Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?

Kenneth Holstein
Carnegie Mellon University
Pittsburgh, PA
kjholste@cs.cmu.edu

Jennifer Wortman Vaughan
Microsoft Research
New York, NY
jenn@microsoft.com

Hal Daumé III
Microsoft Research &
University of Maryland
New York, NY
me@hal3.name

Miroslav Dudík
Microsoft Research
New York, NY
mdudik@microsoft.com

Hanna Wallach
Microsoft Research
New York, NY
wallach@microsoft.com

ABSTRACT

The potential for machine learning (ML) systems to amplify social inequities and unfairness is receiving increasing popular and academic attention. A surge of recent work has focused on the development of algorithmic tools to assess and mitigate such unfairness. If these tools are to have a positive impact on industry practice, however, it is crucial that their design be informed by an understanding of real-world needs. Through 35 semi-structured interviews and an anonymous survey of 267 ML practitioners, we conduct the first systematic investigation of commercial product teams' challenges and needs for support in developing fairer ML systems. We identify areas of alignment and disconnect between the challenges faced by teams in practice and the solutions proposed in the fair ML research literature. Based on these findings, we highlight directions for future ML and HCI research that will better address practitioners' needs.

CCS CONCEPTS

• **Human-centered computing**; • **Social and professional topics → Socio-technical systems**; • **Computing methodologies → Machine learning**;

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI 2019, May 4–9, 2019, Glasgow, Scotland UK
© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300830

KEYWORDS

algorithmic bias, fair machine learning, product teams, need-finding, empirical study, UX of machine learning

ACM Reference Format:
Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miroslav Dudík, and Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3290605.3300830

1 INTRODUCTION

Machine learning (ML) systems increasingly influence every facet of our lives, including the quality of healthcare and education we receive [18, 34, 35, 53], which news or social media posts we see [4, 20, 91], who receives a job [50, 89], who is released from jail [8, 25], and who is subjected to increased policing [70, 78, 106]. With this growth, the potential of ML to amplify social inequities has received growing attention across several research communities, as well as in the popular press. It is now commonplace to see reports in mainstream media of systemic unfair behaviors observed in widely used ML systems—for example, an automated hiring system that is more likely to recommend hires from certain racial, gender, or age groups [43, 107], or a search engine that amplifies negative stereotypes by showing arrest-record ads in response to queries for names predominantly given to African American babies, but not for other names [12, 84].

Substantial effort in the rapidly growing research literature on fairness in ML has centered on the development of statistical definitions of fairness [25, 32, 49, 83] and algorithmic methods to assess and mitigate undesirable biases in relation to these definitions [3, 49, 68]. As the field matures, integrated toolkits are being developed with the aim of making these methods more widely accessible and usable (e.g., [7, 31, 38, 45, 76, 77]). While some fair ML tools are already

RIGHTSLINK

being prototyped with practitioners, their initial design often appears to be driven more by the availability of algorithmic methods than by real-world needs (cf. [52, 111]). If such tools are to have a positive and meaningful impact on industry practice, however, it is crucial that their design be informed by an understanding of practitioners' actual challenges and needs for support in developing fairer ML systems [111].

We investigate the challenges faced by commercial ML product teams—whose products affect the lives of millions of users [21, 98]—in monitoring for unfairness and taking appropriate action [98, 106]. Through semi-structured interviews with 35 practitioners, across 25 ML product teams from 10 major companies, we investigate teams' existing practices and challenges around fairness in ML, as well as their needs for additional support. To better understand the prevalence and generality of the key themes surfaced in our interviews, we then conduct an anonymous survey of 267 industry ML practitioners, across a broader range of contexts. To our knowledge, this is the first systematic investigation of industry ML practitioners' challenges and needs around fairness.

Through our investigation, we identify a range of real-world needs that have been neglected in the literature so far, as well as several areas of alignment. For example, while the fair ML literature has largely focused on "de-biasing" methods and viewed the training data as fixed [23, 59], most of our interviewees report that their teams consider data collection, rather than model development, as the most important place to intervene. Participants also often report struggling to apply existing auditing and de-biasing methods in their contexts. For instance, whereas previously proposed methods typically require access to sensitive demographics at an individual level, such information is frequently available only at coarser levels. The fair ML literature has tended to focus on domains such as recidivism prediction, automated hiring, and face recognition, where fairness can be understood, at least partially, in terms of well-defined quantitative metrics [21, 26, 65], whereas teams working on applications involving richer interactions between the user and the system (e.g., chatbots, web search, and adaptive tutoring) brought up needs for more holistic, system-level fairness auditing methods. Interviewees also stressed the importance of explicitly considering biases and "blind spots" that may be present in the humans embedded throughout the ML development pipeline, such as crowdworkers or user-study participants. Such concerns also extended to product teams' own blind spots. For instance, teams often struggled to anticipate which subpopulations and forms of unfairness they need to consider for specific kinds of ML applications.

Based on these findings, we highlight opportunities for the ML and HCI research communities to have a greater impact on industry and on the fairness of ML systems in practice.

2 BACKGROUND AND RELATED WORK

The design, prototyping, and maintenance of machine learning systems raises many unique challenges [30, 67, 85, 95] not commonly faced with other kinds of intelligent systems or computing systems more broadly [37, 71, 72]. The budding area of "UX for ML" has begun to explore new forms of prototyping for ML systems, to provide earlier insights into the complex, interacting UX impacts of particular dataset, modeling, and system-design choices [30, 44, 51, 109]. In addition, a growing body of research focuses on the design and development of programmer tools that can better support developers in debugging and effectively monitoring the predictive performance of complex ML systems [5, 24, 63, 67, 85].

