such interventions operate to create a group identity, they may facilitate positive interactions among members of the group.[46]

---

43. *See* David S. Pedulla, *The Positive Consequences of Negative Stereotypes: Race, Sexual Orientation, and the Job Application Process*, 77 SOC. PSYCHOL. Q. 75 (2014) (discussing how counter-stereotypical information can alleviate discrimination by providing a counterweight to stereotypes associated with a particular group). Interestingly, counter-stereotypical cues need not necessarily be positive in nature to alleviate negative stereotypes. Pedulla demonstrates empirically that stereotypes about gay men may negatively impact their perceived employability when the men are assumed to be white—but may actually have *positive* consequences for gay black men's perceived employability. Pedulla posits that the stereotype of gay men as weak counteracts the stereotype of black men as threatening, resulting in a net benefit in perceived employability, rather than a "double disadvantage," for this marginalized group.

44. Robert Livingston & Nicholas Pearce, *The Teddy Bear Effect: Does Babyfaceness Benefit Black CEOs?*, 20 PSYCHOL. SCI. 1229, 1229 (2009).

45. In social psychology, the *common in-group identity model* proposes that the creation of such "superordinate" group identifications can be an effective means of reducing inter-group biases. *See* Samuel L. Gaertner et al., *The Common Ingroup Identity Model: Recategorization and the Reduction of Intergroup Bias*, EUR. REV. SOC. PSYCHOL. 4(1): 1–26 (1993).

46. Gaertner & Dovidio, *supra* note 42, at 629–30.

3. Community policies and messaging

Platforms send a variety of messages to their users about what types of conduct are permissible and use multiple methods to communicate these expectations. These tools range from community guidelines and terms of use (which may be styled as voluntary agreements or mandatory commitments), to required trainings on expected norms of conduct, to other forms of messaging and imagery both on and off the platform. Governance of user behavior through these strategies ranges from explicit rules backed by sanctions to persuasive communication of platform-espoused norms.

Community guidelines and terms of use may explicitly enjoin users from biased interactions on platforms. Some platforms style such policies as "commitments" and incentivize or require users to engage directly with them. For instance, Airbnb's nondiscrimination review requires all users to "affirm and uphold the Airbnb Community Commitment" before using the platform—which commits users "to treat all fellow members of this community, regardless of race, religion, national origin, disability, sex, gender identity, sexual orientation or age, with respect, and without judgment or bias"[47]—as well as to accede to a more detailed nondiscrimination policy, which specifies actions that Airbnb hosts may and may not take.[48] That policy notes that hosts who violate it can be suspended from platform use.[49] Airbnb also provides unconscious bias training for its hosts and asserts it will "work to highlight" hosts who undergo it.[50]

Platforms may also style community guidelines as "pledges," which operate both to cultivate desired norms for interaction *and* to serve a signaling function among the user base. For example, Daddyhunt—a location-based dating app for sexual minority men[51]—encourages its members to "live stigma-free" with respect to dating others regardless of HIV status. A user is not required to pledge to live stigma-free in order to use the platform; however, if he decides to do so, an indicator is attached to his profile (see Figure 1).[52] The founder of Daddyhunt suggests that such a

47. Murphy, *supra* note 18, at 10.

48. For example, hosts may not prohibit the use of a guest's mobility devices but *may* require guests to obey restrictions related to keeping a Kosher kitchen. *Id.* at 27–32.

49. *Id.* at 32.

50. Murphy, *supra* note 18, at 22; *see also supra* Part II.A.1 (critiquing implicit bias training with respect to employees).

51. DADDYHUNT, http://www.daddyhunt.com (last visited July 15, 2017).

52. *See infra* Part II.B.4 (noting the use of profile indicators, like badges, more generally).

feature is designed to "foster a nicer and less judgmental environment for men to meet men."[53]

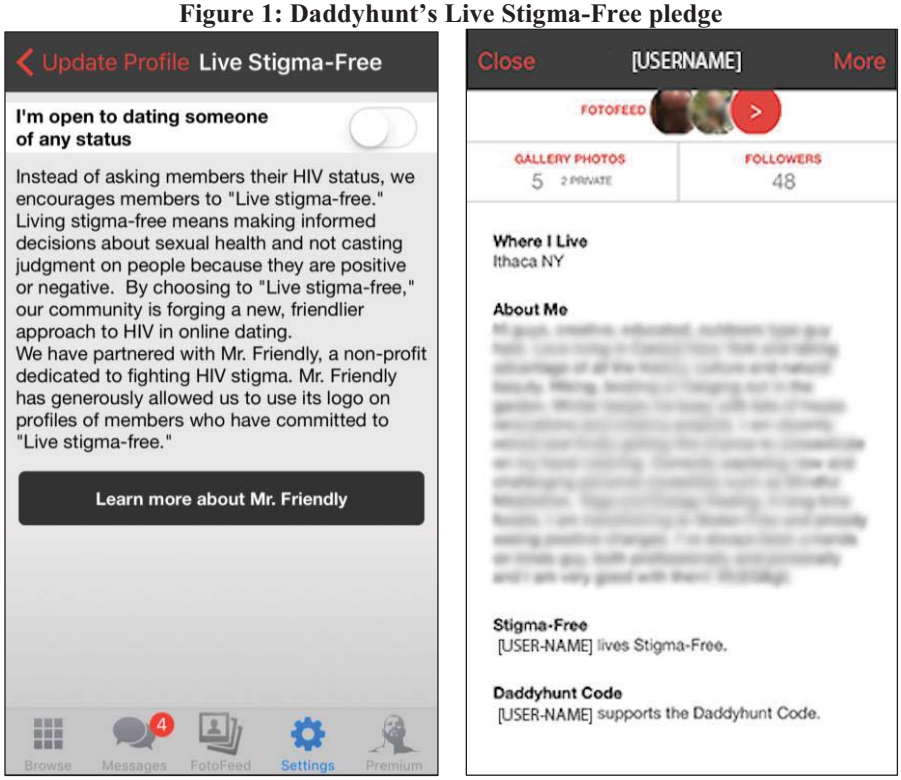
The presence of such a feature ostensibly serves two functions. First, it gives users a tool to learn more about other users' attitudes and behaviors from their profiles, potentially reducing experiences of discrimination and unwelcoming interactions. In addition—even without being required for all users—the presence of the feature serves a broader norm cultivation function by normalizing social interaction with HIV-positive users. Moreover, it represents an interesting alternative means of communicating information about HIV status on Daddyhunt. The platform does not explicitly ask users to disclose HIV status as an element of the user profile,[54] perhaps in light of concerns about increasing stigma by asking such a question directly.[55]

53. Interview with Carl Sandler, CEO of MINSTER, DIGITAL CULTURE & EDUC. (Jul. 17, 2014), http://www.digitalcultureandeducation.com/uncategorized/sandler_html/ (referring to an analogous feature on a related app, MISTER).

54. *See infra* Part II.B.3 on the information revealed in user profiles.

55. In a related vein, the site DUDESNUDE (https://dudesnude.com) opted to create an HIV-friendly community called Poz (short for "positive") rather than asking users about HIV status directly, noting that "not answering the question [about HIV status] may be interpreted as dodging the issue, or putting people in the position where they have to lie." SAN FRANCISCO AIDS FOUNDATION, BUILDING HEALTHY ONLINE COMMUNITIES MEETING REPORT 2 (2014).

Figure 1: Daddyhunt's Live Stigma-Free pledge



Community policies can also be communicated less directly through messaging and imagery, both on and off the platform. For instance, a platform's public statements and advertising can communicate desired norms and evince a desire to attract a user base that shares those norms. Airbnb aired a television advertisement during Super Bowl LI that displayed a diverse range of faces and opined that "[w]e believe no matter who you are, where you're from, who you love or who you worship, we all belong."[56]

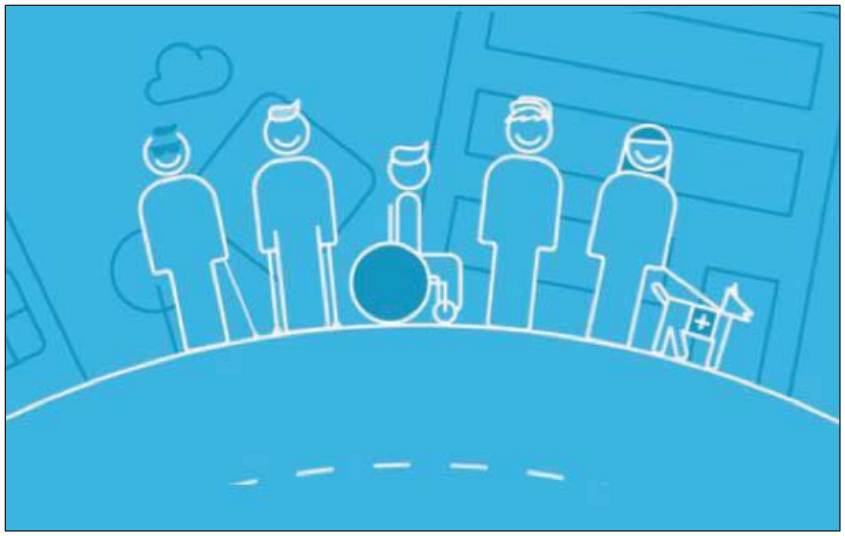
Language and imagery on a site or app may function similarly by affecting how users perceive the norms of transaction. For example, a task site that uses gendered language or images (say, of men assembling furniture) might cue users to view certain categories of people as more appropriate candidates for particular jobs. TaskRabbit, for instance, advertises itself as a tool to "[b]ook a top-rated handyman,"[57] though there

56. Katie Benner, *In Airbnb's Super Bowl Ad, Implied Criticism of Trump's Travel Ban*, N.Y. TIMES (Feb. 6, 2017), at B3.

57. TaskRabbit, http://www.taskrabbit.com (last visited July 15, 2017).

are a number of handywomen working through the platform.[58] Lyft's website features a video on its anti-discrimination policies,[59] which contains imagery of a variety of users with disabilities (including a user in a wheelchair and another with a service dog), providing suggestive visual messaging of the range of users to be welcomed on the platform. The short-term rental site Innclusive[60] describes itself as a "new platform where people of all backgrounds can travel and stay with respect, dignity, and love"[61] and specifically describes the founder's experiences of discrimination on Airbnb as the site's motivation. Even outside of the explicit context of discrimination, visual and textual messaging on a site can cultivate community norms. For instance, Grindr's app includes a "Sexual Health FAQ" that is presented in the context of setting up a user profile. The resources linked to through the FAQ include information on safe sex with HIV-positive partners, potentially reducing such users' stigma on the platform.

Figure 2: Image from Lyft's anti-discrimination policies webpage



58. Elana Lyn Gross, *What It's Like to Be A Female Handywoman on TaskRabbit*, FORBES (Apr. 3, 2017), https://www.forbes.com/sites/elanagross/2017/04/03/what-its-like-to-be-a-female-handywoman-on-taskrabbit/#1828f19115f5.

59. Anti-Discrimination Policies, LYFT, https://help.lyft.com/hc/en-us/articles/214218517-Anti-Discrimination-Policies.

60. INNCLUSIVE, https://www.innclusive.com (last visited July 15, 2017).

61. Our Story, INNCLUSIVE, https://www.innclusive.com/our-story (last visited Mar. 13, 2018).

Community guidelines and messaging can help to address bias in many different ways. Most obviously, published guidelines can offer platforms a consistent basis upon which to assess user behavior and punish those who violate the rules. Should a platform explicitly forbid discriminatory conduct, it could sanction users who engage in such behavior. Efforts to inform users of these guidelines, however, can also deter users from engaging in prohibited behavior in the first place. In such cases, users would consciously self-regulate, abstaining from intentional discrimination for fear of punishment. In contrast, platforms might rely on messaging to change users' underlying beliefs about inclusion, diversity, and non-discrimination. In publicly committing the community to such values, platforms aim to foster such commitments among individual users, encouraging users to shed their prejudicial or biased beliefs. At the same time, messaging of this sort can also invite potential users who hold these beliefs to join the community, thereby changing the composition of the community to include more users consciously committed to keeping their prejudices and biases at bay. Platforms might be more circuitous, however, engendering greater comfort among diverse users by simply exposing users to images and messaging that undermine stereotypes and stigmas. Rather than targeting consciously held beliefs, these interventions address implicit associations, prompting changes in attitudes about which users may remain largely unaware.