In parallel, the potential for undesirable biases in ML systems to exacerbate existing social inequities—or even generate new ones—has received considerable attention across a range of academic disciplines, from ML to HCI to public policy, law, and ethics [11, 98, 106]. Specialized research conferences and initiatives are forming with a focus on biases and unfairness in data-driven algorithmic systems, such as the Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) [2], the nascent FAT* conference [1], AI Now [55], and the Partnership on AI [87]. Significant effort in the fair ML community has focused on the development of statistical definitions of fairness [14, 25, 32, 83] and algorithmic methods to assess and mitigate biases in relation to these definitions [3, 17, 49, 68]. In contrast, the HCI community has studied unfairness in ML systems through political, social, and psychological lenses, among others (e.g., [15, 47, 94, 108]). For example, HCI researchers have empirically studied users' expectations and perceptions related to fairness in algorithmic systems, finding that these do not always align with existing statistical definitions [16, 71, 72, 108]. Other work has focused on auditing widely-used ML products from the outside [8, 21, 29, 62], and often concluded with high-level calls to action aimed at those responsible for developing and maintaining these systems or for regulating their use [36, 93, 97]. Lastly, Crawford [28], Springer et al. [98], and others have highlighted an urgent need for internal processes and tools to support companies in developing fairer systems in the first place.

Despite this widespread attention to biases and unfairness in ML, to the best of our knowledge, only one prior study, by Veale et al. [106], has investigated actual ML practitioners' challenges and needs for support in creating fairer ML systems. Veale et al. conducted exploratory interviews with public-sector ML practitioners working across a range of high-stakes contexts, such as predictive policing [70, 78] and child mistreatment detection [26], to understand the challenges faced by practitioners in aligning the behavior of ML

RIGHTSLINK

systems with public values. Through these interviews, the authors uncovered several disconnects between the real-world challenges that arise in public-sector ML practice compared with those commonly presumed in the fair ML literature.

In the same spirit as Veale et al. [106], this work investigates ML practitioners' needs for support, with the aim of identifying fruitful opportunities for future research [99]. Whereas Veale et al. studied algorithm-assisted decision makers in high-stakes public-sector contexts—who are often experienced in thinking about fairness, yet relatively new to working with ML systems—we study industry ML practitioners, who tend to be experienced in developing ML systems, but relatively new to thinking about fairness. Supporting industry ML practitioners can also be viewed as a critical step towards fairer algorithm-assisted decision making downstream, including in the public sector, where systems are often built on top of ML products and APIs developed in industry [21, 41, 88, 101]. Unlike the public-sector practitioners studied by Veale et al., the industry practitioners studied here work on a much broader range of applications, such as image captioning, web search, chatbots, speech recognition, and personalized retail. In many of these applications, ethical considerations may be even less clear cut than in high-stakes public-sector contexts. Moreover, motivations and organizational priorities can differ considerably in industry contexts.

3 METHODS

To better understand product teams' needs for support in developing fairer ML systems, we conducted a series of semi-structured, one-on-one interviews with a total of 35 practitioners, across 25 ML product teams from 10 major companies. To investigate the prevalence and generality of the key themes that emerged in these interviews, we then conducted an anonymous survey with a broader sample of 267 industry ML practitioners. For both the interviews and the survey, "practitioners" were defined broadly as those who work in any role on a team that develops products or services involving ML. The study went through an ethical review and was IRB-approved. Semi-structured interview protocols and survey questions are provided in the supplementary materials.

Interview Study

Our interview study proceeded in two rounds. First, to get a broad sense of challenges and needs, we conducted six formative interviews. Building on emerging themes, we then conducted more in-depth interviews with a larger sample.

The first six interviews were conducted with product managers (PMs) across different technology areas. Each interview lasted 30 minutes and was conducted by teleconference because PMs were distributed across multiple countries. Each PM was first asked to describe the products their team is responsible for, who the customers of these products are, and how their team is structured. Interviewees were then asked whether fairness is something their team regularly discusses or incorporates into their workflow. The meaning of "fairness" was intentionally left open as we were interested in hearing PMs' notions about what it might mean for their products to be fair. However, if a PM requested clarification at any point, we provided a broad definition: *"Any case where AI/ML systems perform differently for different groups in ways that may be considered undesirable."* PMs were then asked whether their team or customers had ever encountered issues relating to fairness in their products. If so, PMs were asked for concrete examples, and were asked high-level follow-up questions about these experiences. Otherwise, they were asked whether they thought such issues might exist undetected, and whether they had seen *"other surprising or unexpected issues arise"* (cf. [106]). These follow-ups sometimes led PMs to realize they actually did have relevant experiences to share.

Our second, main round of interviews built on themes that emerged during the initial round. We conducted more detailed, semi-structured interviews to investigate teams' current practices, challenges, and needs in greater depth. Interviewees included 29 practitioners, across 19 ML product teams from 10 major technology companies. As shown in Table 1, we interviewed practitioners across a range of technology areas and team roles. Whenever possible, we tried to interview people in different roles on the same team to hear (potentially) different perspectives. Interviewees were recruited using snowball sampling. We searched for news articles related to ML biases and unfairness and contacted members of teams whose products had previously received relevant media coverage (i.e., news articles about unfair behavior observed in these products). In addition, we emailed direct contacts across over 30 major companies. In both cases, we asked contacts to share our interview invitation with any colleagues working on ML products (in any role) at their own company or others. At the end of each interview, interviewees were again encouraged to share any relevant contacts.

Although prospective interviewees were often eager to participate and recruit colleagues, we encountered several challenges resembling those discussed by Veale et al. [106]. For instance, given a recent trend of negative media coverage calling out algorithmic biases and unfairness in widely-used ML systems (e.g., [12, 101, 107, 112]), our contacts often expressed strong fears that their team or company's identity might leak to the popular press, harming their reputation. Some contacts revealed a general distrust of researchers, citing cases where researchers have benefited by publicly critiquing companies' products from the outside instead of engaging to help them improve their products. Finally, some contacts worried that, in diving into the details of their teams'

RIGHTSLINK

Table 1: Interview demographics: Interviewees' self-reported technology areas and team roles. Where multiple people were interviewed from the same product team, interviewee identifiers are grouped in square brackets.

Technology Area	Roles of Interviewees	Interviewee IDs
Adaptive Tutoring & Mentoring	Chief Data Scientist, CTO, Data Scientist, Research Scientist	R10, [R13, R14], R30
Chatbots	CEO, Product Mgr., UX Researcher	[R17, R18], R35
Vision & Multimodal Sensing	CTO, ML Engineer, Product Mgr., Software Engineer	[R2, R3, R4], R6, R7, R9, R26
General-purpose ML (e.g., APIs)	Chief Architect, Director of ML, Product Mgr.	R25, R32, R34
NLP (e.g., Speech, Translation)	Data Mgr., Data Collector, Domain Expert, ML Engineer, Product Mgr., Research Software Eng., Technical Mgr., UX Designer	R1, [R15, R16, R19, R20, R21, R22], R24, [R27, R29], R28, R31
Recommender Systems	Chief Data Scientist, Data Scientist, Head of Diversity Analytics	R8, R12, R23, R33
Web Search	Product Mgr.	R5, R11

prior experiences, they might inadvertently reveal trade secrets. For these reasons, contacts often declined to be interviewed. To allay some of these concerns, we assured contacts that the goal of these interviews was to help us learn about teams' current practices, challenges, and needs around fair ML in general, and that findings would not be linked to specific individuals, teams, or companies. We also asked for interviewees' advance permission to audio record the interviews, noting that these recordings would not be shared outside of the research team. Furthermore, we noted that these recordings would be destroyed following transcription, and that the resulting transcriptions would be de-identified. Finally, we assured interviewees that we would allow them to review any (de-identified) direct quotes before including them in any research publications. All 29 interviewees in the main round of interviews consented to be audio recorded.

Each interview in the main round lasted between 40 and 60 minutes. In each one, interviewees were first reminded of the overall purpose of the research, and were then asked a series of questions about fairness at each stage in their team's ML development pipeline—from collecting data to designing datasets (e.g., curating training and test sets) and developing an ML product to assessing and potentially mitigating fairness issues in that product. For each of these stages, interviewees were asked a broad opening question about critical episodes they had encountered (e.g., *"Can you recall times you or your team have discovered fairness issues in your products?"*), and a series of follow-up questions (where applicable and not previously covered). While a few follow-up questions were specific to particular stages of the pipeline, a core sequence of four follow-ups was used across all stages. First, interviewees were asked to walk through how their team navigated the episode (e.g., *"When you decided there were issues that needed to be addressed... how did your team decide what course of action to take to address them?"*). Next, interviewees were instructed to imagine they could return to these critical episodes, but this time with access to a magical oracle of which they could ask any question to help them in the moment [52, 63]. We asked the question in this way to encourage interviewees to speak freely about their current challenges and needs without feeling constrained to those for which they believed a solution was currently available [52, 63]. After interviewees generated questions for the oracle, they were then asked, for each question, whether and how their team currently goes about trying to answer this question. This follow-up often led interviewees to reflect on gaps between their current and ideal practices. Finally, interviewees were asked whether they saw any other areas for improvement, or opportunities for support, related to the relevant stage of their team's ML development pipeline.

To analyze the interview data, we worked through transcriptions of approximately 25 hours of audio to synthesize findings using standard methodology from contextual design: interpretation sessions and affinity diagramming [48, 54]. Specifically, we employed bottom-up affinity diagramming (using MURAL [82]) to iteratively generate codes for various interviewee utterances and then grouped these codes into successively higher-level themes concerning current practices, challenges, and needs for support. Key themes are presented in detail below, under *Results and Discussion*.

Survey

To validate our findings on a broader population, we conducted an anonymous online survey using Qualtrics [90]. Respondents were recruited using snowball sampling. Specifically, we emailed the survey to direct contacts at over 40 companies and invited them to pass the survey on to colleagues (within or outside of their companies) who are part of a team working on ML products (in any role). We also announced the survey on social media (e.g., Twitter) and online communities related to ML and AI, including Kaggle forums and special interest groups on LinkedIn, Facebook, Reddit, and Slack.

RIGHTSLINK

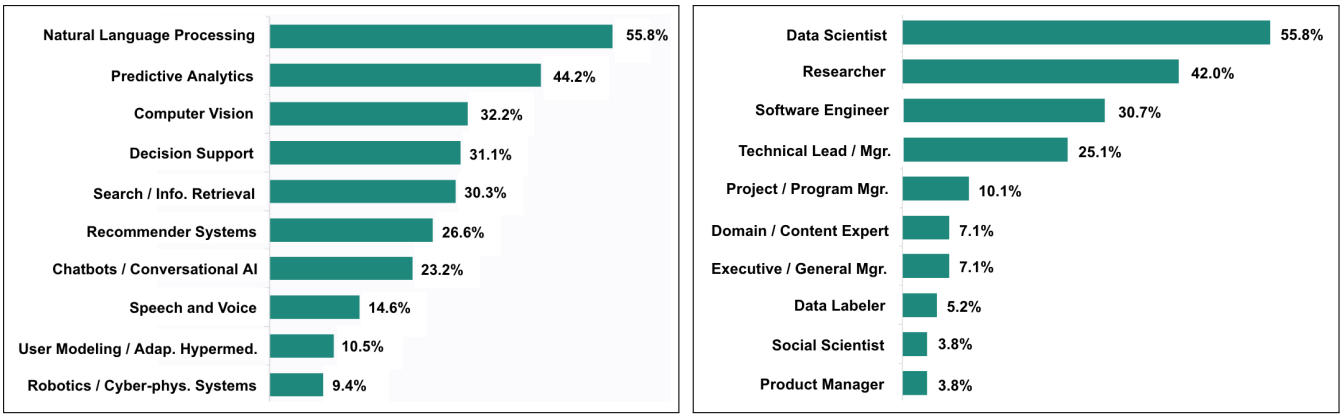

Figure 1: Survey demographics: the top 10 self-reported technology areas (left) and team roles (right).

We structured the survey to act as a quantitative supplement to the interviews. As such, the high-level structure of the survey mirrored that of the main round of interviews. Based on our findings from the interviews, we developed survey questions to investigate the prevalence and generality of emerging themes. First, we asked a set of demographic questions to understand our respondents' backgrounds, including their technology area(s) and team role(s). In a branching sequence of survey sections, respondents were then asked about their team's current practices, challenges, and needs for support around fairness, with each section pertaining to one stage of their team's ML development pipeline. For each question, closed-ended response options were provided based on themes that emerged from the interviews through affinity diagramming, in addition to free-response options that allowed respondents to elaborate on their responses.