B. STRUCTURING INTERACTIONS

Platforms have immense power to scaffold users' encounters with one another. They can offer cues to users about conduct norms in the moment of exchange, just as biases are likely to surface. They can structure markets to determine the degree of choice and discretion users have about their exchange partners, and control what users learn about each other's characteristics throughout an encounter—potentially restricting a user's propensity or ability to discriminate. They can also offer users opportunities to evaluate each other's performance (though these evaluations may themselves be marked by bias) and can make indicia of reputation visible to other potential exchange partners.

1. Prompting and priming

Platforms may prompt users to reflect on their behavior with the intent to minimize bias, often by providing users with some information or pop-up dialog (known in web design as an *interstitial*) that must be acceded to before proceeding with a particular communication or transaction. These measures can take several forms and may be more or less explicit about the reason for the intervention. They may remind users to abide by community

guidelines or a code of conduct[62] to which they have previously agreed, hence priming users to refrain from exhibiting bias in the present transaction by making anti-bias commitments more top-of-mind.[63] Such prompts may be triggered to appear only when likely biased behavior is detected, by default for all interactions, or when some other conditions are met.

For example, the neighborhood-based social networking site Nextdoor developed numerous design strategies to alleviate racial profiling by users reporting suspicious activity in their neighborhoods.[64] Among these was the use of an interstitial prompt (see Figure 3, below) that gives users a variety of tips intended to minimize the role of bias in the report (e.g., "Give a full description, including clothing, to distinguish between similar people. Consider unintended consequences if the description is so vague that an innocent person could be targeted.").[65] Similar approaches have been used in addressing online harassment and incivility. The discussion platform Discourse prompts new users with reminders of community civility guidelines just as they begin to post content to the site.[66] Yik Yak, a social media app in which users post anonymous "Yaks"—short messages visible to others in their local area, often college campuses—prompts users to "pump the brakes" via an interstitial message if the Yak in question seems to contain threatening or offensive language.[67]

Evidence from academic studies lends credence to these types of interventions. Mazar et al. show that people's propensity to lie is affected not by whether they know or believe that dishonest behavior is morally wrong, but "whether they think of these [moral] standards and compare their behavior with them *in the moment of temptation*."[68] In their study, priming research participants with reminders of honor codes and religious edicts just before completing a task deterred cheating behavior. More recent work has shown that immediately censuring people for violations of some norm can also reduce the incidence of such behavior. An experiment on Twitter deployed bots to respond to the use of racial slurs by socially sanctioning

62. *See* Part II.A.3, *supra*.

63. Fisman & Luca, *supra* note 1.

64. *See* Hempel, *supra* note 24. Nextdoor's interventions are discussed in more detail in Part II.C.2, *infra*.

65. Hempel, *supra* note 24.

66. *See* Jeff Atwood, *The "Just in Time" Theory of User Behavior*, CODING HORROR (Jul. 17, 2014), https://blog.codinghorror.com/the-just-in-time-theory/.

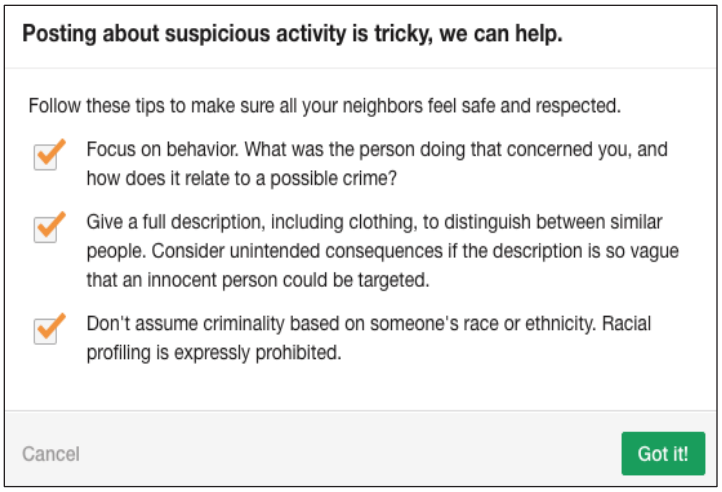
67. Jonathan Mahler, *Who Spewed That Abuse? Anonymous Yik Yak App Isn't Telling*, N.Y. TIMES (Mar. 8, 2015), https://www.nytimes.com/2015/03/09/technology/popular-yik-yak-app-confers-anonymity-and-delivers-abuse.html.

68. Nina Mazar et al., *The Dishonesty of Honest People: A Theory of Self-Concept Maintenance*, 45 J. MKTG. RES. 633, 635 (emphasis added).

users in an @-reply to the offensive message (e.g. "Hey man, just remember that there are real people who are hurt when you harass them with that kind of language."). The experiment found that such sanctions could reduce future propensity to harass, particularly when the bots were perceived to be white males with a large number of Twitter followers.[69]

There remains, however, ongoing debate in social psychology regarding the degree to which interventions aimed at making decisions more deliberative can counter biased judgment.[70] So far, evidence is mixed about the effectiveness of "getting people to think more about, or to attend more closely to, their objectives in an inter-racial interaction" as a means of mitigating the influence of *implicit* bias.[71]

Figure 3: Nextdoor's interstitial prompt, intended to minimize bias



69. *See* Kevin Munger, *Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment*, POL. BEHAV., (Nov. 11, 2016), https://doi.org/10.1007/S11109-16-9373-5.

70. *See* Jennifer L. Eberhardt, *Imaging Race*, 60 AM. PSYCHOL. 181, 181–2 (2005) ("Determining the extent to which racial bias can be automatically triggered versus deliberately controlled is a fundamental issue in social psychology. Better understanding this tension may improve not only theories of social cognition but also interventions designed to reduce bias and minimize racial inequities").

71. Anthony G. Greenwald & Linda Hamilton Krieger, *Implicit Bias: Scientific Foundations*, 94 CAL. L. REV. 945, 962 (2006) (reviewing studies).

2. How users learn about one another

By necessity, platforms scaffold the process by which users find one another, and different design choices can leave more or less leeway for users to exercise biased preferences. Platforms often function like markets, helping supply find demand (and vice versa). Platforms employ a wide range of techniques to facilitate this process. Some assume a relatively passive role in matching supply and demand (leaving a good deal of the process of finding an appropriate partner to the transacting parties themselves), while others take a more active role (automating much of the process of pairing riders and drivers, for example). Some provide users with tools to search and sort; others provide users with recommendations. In many cases, the marketplace that users confront on these platforms is a rank-ordered list; platforms cannot help but play a crucial part in determining what potential exchange partners are listed, and in what order.[72] Broadly speaking, then, platforms structure how users find one another by determining who is in a position to search for exchange partners, how much discretion users have in determining with whom to transact, and what tools are provided to users to facilitate the search process.

When deciding whether to put specific users in the privileged position of choosing with whom to interact, platforms create very different opportunities for discrimination to occur between users. Structurally, platforms can allow buyers to choose among sellers, allow sellers to choose among buyers, or allow both sellers and buyers to choose among themselves. A platform that gives job-seekers the opportunity to find and contact potential employers necessarily ensures that employers cannot limit their search to male candidates *ex ante* (despite the fact that employers may subsequently reject all female applicants) because employers are not the parties doing the searching. In contrast, when employers can search for candidates, they might only consider and contact male candidates, denying female candidates an opportunity to even learn about the job. A requirement to accept all comers in this case could mitigate against the possibility that candidates might reject offers for prejudicial reasons of their own, but such a requirement would also be unworkable in many cases (e.g., compelled work among freelancers).

Platforms can also be designed to provide sellers or buyers with a right of refusal when approached by the other—or deny any such discretion.

72. *See* James H. Moor, *What is Computer Ethics?*, 16 METAPHILOSOPHY 266, 274 (1985) (noting that designers of a search engine cannot avoid making choices about how results are ordered and displayed).

eBay, for example, allows buyers to choose among any sellers of the same or similar item, while sellers must accept all comers.[73] Fiverr, a platform for freelancers, has job-seekers (sellers) list their skills so that employers (buyers) can approach them, though job-seekers retain the right to refuse any particular offer. In contrast, Upwork, another platform for freelancers, features job listings where job-seekers (sellers) choose among job-listers (buyers), where the ultimate hiring decision still rests with the job-lister. Airbnb generally follows this format as well (giving hosts the right to refuse guests' housing requests), though the platform also provides an "Instant Book" feature that allows travelers to book a stay at someone's home without the host having a chance to review the request. The company has recognized that forcing buyers to accept all comers can mitigate against the possibility of bias in assessing users who make contact.[74] In other contexts, both transacting parties have the power to refuse offers as they please. On online dating platforms like OKCupid, the distinction between buyers and sellers breaks down and all users are in a position to find, approach, and accept or reject one another. In these cases, the decisions made by either party—either in who to contact or who to accept—could be biased.

At the same time, platforms frequently attempt to relieve buyers and sellers of much of the burden of finding an appropriate counterparty. At the extreme, some platforms automate the process of matching supply and demand, often using algorithms that pair people according to fixed criteria or patterns learned from historical data. Uber, for instance, does not present riders with a list of nearby drivers from whom riders may choose. Instead, Uber shields from view the process by which the company secures a driver for the specific rider and simply delivers a car to the passenger.[75] Uber does,

73. eBay sellers can, however, cancel bids from or sales to specific buyers, once they have learned about buyers. While the platform lists a small number of reasons for why sellers might want to do this (e.g., "A bidder contacts you to back out of the bid; You cannot verify the identity of the bidder after trying all reasonable means of contact; You end your listing early."), sellers seem to be at liberty to do this whenever they like. Sellers can also block specific buyers from even bidding on items, but sellers need to add specific usernames to a blacklist one-by-one. Managing Bidders and Buyers, eBay Help, http://pages.ebay.com/help/sell/manage_bidders_ov.html#block (last visited September 6, 2017).

74. *See* Murphy, *supra* note 18, at 22 ("Instant Book makes it easier for guests to be accepted by hosts on the platform if they meet some basic qualifications, and hosts can set preferences that serve the purpose of automatically filtering guests, including whether the listing is pet-friendly, suitable for events, or features particular amenities. More importantly, Instant Book reduces the potential for bias because hosts automatically accept guests who meet these objective custom settings they have put in place").

75. While Uber presents users requesting a ride with a map depicting cars in the area, such maps may not represent actual drivers available for pickups. Further, users have no

however, send notifications to nearby drivers who have the option to accept or reject the request. At no time do drivers see *all* requests in the area; Uber instead doles out requests one by one. From the perspective of riders, Uber fully automates the matching of supply and demand; from the perspective of drivers, Uber severely constrains information about nearby demand (to one request at a time), likely to pressure drivers to accept *any* request, given uncertainty about future requests. Notably, the company also withholds information about the intended destination of the passenger requesting a ride, also to limit the flow of information that might dissuade a driver from accepting the request.[76] And to top it off, Uber will penalize drivers who reject too many requests.[77]

Where this strategy is successful, Uber can come close to pairing riders and drivers in an almost fully automated manner. And doing so has been perceived as helping to mitigate bias.[78] Riders and drivers will have little opportunity to make biased assessments of each other because they never engage in any kind of negotiation or interaction; they are simply paired with one another based on some—ideally—rational criteria set by the platform. At the same time, such behavior betrays the belief that platforms operate as free and open markets.[79]

Fully automated matching will not work for all types of platforms. Users may not know exactly what they want in a counterparty or cannot express all their preferences and requirements explicitly or in advance. The process of exploring what is available on a platform may be the process by which users figure out what they want. Platforms facilitate this process by providing recommendations, rank ordering options, offering search functionality, and furnishing users with granular controls to sort and filter results in any number of ways. How platforms present users to each other can dramatically affect whether patterns of interaction exhibit bias. To

---

way of choosing among the cars depicted on the map. Alex Rosenblat, *Uber's Phantom Cabs*, MOTHERBOARD (Jul. 27, 2015), https://motherboard.vice.com/en_us/article/ubers-phantom-cabs.