A total of 287 people started the survey. However, not all respondents completed it, so we analyze only the 267 respondents who completed at least one section beyond demographics. The top 10 self-reported technology areas and team roles are shown in Figure 1. In each case, respondents were able to select multiple options. Additional survey demographics are available in the supplementary materials.

4 RESULTS AND DISCUSSION

Although participants spanned a diverse range of companies, team roles, and application areas, we observed many commonalities. In the following, we discuss teams' current challenges and needs around fairness, organized by top-level themes that emerged through affinity diagramming. These include needs for support in fairness-aware data collection and curation, overcoming teams' blind spots, implementing more proactive fairness auditing processes, auditing complex ML systems, deciding how to address particular instances of unfairness, and addressing biases in the humans embedded throughout the ML development pipeline. Within each of these top-level themes, we present selected sub-themes to highlight research and design opportunities that have received little attention in the fair ML literature thus far.

We supplement our findings from the interviews with corresponding survey responses. Interviewees are identified with an "R," and survey responses are accompanied with percentages. Some survey questions were completed by a subset of respondents because the survey used branching logic, with follow-up sections appearing only as applicable (e.g., respondents were only asked questions about addressing fairness issues if they reported that their team had previously detected such issues in their products). In such cases, question-specific response rates are provided in addition to the percentage of respondents who were asked the question. A graphical summary of selected survey responses is provided in the supplementary materials. To illustrate general themes, we share direct quotes in cases where we received explicit permission from interviewees, in accordance with our IRB approval and consent form. Therefore, although the themes we present are drawn from a wide range of application domains, the domains represented by direct quotes may be narrower.

By recruiting interviewees using snowball sampling and specifically targeting members of teams whose products had previously received media coverage related to ML biases and unfairness, we may have sampled practitioners who are unusually motivated to address fairness issues in their products. Indeed, many interviewees reported having invested significant time and effort into trying to improve fairness in their teams' products, even when they felt unsupported in these initiatives by their team or company leadership. As such, many of the findings presented below may be taken to represent barriers that industry ML practitioners run up against in improving fairness even when they are motivated to do so.

While much of our discussion focuses on needs for tools, we stress that biases and unfairness are fundamentally sociotechnical problems, and that technically focused research

RIGHTSLINK

efforts must go hand-in-hand with efforts to improve organizational processes, policies, and education [96, 102].

Fairness-aware Data Collection

While media coverage of biases and unfairness in ML systems often uses a "bias in, bias out" framing, emphasizing the central role of dataset quality [42, 107], the fair ML research literature has overwhelmingly focused on the development of algorithmic methods to mitigate biases, viewing the dataset as fixed [23, 59]. However, many of our interviewees reported that their teams typically look to their training datasets, not their ML models, as the most important place to intervene to improve fairness in their products. Out of the 65% of survey respondents who reported that their teams have some control over data collection and curation, a majority (58%) reported that they currently consider fairness at these stages of their ML development pipeline. Furthermore, out of the 21% of respondents whose teams had previously tried to address fairness issues found in their products, the most commonly attempted strategy (73%) was "collecting more training data."

"It's almost like the wild west." Most interviewees reported that their teams do not currently have processes in place to support the collection and curation of balanced or representative datasets. A software engineer (R7) described their team's current data collection practices as *"almost like the wild west"* and a data scientist (R10) noted that *"there isn't really a thought process surrounding... 'Should [our team] ingest this data in?' [...] If it is available to us, we ingest it."* An ML engineer (R19) emphasized the value of supporting more effective communication (e.g., around sampling practices and useful metadata to include) between those responsible for decisions about data collection and curation, and those responsible for developing ML models. On a 5-point Likert scale from "Not at all" to "Extremely" useful, 52% of the respondents who were asked the question (79%) indicated that tools to facilitate communication between model developers and data collectors would be "Very" or "Extremely" useful.

Interviewees also shared a range of needs for tools and processes that can actively guide data collection as it occurs, to support fairness in downstream ML models. For example, in response to the "oracle" interview question, an ML engineer (R19) working on automated essay scoring noted:

> *"To score African American students fairly, they need examples of African American students scoring highly. But in the data [the data collection team] collect[s], this is very rare. So, what is the right way to sample [high-scorers] without having to score all the essays? [...] So [we need] some kind of way... to indicate [which schools] to collect from [...] or what to bother spending the extra money to score."*

A majority (60%) of survey respondents, out of the 25% who indicated their team has some control over data collection processes, indicated having such active guidance would be at least "Very" useful. By contrast, in contexts where training data is collected via regular, "in-the-wild" users of a product, challenges can arise when specific user populations are less engaged with the product. To overcome such challenges, a technical manager working on speech recognition (R31) suggested it would help to know more effective strategies to incentivize"in-the-wild" product usage within specific populations, such as targeted gamification techniques [58].

Scaffolding fairness-aware test set design. Several interviewees stressed the central importance of careful test set design to detecting potential fairness issues. For example, R4's team would discuss possible issues to watch out for, and then *"try to design the test set to capture those notions, if we can."* R4 credited having *"the test set be well constructed and not biased"* for the discovery of gender biases in their image captioner (e.g., images of female doctors were often mislabeled as nurses) which they ultimately traced to imbalances in their training data. An ML engineer working on gesture recognition (R6) noted it would be helpful to have better tools to scaffold the design of test sets, to make it easier to

> *"assign tags to data points based on certain characteristics that you want to make sure are fair [...and check] that first of all, each of those tags has a significant number of samples [and] look at the measurements across each slice and [check] if there's a bias issue."*

A majority (66%) of survey respondents, out of the 70% who were asked this question, indicated that such tools to scaffold the design of test sets would be "Very" or "Extremely" useful.

Implications. Although the fair ML literature has focused heavily on algorithmic "de-biasing" methods that assume a fixed training dataset, practitioners in many industry contexts have some control over data collection and curation. Future research should support these practitioners in collecting and curating high quality datasets in the first place (cf. [23, 40, 59]). In contrast to the fair ML literature, HCI research on ML developer tools has often focused on the design of user interfaces to scaffold data collection and curation processes (e.g., [22, 24, 39, 66]). However, this work has tended to focus on improving ML models' overall predictive accuracy, rather than fairness. A promising direction for future research is to explore how such tools might be explicitly designed to support fairness and equity in downstream ML models by interactively guiding data collection, curation, and augmentation, or by supporting the design and use of test sets that can effectively surface biases and unfairness (cf. [19, 57, 110]).

RIGHTSLINK

Challenges Due to Blind Spots

Interviewees expressed many anxieties around their teams' potential "blind spots," which might stand in the way of effectively addressing fairness issues, or even thinking to monitor for some forms of unfairness in the first place.

Data collection and curation challenges due to blind spots. Most of our interviewees highlighted needs for support in identifying which subpopulations their team needs to consider when developing specific kinds of ML applications, to ensure they collect sufficient data from these subpopulations or balance across them when curating existing datasets. A technical director working on general-purpose ML tools (R32) emphasized the challenges of anticipating which subpopulations to consider, noting that this can be highly context and application dependent [46], with potential subpopulations extending well beyond those commonly discussed in the fair ML literature: *"Most of the time, people start thinking about attributes like [ethnicity and gender...]. But the biggest problem I found is that these cohorts should be defined based on the domain and problem. For example, for [automated writing evaluation] maybe it should be defined based on [...whether the person is] a native speaker."* A majority (62%) of survey respondents, out of the 80% who were asked the question, indicated that it would be at least "Very" useful to have additional support in identifying relevant subpopulations.

"You'll know if there's fairness issues if someone raises hell online." Interviewees often reported that their teams do not discover serious fairness issues until they receive customer complaints about products (or, worse, by reading negative media coverage about their products). As a software engineer working on image classification (R7) put it, *"How do you know the unknowns that you're being unfair towards? [...] You just have to put your model out there, and then you'll know if there's fairness issues if someone raises hell online."* Several interviewees expressed needs for support in detecting biases and unfairness prior to deployment, even in cases where they may not have anticipated all relevant subpopulations or all kinds of fairness issues. Despite their efforts running user studies, teams often discovered serious issues only after deploying a system in the real world (51% of survey respondents marked this statement as at least "Very" accurate).

Team biases and limitations. Several of our interviewees' teams have a practice of getting together and trying to imagine everything that could go wrong with their products, so that they can make sure to proactively monitor for those issues. A few teams even reported including fairness-focused quizzes in their interview processes, with the aim of hiring employees who would be effective at spotting biases and unfairness. R2 described much of their team's current process as *"just everyone collecting all the things that they can think of that could be offensive and testing for [them]."* But as R4 emphasized, *"no one person on the team [has expertise] in all types of bias [...] especially when you take into account different cultures."* Interviewees noted that it would be helpful to somehow pool knowledge of potential fairness issues in specific application domains across teams with different backgrounds, who have complementary knowledge and blind spots. A majority (67%) of survey respondents, out of the 71% who were asked the question, indicated that tools to support such knowledge pooling would be at least "Very" useful.

A few interviewees also shared experiences in which efforts to obtain additional training data to address a fairness issue were hampered by their teams' blind spots. A developer working on image captioning (R4) recalled cases where many customers had complained that a globally deployed system performed well for celebrities from some countries, but routinely misidentified major celebrities from others:

> *"It sounds easy to just say like, 'Oh, just add some more images in there,' but [...] there's no person on the team that actually knows what all of [these celebrities] look like, for real [...] if I noticed that there's some celebrity from Taiwan that doesn't have enough images in there, I actually don't know what they look like to go and fix that. [...] But, Beyoncé, I know what she looks like."*

Implications. Assessing and mitigating unfairness in ML systems can depend on nuanced cultural and domain knowledge that no single product team is likely to have. In cases where ML products are deployed globally, efforts to recruit more diverse teams may be helpful yet insufficient. A fruitful area for future research is the design of processes and tools to support effective sharing and re-use of such knowledge across team or company boundaries—for example via shared test sets (cf. [110]) or case studies from other teams who have worked in specific application domains. Another promising direction may be to support teams in the ad-hoc recruitment of diverse, team-external "experts" for particular tasks [103].

Needs for More Proactive Auditing Processes

Detecting potential fairness issues presents many unique auditing challenges. Interviewees often described their teams' current fairness auditing processes as reactive—with efforts tightly focused around specific customer complaints—in contrast to their teams' proactive approaches for detecting potential security risks. As a PM for web search (R11) put it,

> *"It's a little bit of a... manual search to say, 'hey, we think this has a bias, let's go take a look and see if it does,' which I don't know is the right approach [...] because there are a lot of strange ones that you wouldn't expect [...] that we just accidentally stumbled upon."*

Out of the 63% of survey respondents who were asked the question, almost half (49%) reported that their team had

RIGHTSLINK

previously found potential fairness issues in their products. Of the 51% who indicated that their team had not found any issues, most (80%) suspected that there might be issues they had not yet detected, with a majority (55%) reporting they believe undetected issues "Probably" or "Definitely" exist.

Furthermore, most interviewees noted that they are not generally rewarded within their organizations for their efforts around fairness. On a 5-point Likert scale from "Not at all" to "A great deal," only 21% of survey respondents reported that their team prioritizes fairness "A lot" or "A great deal," and 36% indicated "Not at all." Given that interviewees often engaged in ML fairness efforts on their own time and initiative, several of the needs they highlighted focused on ways to reduce the amount of manual labor required to effectively monitor for unfairness, or ways to persuade other team members that a given fairness issue actually exists (e.g., by showing them quantitative metrics) and should be addressed.

Interviewees revealed a range of needs for support in implementing proactive auditing processes. These include domain-specific auditing processes, including metrics and tools; methods to effectively monitor for unfairness when individual-level demographics are unavailable; more scalable and comprehensive auditing processes; and ways to determine if a specific issue is part of a systemic problem.