76. Drivers and riders might get around this by cancelling on one another after learning about their assigned rider or driver. *See infra* note 158.

77. *See* Alex Rosenblat & Luke Stark, *Algorithmic Labor and Information Asymmetries: A Case Study of Uber's Drivers*, 10 INT'L J. COMM. 3758, 3762 (2016).

78. *See* Ruth Igielnik & Monica Anderson, *Ride-Hailing Services are Seen as a Benefit to Areas Underserved by Taxis*, PEW RES. CTR. (Jul. 25, 2016), http://www.pewresearch.org/fact-tank/2016/07/25/ride-hailing-services-are-seen-by-minorities-as-a-benefit-to-areas-underserved-by-taxis/.

79. *See* Tim Hwang & Madeleine Clare Elish, *The Mirage of the Marketplace*, SLATE (Jul. 27, 2015), http://www.slate.com/articles/technology/future_tense/2015/07/uber_s_algorithm_and_the_mirage_of_the_marketplace.html.

begin, if the platform relies on historical patterns of successful user interaction to guide its future recommendations, these suggestions might reproduce or even exacerbate the prejudices and biases that influenced previous users' decisions to interact with others—and their assessments of those people.[80] That said, recommendations can also direct users to focus on others who are more appropriate counterparties than users' biased heuristics might lead them to believe, or than traditional searching methods might uncover. Indeed, such recommendations might even incorporate diversity metrics,[81] which would aim to ensure that recommendations span a range of appropriate users across numerous identified groups. Platforms that aim to generate rank ordered lists that feature descending "best matches" might serve a similar function if the platform computes the rank by looking at demonstrably relevant factors as well as diversity metrics.[82]

Users might also have the ability to perform different types of searches, seeking out a counterparty with specific and discrete qualities (selected using dropdown menus or radio buttons, for example) or by entering freeform text into a dialogue box. Platforms might provide further control to users in the form of sorting mechanisms or filters, allowing users to change the rank order of people in the search results according to certain criteria or remove certain people completely. As a condition of providing these kinds of granular controls, platforms must determine the discrete criteria upon which they will allow users to sort and filter. Tools that grant users greater control over the set of people they will see when searching for a counterparty can both empower and embolden users to discriminate (and may expose platforms to liability for violation of underlying civil rights laws). While online dating platforms are not subject to any such laws, they vary dramatically in whether they allow users to search, sort, or filter by race, ethnicity, or religion, for example.[83] On Match.com, one of the first

80. *See* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CAL. L. REV. 671, 681–84 (2016). *See also* Aniko Hannak et al., *Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr*, Proc. Conf. on Computer-Supported Cooperative Work and Social Computing (2017).

81. *See, e.g.,* Saúl Vargas & Pablo Castells, *Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems*, PROC. FIFTH ACM CONF. RECOMMENDER SYS. 109 (2011).

82. In most cases, platforms that present a range of options to users cannot avoid presenting a list ordered in some particular way. *See* James H. Moor, *What is Computer Ethics?*, 16 METAPHILOSOPHY 266, 274 (1985).

83. *See, e.g.,* Carrie Weisman, *The Casual Racism of our Most Popular Dating Apps*, SALON (Sep. 28, 2015), https://www.salon.com/2015/09/28/sexual_racism_why_people_say_racist_things_on_dating_apps_partner; Patrick Strudwick, *The Founders of This Gay Dating App Won't Stop You Searching By Race*, BUZZFEED (Feb. 9, 2016),

questions posed to new users concerns their ethnic preferences in a partner; later on, users can filter search results on the basis of ethnicity. Yet other sites do not solicit these preferences or allow users to filter accordingly, despite the fact that users' preferences might still guide their ultimate dating choices.[84] And while platforms might not want to limit users' freedom of choice when it comes to romantic and sexual decisions, they may nevertheless refrain from collecting information and providing tools that allow users to effectively remove members of entire racial or ethnic groups from the apparent marketplace of potential partners.

3. What users learn about one another

In designing the interfaces through which users interact, platforms exercise enormous control over the type of information made available to transacting parties. What users see when hiring a worker to complete a task, getting in touch with the owner of a rental property, or contacting a prospective date, for example, can significantly affect how users judge these counterparties. In offline interactions, people cannot help but draw all sorts of inferences about others from readily available indicators. For example, a job applicant may (perhaps unwittingly) signal his personality through the clothing he wears or the way he carries himself. In the mediated interactions facilitated by platforms, such details might not be communicated to users at all—or supplemented by others. Indeed, platforms can make all sorts of choices about what information is disclosed, what is withheld, and how trustworthy that information is deemed to be.[85] For the purposes of online interactions, users are the sum total of the signals that platforms transmit between parties. Platforms, therefore, have powerful capacities to determine what their users learn about one another. This section details four ways in which they do so: by encouraging or requiring the disclosure of user information; by withholding user information; by structuring the input of

https://www.buzzfeed.com/patrickstrudwick/this-is-how-gay-dating-app-bosses-defend-racial-filtering. *See also* Elizabeth F. Emens, *Intimate Discrimination: The State's Role in the Accidents of Sex and Love*, 122 HARV. L. REV. 1307, 1322–23 (2009) (describing a range of dating websites' practices allowing, or requiring, indication of a user's own race and capacity to search for potential partners by race).

84. *See* Russell K. Robinson, *Structural Dimensions of Romantic Preferences*, 76 FORDHAM L. REV. 2787, 2792 (2007).

85. *See* Nicole B. Ellison et al., *Profile as Promise: A Framework for Conceptualizing Veracity in Online Dating Self-Presentations*, 14 NEW MEDIA & SOC. 45, 54 (2011) (describing online platforms as "reduced-cue environments" in which "online daters cannot 'show' characteristics such as age, gender, or location, [and so] are forced to 'tell' them through text-based communication").

user information; and by linking user information to external sources for authentication.

First, some platforms encourage or require[86] users to disclose personal information about themselves that fills them out as people, even if such information is not directly germane to the substance of the transaction. For instance, Airbnb allows users to submit 30-second profile videos and suggests that they include "a fun fact about yourself or why you love Airbnb"; eBay implores users to share "what you're passionate about."[87] TaskRabbit's "about me" section encourages taskers to enter information about hobbies and interests, in response to prompts like "When I'm *not* tasking . . . ."[88] In addition to personal profiles, many platforms encourage or require users to include profile photos or videos of themselves. Airbnb tells users that "[c]lear frontal face photos are an important way for hosts and guests to learn about each other. It's not much fun to host a landscape! Please upload a photo that clearly shows your face."[89]

The disclosure of additional information about a user may serve to mitigate bias. Information that leads others to see a user as a "whole person" might lead them to rely less on discrete signals (gender, ethnicity, etc.) in choosing partners with whom to transact. A recent study of Airbnb host profiles found that the majority of hosts disclose information about their career or education (e.g., the host's current job and where she went to school) and their interests and tastes (e.g., favorite books, music, and hobbies).[90] Moreover, the study found that hosts with longer profiles and who discuss more topics in their profiles are perceived as more trustworthy, and that such perceived trustworthiness can influence guests' choices in deciding with whom to stay.[91] Profiles may provide opportunities to signal

86. Even if platforms do not require users to disclose particular types of information, users who decline to include such information may be subject to adverse inference. *See* Scott R. Peppet, *Unraveling Privacy: The Personal Prospectus and the Threat of a Full-Disclosure Future*, 105 NW. U.L. REV. 1153, 1176–77 (explaining that users who refrain from disclosing information will be assumed to possess undesirable qualities) (2011).

87. *How Do I Make a Profile Video?*, AIRBNB, https://www.airbnb.com/help/article/213/how-do-i-make-a-profile-video (last visited Mar. 18, 2018).

88. Screenshots on file with authors.

89. Airbnb signup process [screenshot on file with authors].

90. Xiao Ma et al., *Self-Disclosure and Perceived Trustworthiness of Airbnb Host Profiles*, PROC. OF THE CONFERENCE ON COMPUTER-SUPPORTED COOPERATIVE WORK & SOC. COMPUTING, March 2017, at 2400–01.

91. *Id.* at 2407.

counter-stereotypical information that may mitigate biases based on protected characteristics.[92]

However, the provision of personal information can also exacerbate bias on platforms, if people rely on such information to render biased decisions. Photos, of course, communicate a great deal of information, and lend themselves to inferences about gender, race, age, and other protected class characteristics. Even information about hobbies, interests, geographic location, or other features may function as proxies for such characteristics, even when they are not explicitly revealed.

Therefore, a second emerging strategy for combating bias is to purposefully *withhold* certain types of user information from other users, at least until a transaction is completed. This strategy is particularly salient for user photos and names, which can strongly indicate race and gender without providing much additional information relevant for choosing an exchange partner, to the great detriment of marginalized groups.[93] An experimental study of racial bias on Airbnb found that prospective guests with distinctively black names were 16% less likely to have their rental requests accepted than equivalent guests with white names,[94] precipitating the recommendation that Airbnb eliminate photos and substitute pseudonyms (such as "Airbnb Host" and "Airbnb Guest") for users' real names.[95] (After

92. Writer Brent Staples writes of walking down the street at night in Chicago as a young black man and "whistl[ing] popular tunes from the Beatles and Vivaldi's *Four Seasons*" to counter the negative stereotypes white pedestrians had about him as a threatening figure. CLAUDE M. STEELE, WHISTLING VIVALDI: HOW STEREOTYPES AFFECT US AND WHAT WE CAN DO 6 (2010). Similarly, the presentation of counter-stereotypical indicators in a profile may be a strategy to alleviate negative treatment in online spaces. Of course, Staples's perceived need to counteract others' stereotypes in order to exist in public space represents an enormous and unfair burden wrought by discrimination, and it must be acknowledged that the judgments rendered on the basis of counter-stereotypical information are likely themselves inflected by bias, as well. *See supra* Part II.A.2.

93. *See* Jennifer L. Doleac & Luke C.D. Stein, *The Visible Hand: Race and Online Market Outcomes*, 123 ECON. J. F469 (2013) (finding experimental evidence on an online classified marketplace that black sellers had worse market outcomes, based on photographs that included a dark-skinned or a light-skinned hand holding an identical product); Ian Ayres et al., *Race Effects on eBay*, 46 RAND J. OF ECON. 891, 910 (2015) (finding, similarly, that baseball cards held by dark-skinned hands generated lower auction prices on eBay than comparable cards held by light-skinned hands).

94. Benjamin Edelman et al., *Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment*, 9 AM. ECON. J.: APPLIED ECON. 1 (2017).

95. Benjamin Edelman, *Preventing Discrimination at Airbnb* (Jun. 23, 2016), http://www.benedelman.org/news/062316-1.html. *See also* Sarah K. Harkness, *Discrimination in Lending Markets: Status and the Intersections of Gender and Race*, 79 SOC. PSYCH. Q. 81 (2016) (finding that "gender and race significantly affect lenders'

an exchange is finalized, it may be more useful to reveal photos and names—for instance, to help transaction partners find and identify one another offline.[96])

In response to this study, Airbnb agreed to "experiment with reducing the prominence" of user photos on its platform—though it stopped short of concealing them entirely, based on the notion that "Airbnb guests should not be asked or required to hide behind curtains of anonymity when trying to find a place to stay. . . . [T]echnology shouldn't ask us to hide who we are."[97] (The authors of the Airbnb experimental study subsequently released a browser plugin, Debias Yourself, which obscures users' faces and names during Airbnb transactions; the plugin's creators encourage Airbnb hosts and guests to indicate their use of the plugin in their user profiles and photos, and the authors provide sample text to this effect, as well as a badge indicating such use to be included on profile photos.[98])

The practice of withholding certain information from a decision-maker in order to diminish the potential for bias to enter into her decisions has longstanding analogues in offline employment contexts. In one well-known study, symphony orchestras that obscured auditioning musicians behind a screen saw a marked increase in the number of women hired for positions—because decision-makers' evaluations were, presumably, less inflected by bias about the inferiority of women musicians[99]—and a number of sites and apps have imported this idea to other contexts of employment-related decision-making, by concealing certain candidate characteristics or otherwise restructuring the candidate assessment process.[100] A number of legal rules, as well as social norms, militate against requesting (or, in some cases, revealing) various types of information, in the interest of avoiding any possibility of making decisions based on contextually improper (or illegal) considerations: for instance, employers may not inquire as to a

---

funding decisions" on a peer-to-peer lending site "because they alter lenders' status beliefs about" applicants).

96. Some platforms reveal photographs of users only once a user is far into a transaction (or has completed it). The vacation rental site HomeAway withholds profile photos of hosts on search results pages, unlike Airbnb. Ray Fisman & Michael Luca, *Fixing Discrimination in Online Marketplaces*, HARV. BUS. REV. (Dec. 2016), https://hbr.org/2016/12/fixing-discrimination-in-online-marketplaces.

97. Murphy, *supra* note 18, at 23.

98. DeBias Yourself, http://debiasyourself.org/get.html. *See also infra* Part II.B.4 on badges/profile credentials.

99. *See generally* Claudia Goldin & Cecelia Rouse, *Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians*, 90 AM. ECON. REV. 715 (2000).

100. *See* Mundy, *supra* note 26 (describing a range of "anti-bias apps" in hiring).

prospective employee's disability, and colleges can lose federal funds if they request information about an applicant's marital status.[101]

However, even well-intentioned limitations on information collection can have unanticipated detrimental consequences. Over the past few years, the federal government, along with a number of cities, states, and private employers, has promulgated "ban the box" policies that prohibit asking on job applications whether prospective hires have criminal offense records. The intuition behind such policies is that employers are likely to be highly biased against former offenders, and that job seekers who answer such a question honestly are likely to be dismissed out-of-hand, making it nearly impossible for ex-offenders to find employment; therefore, these rules are intended to "level the playing field" between those with and without criminal records in the hiring process.[102] However, such measures have had a perverse consequence. Rather than refraining from consideration of offense records in their absence, employers are more likely to fall back on information that is *correlated* with offense records—namely, information about a candidate's race. Hence, ban-the-box measures can lead to statistically *worse* outcomes for black and Hispanic job candidates.[103] Thus, in place of withholding criminal history information, some economists recommend providing affirmative indicia of reliability, such as "employability certificates" that "signal an individual's work-readiness."[104]

---

101. *See* Adam M. Samaha & Lior Jacob Strahilevitz, *Don't Ask, Must Tell—And Other Combinations*, 103 CAL. L. REV. 919, 946 (2015). *See also* Lior Jacob Strahilevitz, *Reputation Nation: Law in an Era of Ubiquitous Personal Information*, 102 NW. L. REV. 1667, 1711–12 (2008) (discussing law's use of "curtains" and "search lights" to, respectively, reduce the observability of certain types of information, or draw attention thereto, for policy purposes).

102. Jennifer L. Doleac, *"Ban the Box" Does More Harm than Good*, BROOKINGS INSTITUTION (May 31, 2016), https://www.brookings.edu/opinions/ban-the-box-does-more-harm-than-good/; Note that employers following ban-the-box policies typically *may* ask prospective hires about their pasts at the interview stage; at that point, it is assumed that candidates will have more opportunity to explain their situations, and employers will have a "fuller picture" of who the candidate is.

103. *See id. See also* Jennifer L. Doleac & Benjamin Hansen, *Does "Ban the Box" Help or Hurt Low-Skilled Workers? Statistical Discrimination and Employment Outcomes When Criminal Histories are Hidden*, NAT. BUR. OF ECON. RES. WORKING PAPER NO. 22469 (Jul. 2016), http://www.nber.org/papers/w22469; Amanda Agan & Sonja Starr, *Ban the Box, Criminal Records, and Statistical Discrimination: A Field Experiment*, BECKER FRIEDMAN INST. FOR RES. IN ECON. WORKING PAPER NO. 2016–17 (Jul. 2016), http://bfi.uchicago.edu/sites/default/files/research/2016-17.pdf.

104. Jennifer Doleac, *More Job Opportunities, Less Recidivism*, REALCLEARPOLICY (Dec. 15, 2016), http://www.realclearpolicy.com/articles/2016/12/15/more_job_opportunities_less_recidivism.html; Peter Leasure & Tia Stevens Andersen, *The Effectiveness of Certificates of Relief as Collateral Consequence Relief Mechanisms: An*

Furthermore, even when online environments intentionally offer few explicit cues of a user's characteristics, users may nevertheless be able to readily infer (and behave differently based on) those characteristics. For example, an experimental study of users on eBay found that users were able to accurately identify the gender of an eBay seller the majority of the time, even in the absence of user photos, real names, or an explicit profile indicator of gender.[105] The study's authors also found that women sellers on eBay suffered significant penalties for their gender, earning about 80 cents on the dollar earned by male sellers for identical new products.[106]

The eBay study and the ban-the-box case suggest that strategies premised on suppressing information about users, while holding promise for reducing the potential for bias, must be carefully considered in the broader context of what information *is* visible on the platform. Users may readily default to what information is available about a counterparty, resulting in less effective (or potentially even detrimental) interventions.

Third, in addition to the amount of personal disclosure permitted (or required or disallowed), the forms that such presentation is permitted to take may influence its role in supporting users' biases. For instance, structuring information via predefined fields—as opposed, say, to free text—allows platforms to define input categories in advance as well as acceptable inputs for each category. In so doing, platforms may attempt to delimit what information they want to permit users to draw from in making decisions:[107] for instance, by excluding the capability to list one's religion, platforms may attempt to insulate user activity from consideration thereof.[108] However, decisions about how information is structured can also introduce other forms of bias: consider, for instance, asking for a user's gender as a binary

Experimental Study, YALE L. & POL'Y REV. INTER ALIA (Nov. 7, 2016), http://ylpr.yale.edu/inter_alia/effectiveness-certificates-relief-collateral-consequence-relief-mechanisms-experimental; Peter Leasure & Tara Martin, *Criminal Records and Housing: An Experimental Study*, J. EXP. CRIMINOL. (2017); *see also infra* Part II.B.4 on profile badges and credentials.

105. Tamar Kricheli-Katz & Tali Regev, *How Many Cents on the Dollar? Women and Men in Product Markets*, 2 SCI. ADVANCES 1 (2016). The array of goods the seller had for sale seemed to be a reliable proxy for gender, as "the probability of correctly identifying the gender of the seller increased by 5% with every additional item for sale on display on the seller's profile." *Id.* at 6.

106. *Id.* at 1.

107. This is tempered, however, by the aforementioned "ban-the-box" effects, in which users glean signals from correlated variables.

108. However, structuring inputs in this manner may place a platform at a greater risk for liability based on its users' behavior; *see supra* discussion of *Roommates.com*, notes 10–14.

variable, and how such structuring can operate as a vehicle for exclusion of users with other gender identities.[109]

Platforms may also discretize how information is displayed to users at various points in decision processes (i.e., on separate pages of a user interface). "Chunking" involves pulling out specific attributes individually and asking decision-makers to compare across them. In the employment context, the hiring software beApplied "reorders applications horizontally so that reviewers just focus on comparing responses to individual questions. Since applications are also blind, you know you're assessing the quality of each response fairly."[110] The goal of doing so is to minimize "halo effects" from previous knowledge about a candidate (i.e., letting one's judgment be colored by knowledge about a person's gender, age, etc.); in a sense, such discretization is the logical opposite of platforms' "whole person" design strategies described above.

Finally, platforms may authenticate users' identities by linking them to external profiles or by requiring users to use their real names on the platform. Many platforms allow, or require, users to link accounts to a Facebook, LinkedIn, or other social media profile (see Figure 4 below), often in the interest of preventing scams.[111] Others—most notably, Facebook—require that users' profiles use their "real names," as opposed to a pseudonymous handle.[112]

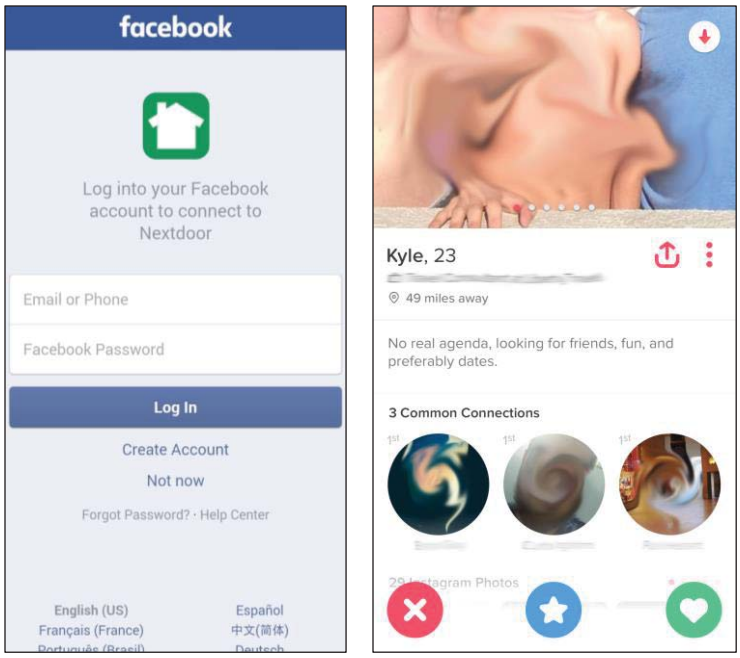
109. Sabrina Fonseca, *Designing Forms for Gender Diversity and Inclusion*, UXDESIGN.CC (Apr. 24, 2017), https://uxdesign.cc/designing-forms-for-gender-diversity-and-inclusion-d8194cf1f51; *see also* discussion of Tinder's gender categories *infra* notes 170–183.

110. BEAPPLIED, https://www.beapplied.com/features (last visited September 6, 2017).

111. For instance, the localized goods exchange app 5miles requires two forms of external identification (phone number, email address, or Facebook profile) to be linked to a user; the related app OfferUp requires Facebook access and a photograph of the user's driver's license. Roy Furchgott, *Decluttering? Yes, There's an App*, N.Y. TIMES (Apr. 7, 2017), https://www.nytimes.com/2017/04/07/realestate/spring-cleaning-and-decluttering-help-apps.html.

112. Facebook's real name policy states that "Facebook is a community where everyone uses the name they go by in everyday life. This makes it so that you always know who you're connecting with and helps keep our community safe." *What names are allowed on Facebook?*, FACEBOOK HELP CENTER, https://www.facebook.com/help/112146705538576 (last visited September 6, 2017).

Figure 4: Nextdoor allows users to authenticate themselves using their Facebook login credentials; Tinder allows users to link their accounts to their Facebook accounts, highlighting if users share friends in common.