Needs for (domain-specific) standard auditing processes. To progress beyond *"having each team do some sort of ad hoc [testing]"* (R4), several interviewees expressed desires for greater sharing of guidelines and processes. As R30 stressed,

> *"If you're developing [a model], there is a type of checklist that you go through for accuracy and so on. But there isn't anything like that [for fairness], or at least it hasn't been disseminated. What we need is a good way of incorporating [fairness] as part of the workflow."*

Most of our interviewees' teams do not currently have fairness metrics against which they can monitor performance and progress. Only 23% and 20% of survey respondents, respectively, out of the 26% who were asked these questions, marked as at least "Very" accurate the statements *"We have [fairness] metrics / key performance indicators (KPIs)"* and "*We run automated tests [for fairness]."* Yet, as R2 noted, *"it's really hard to fix things that you can't measure."* Similarly, R1 said, *"it would be really nice to learn more about how unfair we actually are, because only then can we start tackling that."*

Several of our interviewees reported that their team had searched the fair ML literature for existing fairness metrics. However, they often failed to find metrics that readily applied to their specific application domains. For example, although much of the fair ML literature has focused on metrics for quantifying "allocative harms" (relating to the allocation of limited opportunities, resources, or information), many of the fairness issues that arise in domains such as web search, conversational AI, or image captioning are "representational harms" (e.g., perpetuating undesirable associations) [28].

Some interviewees reported that they had tried to hold meetings with other teams within their company, to learn from one another's experiences and avoid duplicated effort. However, as R2 explained, even when practitioners are working on different problems in the same application domain,

> *"It doesn't necessarily result in a best practices list. [...] We've all tried to make these ways to measure [unfairness ... but] with each problem comes nuances that make it difficult to have one general way of testing."*

Given that most interviewees had limited time and resources to devote to developing their own solutions, they often emphasized the value of resources that could help them learn from others' experiences more efficiently. For example, R21 suggested it would be ideal to have *"a nice white paper that's just like... 'Here's a summary of research people have done on fairness [specifically] in NLP models.' "* Other interviewees suggested it would be extremely helpful to have access to tools and resources that would help their team anticipate what kinds of fairness issues can arise in their specific application domain, together with domain-specific frameworks that can help them navigate any associated complexities.

Fairness auditing without access to individual-level demographics. Although most auditing methods in the fair ML literature assume access to sensitive demographics (such as gender or race) at an individual level [105], many of our interviewees reported that their teams are only able to collect such information at coarser levels, if at all (cf. [11, 64, 105]). For example, companies working with K-12 student populations in the US are typically prohibited from collecting such demographics by school or district policies and FERPA laws [86]. A majority (70%) of survey respondents, out of the 69% who were asked the question, indicated that the availability of tools to support fairness auditing without access to demographics at an individual level would be at least "Very" useful.

A few interviewees reported that their teams had attempted to use coarse-grained demographic information (e.g., region- or organization-level demographics) for fairness auditing. However, each of these teams had quickly abandoned these efforts, citing limited time and resources to spend on building their own solutions. R21 said, *"If we had more people who we could throw at this... 'Can we leverage this fuzzy data to [audit]?' that would be great [...] It's a fairly intimidating research problem I think, for us."* Other interviewees noted that, while it would be helpful to have support in efficiently using coarse-grained demographic information for fairness auditing, several challenges would remain. For example, as R14 noted, *"even when you have those data [...] you may know a bunch about the demographics of a school, but then, you*

RIGHTSLINK

know, it turns out [our product] is only used by the gifted [or remedial] students, and you may not have means [to check]."

Some interviewees shared that their teams had experimented with developing ML models to infer individual-level demographics from available proxies, so that they could then use these inferred demographics to audit their ML products. However, interviewees worried that the use of proxies may in itself introduce undesirable biases, introducing a need to audit the auditing tool. A data scientist working on automated hiring (R23) recounted a time when their team had developed such a tool, but ultimately decided against using it:

> *"We called it the SETHtimator, a sex and ethnicity estimator. [...with] one dataset, we [only] had a list of people's names and their IP addresses. So we were able to sort of cross-reference their IP addresses with a name database, and from there use a [classifier] to list a probability that someone with that name in that region would have a certain gender or ethnicity. [...] It's buggy. If there was a tool [...to] do this automatically and with a trusted data source... that would be super useful."*

Interviewees often commented that it would be ideal to be able to *"get the demographic information in the first place"* (R15). Recent work in the fair ML literature has proposed encryption mechanisms that ensure any collected individual-level demographics can only be used for auditing [64, 105]. But interviewees emphasized that, while such technical solutions are an important prerequisite, half the battle would lie in convincing stakeholders and policymakers that these mechanisms are truly secure and that the benefits outweigh the risks of a potential data leak. Anticipating such challenges, a couple of interviewees expressed interest in mechanisms that might allow local decision makers, such as healthcare professionals in hospitals, to use their "on the ground" knowledge of individual-level demographics to improve fairness in an ML system without revealing these demographics.

Needs for greater scalability and comprehensiveness. Interviewees often complained about limitations of their current auditing processes, given the enormous space of potential fairness issues. For example, a UX researcher working on chatbots (R18) highlighted the challenges of recruiting a sizable, diverse sample of user-study participants as one reason their team sometimes fails to detect fairness issues early on: *"because of just logistics... we get [8 or 10] participants at a time, and even though we recruit for a 'diverse' group... I mean, we're not representing everybody."* Similarly, R4 described their team's current user-testing practices as *"more of a spot check,"* noting *"what I would rather have is a more comprehensive... full bias scan."* Drawing parallels to existing, automated tools that scan natural language datasets for potentially sensitive terms, R1 suggested having tools that *"at least flag things that seem potentially 'unfair' would be helpful."* While several other interviewees generated similar ideas, they also often pointed out that developing scalable, automated processes would be a highly challenging research problem, given that fairness can be so context dependent. R5 emphasized that domains like image captioning or web search are particularly challenging, because fairness can depend jointly on the system's output and the user-provided input (e.g., a query).

Diagnosing systemic problems from specific issues. Interviewees also highlighted challenges in diagnosing whether specific issues (e.g., complaints from customers) are symptomatic of broader, systemic problems or just "one-offs." A product and data manager for a translation system (R1) said

> *"If an oracle was able to tell me, 'look, this is a severe problem and I can give you a hundred examples [of this problem],' [...] then it's much easier internally to get enough people to accept this and to solve it. So having a process which gives you more data points where you mess up [in this way] would be really helpful."*

A majority (62%) of survey respondents, out of the 70% who were asked this question, indicated that tools to help find other instances of an issue would be at least "Very" useful.