Authentication mechanisms might mitigate bias in at least three ways. Linkage with activity on other platforms or with offline identities could coax users into self-modulating their conduct in socially desirable ways, out of a sense of greater accountability for their actions—though whether such accountability would extend to implicit biases is an open question. Second, seeing *other* users' linkages to other arenas might operate as a humanizing signal of their identity as a "whole person" that might mitigate reliance on stereotypes (much like user profile information, discussed *supra*) and increase trust. Finally, certain platforms attempt to cultivate even greater confidence among users by showcasing when they share friends in common on the outside platform through which they have authenticated their identities. For example, Tinder notes the number of Facebook friends that a user shares with the person whose profile the user is viewing. Making such connections explicit might encourage users to consider potential matches that they would have been dismissed instinctively otherwise, potentially on prejudicial or biased grounds. At the same time, identification mechanisms could provide additional fodder on which to base biased decisions. For instance, real name policies might prove detrimental to platform users whose names are strongly associated with particular races or ethnicities, and users may be limited from adopting indicators of identity that make them

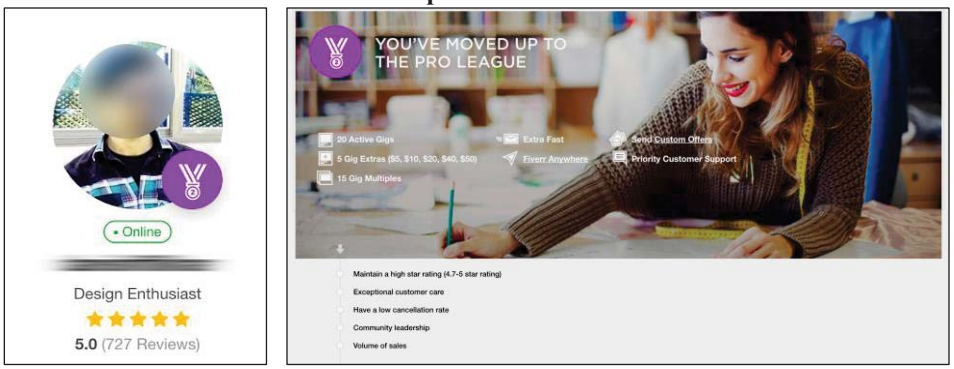
appear less "marked." Identifying common connections can also have the effect of encouraging users to interact with people already in their social network, thereby reproducing disparities in social capital along the lines of race and other protected characteristics.

4. Reputation, reliability, ratings

The strategies described in the previous section help users learn about each other's identities, to decide whether and how to interact with one another. A related set of platform strategies aims to equip users with indicators about what to expect from a transacting partner on the site based on *past behavior*. These indicators can confer a sense of expertise, reputation, or trustworthiness that may increase counterparties' trust in the exchange,[113] and may provide countervailing signals that can mitigate users' implicit biases.

Some indicia of reliability take the form of badges or other graphic elements on users' profiles. Badges may indicate a certain amount of engagement or longevity with a platform (e.g., a certain number of tasks completed), or a certain degree of quality, perhaps operationalized as a high composite rating. The labor marketplace Fiverr, for instance, allows sellers to "level up" based on experience and ratings; when a new level is achieved, a badge is displayed prominently on the user's profile (see Figure 5).[114]

Figure 5: Fiverr displays badges on a user's profile based on experience on the platform



113. Audun Jøsang et al., *A Survey of Trust and Reputation Systems for Online Service Provision*, 43 DECISION SUPPORT SYS. 618, 621–22 (2007); *see also* Part II.A.2 *supra*.

114. FIVERR'S LEVELS, https://www.fiverr.com/levels (last visited September 6, 2017).

Testimonials, references, and reviews may also serve as signals of reliability because they may pertain to a user's general character and reputation. For instance, Airbnb permits users to post "references from your personal network . . . from people who know you well" which "will help other members get to know you"[115], and LinkedIn permits users to offer recommendations[116] and endorsements about another user's skills.[117] Alternatively, these indicators may be based on particular past interactions on the platform; eBay, Airbnb, Etsy, and many other platforms allow users to write reviews of past interactions, and make these reviews visible to the broader user base as a means of broadcasting reputation. A recent field experiment on Airbnb found that booking requests from black guests (as indicated by distinctively African-American names) had a lower acceptance rate than those from white guests; however, when each guest's profile had one positive review, the acceptance rate was almost identical, suggesting that the presence of a review acted as a counter-stereotypical signal that alleviated bias.[118] Matched negative reviews had the same effect at removing disparities in acceptance rates between black and white guests.[119]

One of the most ubiquitous forms of user evaluation involves rating a counterparty on the quality of an exchange, most commonly by assigning them a number of stars. Uber, Lyft, eBay, Instacart, Postmates, and myriad other platforms have ratings systems for evaluation of tasks; individual ratings contribute to a composite rating that is typically displayed on the user's profile, operating as an indicator of reliability and satisfaction.[120] Ratings are typically platforms' strongest signals of customer satisfaction, and are relied upon for a number of purposes, including as a threshold for

115. Airbnb host sign up process [screenshot on file with authors]. *See generally What are References on Airbnb?*, AIRBNB, https://www.airbnb.com/help/article/173/what-are-references-on-airbnb (last visited Mar. 18, 2018).

116. LINKEDIN, *Recommending Someone, LinkedIn Help*, https://www.linkedin.com/help/linkedin/answer/97 (last visited September 6, 2017).

117. LINKEDIN, *Skill Endorsements -- Overview, LinkedIn Help*, https://www.linkedin.com/help/linkedin/answer/31888/skill-endorsements-overview (last visited September 6, 2017).

118. Ruomeng Cui et al., *Discrimination with Incomplete Information in the Sharing Economy: Field Evidence from Airbnb* (Jan. 9, 2017), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2882982; *see also* Jun Li et al., *A Better Way to Fight Discrimination in the Sharing Economy*, HARV. BUS. REV. (Feb. 27, 2017), https://hbr.org/2017/02/a-better-way-to-fight-discrimination-in-the-sharing-economy.

119. *Id.*

120. Caroline O'Donovan, *That Four-Star Rating You Left Could Cost Your Uber Driver Her Job*, BUZZFEED (Apr. 11, 2017), https://www.buzzfeed.com/carolineodonovan/the-fault-in-five-stars.

deactivation,[121] an indicator of problematic transactions, or a basis for subsequent sorting, matching, and filtering by users.

Ratings tend to be a coarse form of evaluation; they require the distillation of multifaceted experiences into a discrete value, and are very rarely accompanied by more specific justification or explanation of the rater's source of (dis)satisfaction.[122] Further, rating interfaces seldom include guidelines about what qualities of the interaction a rater ought, or ought not, to consider. For instance, should the rating encompass only the timeliness of a service or the quality of goods delivered? Should it also pertain to the personal interaction between users (i.e., how much the rater liked or felt affinity for the ratee)? This coarseness makes the act of rating a ready conduit for bias to enter into user interactions. As we have described at length elsewhere,[123] there is ample risk that rating processes on platforms may systematically disadvantage marginalized groups, who may receive lower aggregate ratings than other groups; social science research on workplace evaluations finds such effects.[124] In some cases, if platforms make material employment determinations based on consumer-sourced ratings, they may create a facially neutral avenue through which discrimination can creep into employment decisions, despite the fact that a company would be prohibited from making such biased assessments directly.[125]

C. MONITORING AND EVALUATING

Platforms may rely on a diverse set of methods to identify, sanction, and correct for biased behavior among their users. They may create infrastructures through which users can report apparent cases of discrimination that may serve as the basis for sanction. However, such reports can be a way for users to discriminate against one another if users abuse the reporting mechanism to falsely accuse others. Platforms may also take steps to improve the quality of evaluations rendered by users—for instance, by requiring users to submit more granular information in suspect cases, by validating evaluations with independent data sources, or by

121. *See* Rosenblat et al., *supra* note 5.

122. However, some platforms may seek more granular evaluations for particularly poor ratings; *see* Part II.C.2, *infra*.

123. Rosenblat et. al., *supra* note 5.

124. *See, e.g.*, Emilio J. Castilla, *Gender, Race, and Meritocracy in Organizational Careers*, 113 AM. J. SOC. 1479 (2008); Marta Elvira & Robert Town, *The Effects of Race and Worker Productivity on Performance Evaluations*, 40 INDUS. REL. 571 (2001).

125. Rosenblat et. al, *supra* note 5; *see also* Dallan F. Flake, *When Should Employers Be Liable for Factoring Biased Customer Feedback into Employment Decisions?*, 102 MINN. L. REV. __ (Forthcoming 2018).

reweighting evaluations suspected to be influenced by bias. Finally, platforms can measure how certain key outcomes vary according to users' race, gender, and other protected characteristics that they can report publicly or make accessible to regulators and outside researchers. They can also perform controlled experiments or rely on natural experiments to assess whether disparities in outcome owe to differences in these characteristics alone.

1. Reporting and sanctioning

Another set of strategies involves infrastructures for reporting behavior that seems to exhibit bias, and sanctioning users who propagate it. In creating these mechanisms, platforms often take their cues from users who report witnessing or being subject to perceived biased behavior. Such reporting systems are common on social media platforms for marking *explicit* manifestations of bias, offensive content, and overt harassment—often through a flagging system, which may automatically remove content or refer it to a site moderator for review.[126]

Reporting and sanctioning mechanisms have been implemented in attempts to mitigate implicit bias *offline* as well. Complaints to the Equal Employment Opportunity Commission commonly allege that an employer (or prospective employer) behaved in a manner that disproportionately impacted members of a protected class; these complaints act as a trigger for further investigation by the EEOC. A number of colleges and universities have recently launched bias response hotlines and reporting mechanisms aimed at improving campus climate, often including both overt manifestations of bias (e.g., hate speech) as well as implicitly or unintentionally biased behavior.[127]

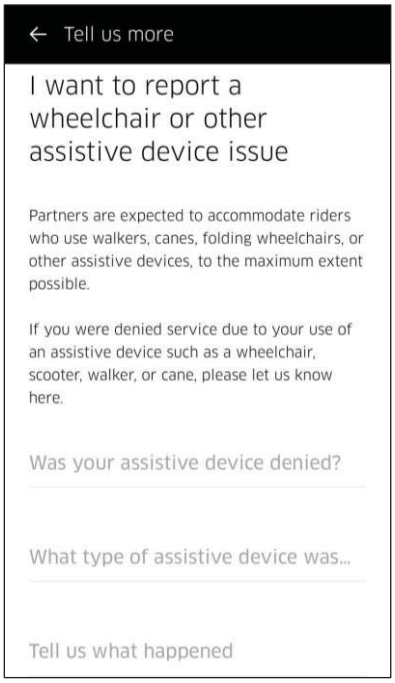
Like these offline analogues, platform bias reporting mechanisms typically elevate concerns institutionally by referring them to a platform

126. *See, e.g.,* J. Nathan Matias et al., *Reporting, Reviewing, and Responding to Harassment on Twitter*, WOMEN, ACTION, AND THE MEDIA REPORT (May 13, 2015), http://womenactionmedia.org/twitter-report; Kate Crawford and Tarleton Gillespie, *What is a Flag For? Social Media Reporting Tools and the Vocabulary of Complaint*, 18 NEW MEDIA & SOC'Y 410, 411 (2016) ("'Flagging'—a mechanism for reporting offensive content to a social media platform—is found on nearly all sites that host user-generated content, including Facebook, Twitter, Vine, Flickr, YouTube, Instagram, and Foursquare, as well as in the comments sections on most blogs and news sites.").

127. At some campuses, such mechanisms have instigated controversy around concerns about their potential chilling effects on academic freedom. *See* Jake New, *Defending BARTs*, INSIDE HIGHER ED (Sep. 12, 2016), https://www.insidehighered.com/news/2016/09/12/despite-recent-criticism-college-officials-say-bias-response-teams-fill-important.

representative or site moderator with authority to adjudicate or investigate the issue. Airbnb's new Open Doors policy, for instance, ensures that guests who report having been unable to find a rental due to discrimination can receive "timely, 24/7, personalized, hands-on support from a specially trained Airbnb employee" who will find the guest a similar listing or an "alternative accommodation option" (presumably, a hotel).[128] Uber riders who report, for instance, a driver's refusal to accommodate a walker or other assistive device can submit a report (see Figure 6); such a report temporarily deactivates the driver account while the company reviews the incident, and confirmed violations of the law may result in the driver's deactivation from the platform.[129]

Figure 6: Uber's mechanism for reporting a driver's refusal to accommodate assistive devices



Systems that rely on users to report manifestations of implicit bias may be difficult to implement in practice. Unlike explicitly discriminatory or harassing conduct, users may lack access to signals that indicate when implicit bias is likely at work or how the design of a platform might

128. Murphy, *supra* note 18, at 21.

129. *Accessibility at Uber*, UBER, https://accessibility.uber.com/#our-policies (last visited September 6, 2017).

exacerbate or mitigate it. Users may have very little insight into how similarly situated users are treated on the platform relative to others; and because they may not interact with the same user on the platform repeatedly, it may be difficult to assess a pattern of behavior (as compared, say, to a managerial relationship in a traditional employment context). In addition, it may be difficult for users to understand what constitutes bias, so that they can usefully report it. Nextdoor, in its attempts to mitigate racial profiling on its platform, initially asked users to flag posts for "racial profiling," a solution that was later deemed inadequate because "many people didn't understand what it was, and Nextdoor members began reporting all kinds of unrelated slights as racial profiling."[130]

In addition, bias reporting systems may *themselves* operate as mechanisms through which bias can be instantiated on a platform. Users may report each other as a means of personal attack, retribution, or to police viewpoints with which they disagree[131]—and different groups of users may be differentially reported on platforms, reducing their ability to participate as part of the community.

Platforms may also sanction users for behavior that appears to exhibit bias *without* reliance on the user community to report it.[132] Airbnb noticed a problem wherein potential guests would attempt to book a listing listed as available, only to be told by hosts that it was not, in fact, vacant for the dates in question—and that those listings were then sometimes booked by guests of a different race.[133] In response, Airbnb changed its platform to automatically prevent a listing from being subsequently booked for a given date if a host tells a potential guest that the space is unavailable. By making it structurally impossible for a host to rebook a space for a "more desirable" guest, Airbnb aims to discourage behavior likely to be inflected with bias.

2. Data quality and validation

Users commonly provide feedback on each other's performance in the course of using a platform—by rating one another, leaving reviews on past transactions, and the like. Though such activity can provide a basis for trust and reliability with unknown partners, it is also likely to be inflected by users' implicit biases and may therefore result in systematically worse

130. Hempel, *supra* note 24.

131. Crawford & Gillespie, *supra* note 126, at 420, 423–24. *See also* our discussion of reporting and transphobia on Tinder, *infra*, footnotes 170–83.

132. *See also infra* Part II.C.2.

133. Murphy, *supra* note 18, at 20.

outcomes for users from marginalized groups.[134] To ameliorate these effects, platforms may seek to improve the quality of evaluations that users tender to one another on a platform. They may do this by requiring more granular information of users in suspect cases, in efforts to make users reflect more precisely on the factors on which their evaluations depend.[135] They may also mitigate the effects of bias (if not bias itself) by adjusting ratings or de-listing reviews likely to be impacted by bias, perhaps using machine learning techniques to detect high- or low-quality evaluations. Finally, platforms may require validation of poor evaluations with external sources of data.

Nextdoor has been the subject of significant controversy in recent years, following media coverage of the platform's users engaging in racial profiling when reporting nearby crimes or suspicious activities.[136] The platform explored a number of strategies to address the problem, ultimately adopting a number of different approaches,[137] including changes to the interface where users report such activity. In particular, Nextdoor now notes when users rely on race to report a crime or suspicious activity,[138] operating under the assumption that such reports are likely to be biased. If this occurs, Nextdoor prompts users to first describe the incident without describing the people involved in the incident. Once users have submitted this information, they are then taken to a second prompt where Nextdoor asks users to fill in predefined fields describing those involved in the incident—and users must fill in at least two of four fields, none of which are related to race (see Figure 7). By forcing users to provide more specific and granular information, Nextdoor limits the degree to which the reporting of crime or suspicious activity can rely solely on the race of the person involved in the incident. While the platform has found that imposing this additional burden seems to discourage users from reporting such events, its leadership believes that it encourages users to think more carefully about the cause of their suspicion and provide more accurate and useful reports.[139]

134. *Id.*

135. Hempel, *supra* note 24.

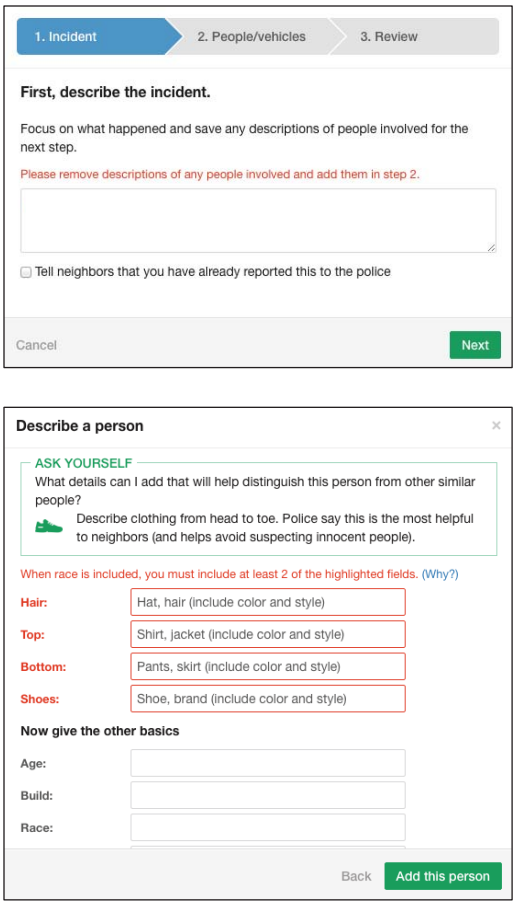
136. Pendarvis Harshaw, *Nextdoor, The Social Network for Neighbors, Is Becoming a Home for Racial Profiling*, FUSION (Mar. 24, 2015), https://fusion.kinja.com/nextdoor-the-social-network-for-neighbors-is-becoming-1793846596.

137. *See infra* Parts II.A.2 and II.B.1.

138. Hempel, *supra* note 24 ("If you refer to race in your description of the incident, Nextdoor's algorithms detect it and prompt you to save this part of the description for the next screen.").

139. *Id.*

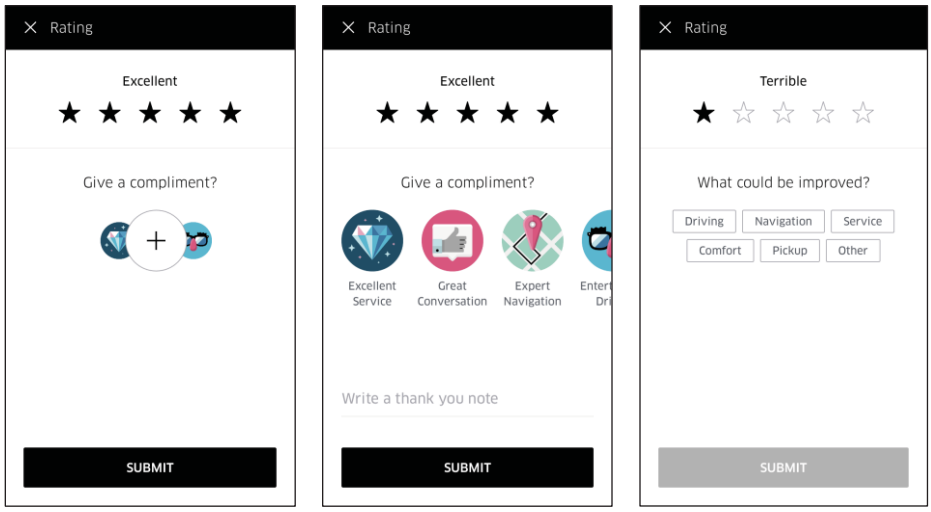
Figure 7: Nextdoor's prompts when users rely on race to report suspicious activity



Uber employs a similar strategy in its rating system, where riders can choose among a set of predetermined and specific compliments to accompany their five-star ratings of drivers, ranging from "Great Conversation" to "Expert Navigation" to "Neat and Tidy". When riders provide ratings lower than five stars, Uber asks "What could be improved?" and provides riders with a set of predefined answers (Comfort, Driving, Navigation, Pickup, Service, and Other) (see Figure 8). While Uber prompts and often requires riders to provide a star rating to drivers, giving a specific compliment is entirely optional. In contrast, Uber may require riders to specify what drivers could have improved when riders give drivers less than five stars. In both cases, Uber seems to want to find a way to solicit more precise and actionable information from riders than the company and drivers might glean from stars on their own. As with Nextdoor, requesting this additional information may impose an additional burden that users

might not always be willing to shoulder, but it can also help to reduce the likelihood that riders are assessing drivers merely on the basis of who the drivers happen to be and not on the quality of their service.[140]

Figure 8: Uber's interface allows riders to give compliments following 5-star ratings, and requires riders to provide suggestions for improvement following 1-star ratings



The quality of users' ratings and reviews can vary dramatically.[141] Some users may invest considerable time and thought in their evaluation while others might pass quick judgment.[142] Less deliberative assessments are likely to be of poorer quality, of course, but also more likely to rely on crass heuristics and thus involve implicit bias. Platforms that rely on users' ratings and reviews tend to be well aware of this problem, and researchers

140. Rosenblat, *supra* note 121.

141. Susan M. Mudambi & David Schuff, *What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com*, 34 MGMT. INFO. SYS. Q. 185, 186 (2010); Stefan Siersdorfer et al., *How Useful Are Your Comments?: Analyzing and Predicting YouTube Comments and Comment Ratings*, PROC. ACM INT'L CONF. ON WORLD WIDE WEB 891, 892 (2010).

142. Data quality adjustments are commonly made in related data-collection contexts to guard against manipulation, inattention, and other sources of inaccuracy. Daniel M. Oppenheimer et al., *Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power*, 45 J. EXPERIMENTAL SOC. PSYCH. 867, 868 (2009); Chrysanthos Dellarocas, *Immunizing Online Reputation Reporting Systems Against Unfair Ratings and Discriminatory Behavior*, PROC. ACM CONF. ON ELECTRONIC COMMERCE 150 (2000); Andrew Whitby et al., *Filtering Out Unfair Ratings in Bayesian Reputation Systems*, 6 PROC. WORKSHOP ON TRUST IN AGENT SOCIETIES 106 (2004).

have developed a variety of techniques to address it.[143] Yelp, for example, automatically evaluates the quality of users' reviews and prioritizes them accordingly.[144] The platform purposefully does not highlight reviews from new users or users about whom Yelp knows very little; it attempts to weed out reviews from "family, friends, or favored customers" or reviews purchased by the business owner (in an attempt to either benefit the business or hurt a competitor); and it tries to avoid "unhelpful rants and raves," which the company does not define.[145] While many of these reviews remain accessible to interested users, Yelp itself will not factor the scores from these reviews into a business' average score. To the extent that biased assessments are generally more likely to occur in assessments of poor quality, automated systems that aim to remove such reviews and ratings or prioritize high quality evaluations will likely reduce how much bias affects those subject to such evaluations.[146] Yet Yelp's application of such techniques has not been without controversy, in large part because it reveals just how much power it wields in deciding how businesses ultimately fare on its platform.[147]

Adjusting for data quality in terms of removing user *bias* is even more complicated normatively, in that such corrections imply that users' biased judgments are less valid and ought not be considered.[148] Despite this complexity, platforms may still see fit to identify and adjust biased data, to the extent that they can, to diminish the systemic effects of bias on marginalized users. Or they might decide that certain decisions are too consequential to hinge on ratings and reviews from which potential bias

143. *See, e.g.,* Yang Liu et al., *Modeling and Predicting the Helpfulness of Online Reviews*, IEEE INT'L CONF. ON DATA MINING 443 (2008); Dellarocas, *supra* note 142.