Implications. Given that fairness can be highly context and application dependent [46, 71], there is an urgent need for domain-specific educational resources, metrics, processes, and tools to help practitioners navigate the unique challenges that can arise in their specific application domains. Such resources might include, for example, accessible summaries of state-of-the-art research around fairness in machine translation, or case studies of challenges faced by teams working on particular kinds of recommender systems (e.g., [27]).

Future research should also explore processes and tools to support fairness auditing with access to only coarse-grained demographic information (e.g., neighborhood- or school-level demographics). Recent work [64, 105] has begun to explore the design of encryption mechanisms that ensure any collected individual-level demographics can only be used for auditing. However, our interviewees noted that in certain sensitive contexts, stakeholders may be unwilling to reveal individual-level demographics, even with such mechanisms.

Although interviewees highlighted needs for more scalable and comprehensive auditing processes, they also often expressed skepticism that fairness auditing could be fully automated in their domain due to challenges in quantifying fairness. Therefore, in addition to developing domain-specific metrics and tools, a direction for future research may be to explore and evaluate human-in-the-loop approaches to fairness auditing that combine the strengths of automated methods and human inspection (cf. [24, 53, 100, 106]), perhaps aided by tools for visualization and guided exploration [24, 69].

RIGHTSLINK

Finally, at a broader, organizational level, future research should explore how teams and companies can best be motivated to adopt fairness as a central priority. Efforts to develop easily-trackable fairness metrics in particular application domains may help towards this goal in strongly metric-driven companies. Relatedly, future tools for fairness auditing should be designed to support practitioners in both (1) determining whether specific issues are instances of broader, systemic problems and (2) effectively persuading other team members that there are issues that need to be addressed.

Needs for More Holistic Auditing Methods

Much of the existing fair ML literature has focused on domains such as recidivism prediction, automated hiring, and face recognition, where fairness can be at least partially understood in terms of well-defined quantitative metrics (e.g., between-group parity in error rates or decisions [21, 26, 65]). However, interviewees working on applications involving richer, complex interactions between the user and the system—such as chatbots, automated writing evaluation, adaptive tutoring and mentoring, and web search—brought up needs for more holistic, system-level auditing methods.

Fairness as a system-level property. Many interviewees noted disconnects between the way they tend to think about fairness in their application domain and the discourse they have observed in both the popular press and academic literature. For example, a technical manager working on adaptive learning technologies (R30) noted that their team does not think about fairness in terms of monitoring individual ML models for unfairness, but instead evaluating the real-world impacts of ML systems, mentioning that *"If we think about educational interventions as analogous to medical interventions or drug trials [...] we know and [expect] a particular intervention will have different effects on different subpopulations."*

In a complex, multi-component ML system, there is not always a clean mapping between performance metrics for an individual ML model and the system's utility for users. System components may interact with one another [33, 85] and with other aspects of a system's design in ways that can be difficult to predict absent an empirical study with actual users (cf. [37]). Furthermore, it may not be straightforward to even define fair system behavior without first understanding users' expectations and beliefs about the system (cf. [71, 72]). For example, a PM working on web search (R11) shared that their team had previously experimented with ways to address a fairness issue involving image search (a search for the term "CEO" resulted predominantly in images of white men). However, through user studies, the team learned that many users were uncomfortable with the idea of the company "manipulating" search results, viewing this behavior as unethical:

> *"Users right now are seeing [image search] as 'We show you [an objective] window into [...] society,' whereas we do have a strong argument [instead] for, 'We should show you as many different types of images as possible, [to] try to get something for everyone.'"*

Needs for simulation-based approaches for complex systems. In applications involving sequences of complex interactions between the user and the system, fairness can be heavily context dependent. R17 suggested it would be valuable to have ways to prototype conversational agents more rapidly, including methods for simulating conversational trajectories (cf. [56, 113]) *"and then find[ing] ways to automate the identification of risky conversation patterns that emerge."* Similarly, a data scientist working on adaptive mentoring software (R10) suggested that because their product involves a long-term feedback loop, it would be ideal to run it on a population of "simulated mentees" (cf. [80]) to see if certain forms of personalization might be harmful with respect to equity.

Implications. In applications involving richer, complex interactions between the users and the system, assessing fairness issues via de-contextualized quantitative metrics for individual ML models may be insufficient (cf. [96]). Future research should explore new approaches to auditing and prototyping ML systems (cf. [30]), perhaps aided by simulation tools that can help developers anticipate sensitive contexts or real-world impacts of using an ML system (cf. [44, 75]).

Addressing Detected Issues

Interviewees revealed a range of challenges and needs around debugging and remediation of fairness issues once detected. These included, among others, needs for support in identifying the cheapest, most effective strategies to address particular issues; methods to estimate how much additional data to collect for particular subpopulations; processes to anticipate potential trade-offs between specific definitions of fairness and other desiderata for an ML system (not limited to predictive accuracy); and frameworks to help navigate complex ethical decisions (e.g., the fairness of fairness interventions).

Needs for support in strategy selection. Interviewees reported that their teams often struggle to isolate the causes of unexpected fairness issues, especially when working with ML models that the team consider to be "black boxes." As a result, it is often difficult for teams to decide where to focus their efforts—switching to a different model, augmenting the training data in some way, collecting more or different kinds of data, post-processing outputs, changing the objective function, or something else (cf. [23]). Of the 21% of survey respondents whose teams had previously tried to address detected issues, a majority (54%) indicated that it would be at least "Very" useful to have better support in comparing specific

RIGHTSLINK

strategies for addressing particular fairness issues. Different costs may be associated with different strategies in different contexts. For example, a developer working on image captioning (R7) noted that collecting additional data is typically a *"last resort"* option for their team, given data collection costs. In contrast, a PM working on image captioning on a different team (R2) cited data collection as their team's default starting place when trying to address a fairness issue. These interviewees and others also shared experiences where their teams had wasted significant time and resources pursuing various dead ends. As such, interviewees highlighted several needs for "fair ML debugging" processes and tools (cf. [23, 38, 45]) to support their teams in identifying the cheapest, yet most promising strategies to address particular issues. For example, R1 highlighted needs for support in *"identify[ing] the component where we mess up"* in complex, multi-component ML systems [85], and in deciding whether to focus their mitigation efforts on training data or on models. A majority (63%) of survey respondents indicated that tools to aid in these decisions (cf. [23, 45]), would be at least "Very" useful.