144. Yelp, Inc., *How Yelp Helps You Find the Right Local Business*, YELP BLOG (Nov. 13, 2013), https://www.yelpblog.com/2013/11/yelp-recommended-reviews.

145. Yelp stresses that its "recommendation software is entirely automated so that it can apply the same objective standards to every business and every review without being overridden by someone's personal preferences." *Id.*

146. Yelp reviews nevertheless continue to exhibit bias. *See* Sharon Zukin et al., *The Omnivore's Neighborhood? Online Restaurant Reviews, Race, and Gentrification*, J. CONSUMER CULTURE 1 (2015).

147. Jay Barmann, *Yelp is Allowed to Manipulate Ratings and Remove Good Reviews, Says Court*, SFIST (Sep. 4, 2014), http://sfist.com/2014/09/04/yelp_is_allowed_to_manipulate_ratin.php.

148. *See* Rosenblat, *supra* note 5, at 15 ("[t]he suggestion that implicit or explicit consumer biases ought not inflect [user-to-user] ratings … —or at least, that platforms ought to account and correct for the likely presence of such biases—represents a complex normative judgment, and we must acknowledge that adjustments to correct for bias in this context are therefore more normatively laden than adjustments made to correct for systematic error (e.g., sampling bias) in standard data analysis").

cannot be completely purged, in an attempt to limit the effects of biased ratings rather than address the bias itself.[149]

Finally, platforms might rely on alternative sources of information to confirm or validate users' claims. For example, should one user give another a low rating, the platform might ask the initial user to furnish documentary evidence to support broad claims about the quality of the service she received from the other user. Airbnb, for example, might ask the user to photograph any problems with the property. Or the platform itself might try to collect or repurpose data that would allow it to serve as a reliable source against which to judge the validity of users' claims. Uber might examine the location data it collects from riders' phones to see whether they were late to meet their drivers; the data Uber collects from drivers' phones might help to confirm whether they made any unsafe or erratic maneuvers.[150]

While soliciting high quality and informative feedback from users can help to mitigate bias, doing so is not without costs. If platforms ask users to complete more detailed reviews—and therefore spend more time and thought on their assessments—platforms may find that fewer users are willing to even complete the process. If platforms instead attempt to evaluate the quality and reliability of users' assessments and adjust or discount these accordingly, platforms may court controversy by exercising direct control over the relative standing of different users, even if the effect may be to reduce the influence of bias in these users' ratings. And because much of the success of platforms owes to the fact that they have been able to push a good deal of the bureaucracy that comes along with traditional service providers onto users themselves, platforms might balk at the idea of investing the resources necessary to perform user evaluations themselves, even if these might be much less biased than those currently performed by platforms' untrained users.

3. Measurement and detection

Finally, platforms may make independent efforts to measure any potentially disparate effects of their design decisions or to detect bias in the behavior of their users. These approaches draw from offline analogues like the collection of demographic data[151] and the use of audit or correspondence

149. Rosenblat, *supra* note 5.

150. *Id.*

151. EQUAL EMPLOYMENT OPPORTUNITY COMMISSION, *Race/Ethnicity Self-Identification Forms*, https://www.eeoc.gov/employers/eeo1survey/sample_self_identification.cfm (last visited September 6, 2017).

studies in such areas as employment, housing, and credit, among others.[152] Platforms might seek to collect demographic data from their users directly or infer it from other disclosed information.[153] In either case, such details would serve as the necessary foundation to establish whether users from different protected groups fare differently on these platforms, but not whether such differences were the result of biased decision-making.[154] These findings could be reported publicly, much like the transparency reports about government requests for user data that have become common among the major online platforms.[155] As part of these reports, platforms might also describe their methodologies and release the underlying data.[156] In April 2017, California's Department of Fair Employment and Housing entered into a voluntary agreement with Airbnb to resolve the agency's prior complaint against the company for violations of the California Fair Employment and Housing Act and Unruh Civil Rights Act.[157] As part of this agreement, Airbnb assented to generating and sharing such reports with DFEH, noting the average "relative acceptance rate" for users of different races, among other things.[158] The Agreement even suggests that Airbnb

152. Devah Pager & Hana Shepherd, *The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets*, 34 ANN. REV. SOC. 181, 184–85 (2008).

153. Platforms might be reluctant to collect this information or attempt to infer it, as users might perceive such activities as a privacy violation or posing risks of discrimination. Separately, users might not want to volunteer demographic information, especially if they are concerned with discrimination.

154. A disparate impact case would start with the same analysis: a showing, for example, that female job, housing, or credit applicants fare systematically worse than male applicants.

155. Laura DeNardis and Andrea M. Hackl, *Internet Governance by Social Media Platforms*, 9 TELECOMM. POL'Y 39, 761–70 (2015). *See also* Aaron Belzer & Nancy Leong, *The New Public Accommodations*, 105 GEO. L.J. 1271, 1319 (2017) (proposing federal legislation to mandate such disclosure by platforms and noting the benefits of providing such data for researchers).

156. Benjamin Edelman, *Response to Airbnb's Report on Discrimination* (Sep. 19, 2016), http://www.benedelman.org/news/091916-1.html ("Certainly Airbnb could provide the interested public with aggregate data measuring discrimination and showing the differential outcomes experienced by white versus black users. If Airbnb now has mechanisms to measure discrimination internally, as the report suggests, it's all the more important that the company explain its approach and detail its methodology and numerical findings—so past outcomes can be compared with future measurements.")

157. California Dep't of Fair Emp. and Housing (DFEH), *Voluntary Agreement between Airbnb and DFEH* (Apr. 19, 2017), https://www.dfeh.ca.gov/files/2017/04/04-19-17-Airbnb-DFEH-Agreement-Signed-DFEH-1-1.pdf.

158. Lyft has adopted a similar approach in response to a letter from former Senator Al Franken, raising concerns about disparate rates at which drivers cancelled rides for black and white passengers. *See* Alan Franken, *Letter to Travis Kalanick and Logan Green* (Nov.

consider creating something functionally akin to transparency reports, but for hosts: a gallery of the guests that hosts have rejected. Such a gallery would act as a kind of mirror through which hosts would be able to take a hard look at their past decisions, possibly revealing patterns of prejudice that would shame hosts or highlighting apparent, but unrecognized, bias that would spur them to alter their behavior. Where guests' races are known, the platform could instead communicate to hosts the relative rates at which they accept guests of different races.

Platforms could also experiment with design choices and observe if they result in any corresponding change in outcomes for minority and marginalized populations. For example, Airbnb has publicly committed to "perform[ing] tests[,]…examin[ing] algorithms, and mak[ing] ongoing adjustments to the technical underpinnings of [its] platform" to explore what might help address the incidence of apparent discrimination. The goal of such experimentation need not be to determine when and where users act in a prejudicial or biased manner; rather, it could simply be to assess whether adjustments to the user experience can help minimize or eliminate disparities in outcomes, regardless of the underlying and ultimate cause of the disparities. Of course, platforms could also experiment to determine which types of interventions are most effective in addressing users' biases more directly. Over a three-month period, Nextdoor tried a number of approaches, which they evaluated through A/B testing, before settling on a final set of strategies.[159] Ray Fisman and Michael Luca have described this as "[maintaining] an experimental mindset," calling on companies that make extensive use of such experimental techniques in product and service development to apply them to the problem of discrimination as well.[160]

Platforms might be more ambitious and attempt to estimate the extent to which bias affects users' decisions by relying on either natural or controlled experiments. In the former, platforms might seek out seemingly

---

2, 2016), https://www.franken.senate.gov/files/letter/161102_UberLyft.pdf. As a means to address the issue, Lyft stated it would enhance its regular review of ride cancellations by "including a focus on cancellation rates and quality of service in 'minority census tracts.'" Logan Green, *Letter in Response to Nov. 2, 2016 Letter to Travis Kalanick and Logan Green* (Dec. 16, 2016), https://www.franken.senate.gov/files/letter/161216LyftResponseLtr.pdf. Franken's letter followed a paper published by the National Bureau of Economic Research establishing patterns of discrimination on both Uber and Lyft in Seattle, WA, and Boston, MA. Yanbo Ge et al., *Racial and Gender Discrimination in Transportation Network Companies* (Nat'l Bureau of Econ. Research, Working Paper. No. w22776, 2016). Many of the paper's recommendations appear directly in Lyft's response to Franken.

159. Hempel, *supra* note 24.

160. Fisman & Luca, *supra* note 1, at 94.

equivalent cases that resulted in different outcomes, where only a difference in race, for example, seems to account for differences in decision-making. In the latter, platforms might devise experiments of their own where they purposefully generate cases that only differ in the race of test users. At the extreme, platforms could even administer psychological tests to directly measure implicit bias among a subset of their users.[161] Armed with the results of these experiments or tests, platforms could then use machine learning to uncover relationships between the more easily observable qualities or behaviors of users and their propensity for biased decision-making. In effect, these platforms would be able to estimate how much bias likely influences each user's decisions. These strategies are not entirely hypothetical: Airbnb has stated publicly that as part of its effort to address discrimination, the platform is exploring how machine learning might "help enforce our anti-discrimination policy."[162]

In addition to taking steps to measure bias and its effects *themselves*, platforms might also take steps to open their data to independent scrutiny by researchers or government entities. A number of social science researchers are interested in conducting studies to detect and measure discrimination on platforms, but often the methods required for doing so are expressly prohibited by a platform's terms of service.[163] Platforms routinely block researchers' accounts when they are suspected of engaging in such research.[164] What's more, researchers who try to detect discrimination on platforms may be subject to criminal penalties; the Computer Fraud and Abuse Act (CFAA)[165] prohibits "unauthorized access" to a computer, which has been interpreted to include terms-of-service violations.[166] Platforms

161. Brian Uzzi, *How Human-Machine Learning Partnerships Can Reduce Unconscious Bias*, ENTREPRENEUR (Jul. 31, 2016), https://www.entrepreneur.com/article/278214.

162. David King, *A Fair Community for Everyone*, AIRBNBCITIZEN (May 11, 2016), https://www.airbnbcitizen.com/a-fair-community-for-everyone/.

163. Christian Sandvig et al., *Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms*, PROC. WORKSHOP ON DATA AND DISCRIMINATION, INTL. COMM. ASSOC. (2014), http://www-personal.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf.

164. *See, e.g.*, Sam Levin, *Airbnb Blocked Discrimination Researcher Over Multiple Accounts*, THE GUARDIAN (Nov. 17, 2016), https://www.theguardian.com/technology/2016/nov/17/airbnb-while-black-discrimination-harvard-researcher-banned.

165. 18 U.S.C. § 1030.

166. The American Civil Liberties Union and several social scientists are currently challenging the constitutionality of the CFAA in light of this issue. *See* Russell Brandom, *New ACLU Lawsuit Takes on the Internet's Most Hated Hacking Law*, THE VERGE (Jun.

might therefore alter their terms of service to permit "bona fide testing"[167] in the service of removing barriers to researchers' detection of bias (for instance, by permitting researchers to operate multiple accounts in order to compare outcomes across race or gender). In addition to facilitating researchers' access, platforms might open their data to scrutiny by government regulators. Airbnb's agreement with the California Department of Fair Employment and Housing requires Airbnb to permit the agency to conduct fair housing testing—essentially, an audit study—through which the agency will set up multiple profiles to discern differential treatment.[168] Under the agreement, Airbnb will further provide the agency with the names of hosts who are suspected of discrimination for testing purposes.[169]

* * *

The ten categories we describe above are ideal types; in practice, platforms' structures are likely to encompass multiple categories. Some features are likely to function in combination—a platform that requests or requires that users disclose particular fields of information about themselves, for instance, may concomitantly allow other users to search and filter by those fields.

But interactions among design features can be more complex as well, with implications for how bias is instantiated on the platform. Consider, for instance, the dating app Tinder's treatment of its transgender users.[170] Tinder's user interface is designed to be minimal and low-friction, such that users simply swipe left or right on each other's profiles to indicate interest in one another, often based on little more information than a profile picture.[171] Until recently, Tinder permitted people to list one of two options for their gender—male or female—without further specification of gender identity; users could specify if they were interested in being matched with

29, 2016), https://www.theverge.com/2016/6/29/12058346/aclu-cfaa-lawsuit-algorithm-research-first-amendment.

167. Benjamin Edelman, *Response to Airbnb's Report on Discrimination* (Sep. 19, 2016), http://www.benedelman.org/news/091916-1.html.

168. California Dep't of Fair Emp. and Housing (DFEH), *Voluntary Agreement between Airbnb and DFEH* (Apr. 19, 2017), https://www.dfeh.ca.gov/files/2017/04/04-19-17-Airbnb-DFEH-Agreement-Signed-DFEH-1-1.pdf, at 16.

169. *Id.* at 17.

170. We gratefully acknowledge Anna Lauren Hoffmann for bringing this example to our attention.

171. Carson Griffith, *On a Phone App Called Tinder, Looks Are Everything*, N.Y. TIMES (Apr. 24, 2013), http://www.nytimes.com/2013/04/25/fashion/on-a-phone-app-called-tinder-looks-are-everything.html.

men, women, or both.[172] As a result, transgender users were often matched with other users who had not realized they had indicated interest in a transgender user. Some users responded to this information negatively, and as a result, reported transgender users (potentially leading to the suspension of their accounts) or subjected them to offensive and abusive language using the app's messaging feature.[173] In order to prevent such abuse, some transgender users took extra steps to make their gender identity as visible as possible in their profile pictures (e.g., by displaying a tote bag reading "PROUD TO BE TRANS");[174] others refrained from using the platform altogether.[175]

The Tinder case demonstrates how design features interact in complicated ways and how addressing bias effectively requires a multi-pronged approach. Here, a limitation on what users reveal about themselves[176] and on how users find one another[177] resulted in abusive reporting[178] that ultimately affected the composition of the community.[179] In response, Tinder eventually made design changes to address this issue, expanding selectable gender options to a list of over 35 suggestions, plus a free-text field, and giving all users the option of whether they want their gender displayed on their profile[180]; however, users cannot yet filter their matches according to these options.[181] In addition, Tinder has engaged in messaging to reframe the norms of the community around diversity and inclusivity (including a campaign around the hashtag #AllTypesAllSwipes),[182] has conducted training for its staff and allocated

172. Megan Rose Dickey, *Tinder Finally Adds Options for Trans and Gender Non-Conforming People*, TECHCRUNCH (Nov. 15, 2016), https://techcrunch.com/2016/11/15/tinder-finally-adds-options-for-trans-and-gender-non-conforming-people/.

173. Madison Malone Kircher, *Transgender People are Reportedly Being Banned from Tinder*, BUSINESS INSIDER (Jun. 3, 2015), http://www.businessinsider.com/transgender-tinder-users-reported-and-banned-2015-6; Addison Rose Vincent, *Does Tinder Have a Transphobia Problem?*, HUFFINGTON POST (Mar. 25, 2016), http://www.huffingtonpost.com/addison-rose-vincent/does-tinder-have-a-transp_b_9528554.html.

174. Vincent, *supra* note 173.

175. Kircher, *supra* note 173.

176. *See supra* Part II.B.3.

177. *See supra* Part II.B.2.

178. *See supra* Part II.C.1.

179. *See supra* Part II.A.2.

180. *See supra* Part II.B.3.

181. Sophie Kleeman, *Tinder Introduces More Inclusive Gender Options*, GIZMODO (Nov. 15, 2016), http://gizmodo.com/tinder-introduces-more-inclusive-gender-options-1788992315.

182. *See supra* Part II.A.3. *See also Introducing More Genders on Tinder*, TINDER BLOG: BEHIND THE SCENES (Nov. 15, 2016), http://blog.gotinder.com/genders/.

additional resources to its support team, and has been working in consultation with gender non-conforming users and representatives from GLAAD.[183]

A naive attempt to addressing bias on platforms—say, one that focuses on a single strategy—might not acknowledge how users' biases, even if thwarted by one feature, can readily migrate to another feature of the platform, and can even lead to abusive encounters.[184] A coherent approach must acknowledge complexities and interactions among platform features, and consider their normative dimensions, which we discuss below.

III. CONCLUSION: ETHICAL DIMENSIONS OF PLATFORM DESIGN

As we have noted, platforms face emergent and uncertain legal obligations in the face of their users' discriminatory behaviors. But even absent legal requirements, platforms may feel an ethical responsibility or public pressure to confront bias exhibited in user-to-user interactions; few platforms want to condone discrimination or develop reputations as bastions of unfair treatment. At the same time, platforms might hesitate to interfere in the business of consenting users, or to identify which of their users seem to exhibit prejudice or bias. Even attempting to minimize the degree to which implicit bias might affect users' decisions raises several normative questions, for which there may be no easy or obvious answers.

First, in the absence of laws that proscribe or prescribe certain behavior, platforms might question whether they possess—or have been granted—the moral authority to decide which types of user preferences are acceptable and which are objectionable, even if they make such decisions unintentionally in developing their products and services. Answering such questions explicitly will require normative principles that can help distinguish cases in which platforms would be wrong to infringe on users' personal autonomy from those in which platforms can override users' preferences in the interest of combating discrimination. In matters of employment, housing, and credit, platforms might feel at ease looking to discrimination law as a source of moral authority and practical guidance in deciding how to regulate the way users can treat one another. In commerce more generally and in more intimate affairs, platforms will have less obvious places to look. In the case of online commerce, platforms might

183. *See supra* Part II.A.1.

184. *See supra* Part II.B.3 (discussing the withholding of user profile photos); *see also infra* note 158 and accompanying text (discussing the relationship between withholding passenger information and low ratings on rideshare sites).

default to a position that leaves users with considerable freedom to contract as they please, even though certain users' preferences might rest on prejudicial or biased beliefs. Indeed, the very point of many platforms' business models is to allow users to choose with whom to transact. Practically, platforms like Craigslist might attempt to compel sellers to accept all comers, but no intervention could force potential buyers to transact with particular sellers. At best, Craigslist might steer potential buyers to a diverse set of sellers; it cannot command potential buyers to do business with these sellers.

Online dating presents an even more charged situation. These platforms are expressly in the business of catering to the preferences of their users, even though they cannot avoid influencing these preferences through their recommendations and other design choices.[185] How platforms should go about influencing these tastes is controversial, to say the least: which predilections should they cultivate and which should they challenge? In particular, should platforms attempt to counteract the tendency toward assortative mating and the preference to date within one's own racial group? Platforms may hesitate to publicly interfere in decisions that users perceive as deeply personal and intimate, preferring, instead, to present themselves as vehicles for satisfying one's predetermined romantic or erotic desires. Moreover, accommodating users' preferences may serve positive ends by shielding people from experiences of prejudice and facilitating efficient matches, as Emens argues: "people's explicit articulation of their dating preferences as to race, (dis)ability, and sex may be efficient for—or even, in some cases, appreciated by—prospective mates (and non-mates). Gays and lesbians, for example, have long understood the utility of creating distinctive spaces for gay socializing; even in the absence of a need to avoid detection or violence, queer-only spaces save time and energy, not to mention needless rejection."[186] In fact, platforms that purposefully limit the efficiency of searches for sexual minorities—by, for example, refusing to provide tools to filter by sexual orientation—may end up discriminating against these populations.[187]

185. *See, e.g.,* Christian Rudder, *We Experiment On Human Beings!*, THE OKCUPID BLOG (Jul. 28, 2014), https://theblog.okcupid.com/we-experiment-on-human-beings-5dd9fe280cd5 (describing how experiments with the information available to Okcupid users affected user interactions).

186. Emens, *supra* note 83, at 1353.

187. *See, e.g.*, *eHarmony, Inc. Settles Class Action Lawsuit over Same-Sex Matching*, BUSINESS WIRE (Jan. 26, 2010), http://www.businesswire.com/news/home/20100126007340/en/eHarmony-Settles-Class-Action-Lawsuit-Same-Sex-Matching (describing the settlement of a lawsuit asserting that the platform had violated state civil rights law by failing to allow users to search for same-sex partners on the site).

Second, platforms might wonder whether they should attempt to limit how easily users can exercise their biased preferences if the platforms cannot prevent users from *holding* such preferences. Specifically, platforms might find that attempts to address bias do not eliminate or diminish bias, but simply push it to other parts of users' interactions where platforms exercise less control or where the bias is less obvious. Ge et al., for example, argue that transportation platforms that deny drivers information about would-be passengers may limit the degree to which drivers can discriminate between requesters, but that prejudiced or biased drivers might nevertheless give certain passengers low ratings, hurting these passengers' abilities to attract even unbiased drivers in the future.[188] In this case, platforms committed to combatting discrimination might allow users to be biased in their choice of a counterparty, if only to ensure that parties unfairly rejected at the outset do not face even greater penalties later in the process or more severe challenges on the platform in the future. Forcing encounters or interactions between users that one or both parties would prefer to avoid may have unintended effects if either party can punish the other in parts of the process over which platforms maintain less effective control.

Finally, because there are few more serious accusations than bias, platforms run considerable risk when they set out to identify bias in their users' behavior. Cultural and political norms are such that almost no one will readily admit to overt prejudice or intentional discrimination, and most will resist claims that their decisions might have been swayed by implicit bias. And yet platforms have begun to explore many ways to uncover just how much their users discriminate against one another. First, they have begun to track differences in users' experience by race, for example, on the belief that such differences must reflect unfair treatment on the platform. Second, some have begun to develop more sophisticated methods to establish the degree to which other users' biases *cause* these differences. And third, some have even committed to using related methods to identify the *specific* users who exhibit these biases. Each can raise very different concerns. In the first case, simply ascribing differences in users' experiences on the platform to bias treats any difference as necessarily suspect and may foster an environment where bias serves as the presumed explanation for any adverse outcome. The corresponding interventions would not be able to target the source of bias and would justify changes to the platform that equalize the experience of users from different racial groups, for example, even if these interventions impose costs and unwelcome changes on others. Platforms that attempt to determine whether

188. Yanbo Ge et al., *supra* note 158, at 20.

bias actually accounts for these differences will fare much better, especially if they do not attempt to pin bias on specific users. Executing such studies will be challenging, however, especially with observational data alone, as platforms rarely have users that resemble each other on every relevant dimension but race, for example, which would be necessary to establish that race explains the difference in outcome. But the dangers are greatest when platforms aim to identify when *specific* users are behaving in a prejudicial or biased manner. Platforms could easily find themselves building models to estimate the extent to which racial bias, for example, influences users' decisions. Discounting, penalizing, or expelling users from the platform on the basis of these inferences, whether in secret or in public, could be highly problematic. Public accusations of bias are very serious, especially should these prove incorrect, but so too are unexplained actions on the part of platforms, driven by suspicions of bias, that shape people's life chances and everyday experience of the world.

As more of the exchanges that comprise daily life—from finding work to finding a date, getting a ride to getting a loan—move online, platforms cannot help but wield great influence over how their users interact. In scaffolding these exchanges, they have no choice but to interact with the biases users bring to the table. Platforms' dominance in so many domains of daily life puts them in a unique position in which they have power to perpetuate, exacerbate, or alleviate its effects in society. As we established in this Article, the levers at platforms' disposal are numerous, and may be mutually reinforcing. Our goal was to highlight the primacy of policy and design in how bias plays out on platforms, to provide a conceptual framework to identify the strategies available to them, and to draw attention to the legal and ethical considerations that adoption of different strategies might entail. Determining the efficacy of these interventions will require further empirical research, and these findings will help platforms to ascertain the most effective solutions for alleviating discrimination.