Avoiding unexpected side effects of fairness interventions. In addition to costs, interviewees often cited fears of unexpected side effects as a deterrent to addressing fairness issues. R4 shared prior experiences where, after making changes to datasets or models to improve fairness, their system changed in subtle, unexpected ways that harmed users' experiences:

> *"Even if your [quantitative metrics] come out better... at the end of the day, it's really just different from what you had before [...] and [users] notice that for their particular scenario, it's different in a negative way."*

A majority (71%) of survey respondents indicated that it would be at least "Very" useful to have tools to help their team understand potential UX side effects caused by a particular strategy for addressing a fairness issue. To minimize the risk of side effects, several teams had a practice of implementing many local, "band-aid," fixes rather than trying to address the root cause of an issue. In some cases (e.g., when the issue is a "one-off"), such local fixes may be sufficient. However, interviewees also reported that these fixes, such as censoring specific system outputs or responses to certain user inputs, sometimes resulted in other forms of unfairness (e.g., harms to certain user populations caused by the censoring itself).

How much more data would we need to collect? In cases where teams had considered addressing a fairness issue by collecting additional data, interviewees often shared needs for support in estimating the minimum number of additional data points per subpopulation that they would need to address the issue (66% of survey respondents indicated that they would find such support at least "Very" useful). Most of our interviewees' teams currently rely on developer intuition. For example, a PM working on image captioning (R2) said,

> *"It's just hope and trial and error... [the developers have] experimented so much with these models [...] that they can say 'generally, [we will need] this much data to make an impact on this type of model to change things this much.' "*

However, a developer on the same team (R4) noted that their initial estimates are often wrong, which can be costly,

> *"especially when it takes two weeks to get an answer. [...] I always would just really want to know how much was enough."*

Concerns about the fairness of fairness interventions. Interviewees often expressed unease with the idea that their teams' technical choices can have major societal impacts. For example, a technical director working on general-purpose ML tools (R32) said, *"[ML] models' main assumption [is] that the past is similar to the future. [...] if I don't want to have the same future, am I in the position to define the future for society or not?"* Another interviewee (R6) expressed doubts about the morality of targeting specific subpopulations for additional data collection, even if this data collection ultimately serves to improve fairness for those subpopulations (cf. [92]):

> *"Targeting people based on certain aspects of their person... I don't know how we would go about doing that [in the] most morally and ethically and even vaguely responsible way."*

Several interviewees suggested it would be helpful to have access to domain-specific resources, such as ethical frameworks and case studies, to guide their teams' ongoing efforts around fairness (55% of survey respondents indicated that having access to such resources would be at least "Very" useful).

Changes to the broader system design. The fair ML research literature has tended to focus heavily on the development of algorithmic methods to assess and mitigate biases in individual ML models. However, interviewees emphasized that many fairness issues that arise in real-world ML systems may be most effectively addressed through changes to the broader system design. For example, R3 recalled a case where their image captioner was systematically mislabeling images of female doctors as "nurses," in accordance with historical stereotypes. The team resolved the issue by replacing the system outputs "nurse" and "doctor" with the more generic "healthcare professional." Several other interviewees described "fail-soft" design strategies their team employs to try to ensure that the worst-case harm is minimized (cf. [10, 74]). As a PM for web search (R5) put it, *"Sometimes, you start with what you know won't cause more harm, and [then] iterate."* R14 noted that when their team designs actions (e.g., personalized messages) taken in response to particular model

RIGHTSLINK

outputs, they try to imagine the impacts these actions might have in specific false-positive and false-negative scenarios. Indeed, 40% of survey respondents reported that they had tried such fail-soft strategies to address fairness issues.

Implications. Future research should explore educational resources, processes, and tools to help teams identify the cheapest, most effective strategies to address particular fairness issues. This might include helping teams estimate the minimum amount of additional data required to address a fairness issue, or helping teams anticipate trade-offs between definitions of fairness and other desiderata for an ML system, such as user satisfaction (moving beyond the fair ML literature's focus on trade-offs between fairness and predictive accuracy). In some cases, the best option available to a team may be to refrain from applying a particular fairness intervention (e.g., to avoid greater harms to users) [96].

Biases in the Humans in the Loop

Finally, several interviewees stressed the importance of explicitly considering biases that may be present in the humans embedded in the various stages of the ML development pipeline, such as crowdworkers who annotate training data or user-study participants tasked with surfacing undesirable biases in ML systems [9, 51, 60, 85, 104]. For example, a UX designer working on automated essay scoring (R20) noted that their training data is collected by hiring human scorers to evaluate essays according to a detailed rubric. However, their team suspects that irrelevant factors may influence scorers' judgments [6, 81]. An ML engineer on the team (R19) suggested it would be valuable to have support in auditing their human scoring process. R19 proposed auditing scorers by injecting artificially generated essays into the scoring pool, using a hypothetical tool that can *"paraphrase [an essay] in another subgroup's style [...] a different voice [or] vernacular [...without] chang[ing] the linguistic content otherwise... and say, 'If you apply this linguistic feature, do the scores change?' "* Similarly, 68% of survey respondents marked tools to simulate counterfactuals (cf. [45, 68]) as at least "Very" useful.

More broadly, 69% of survey respondents, out of the 79% who were asked the question, marked tools to reduce the influence of human biases on their labeling or scoring processes (cf. [61]) at least "Very" useful. Furthermore, 69% of respondents, out of the 32% who were asked the question, reported that their teams already actively try to mitigate biases in their labeling or scoring processes at least "Sometimes."

Implications. Future research should explore ways to help teams better understand biases that may be present in the humans embedded throughout the ML development pipeline, as well as ways to mitigate these biases (see [61, 79, 104]).

5 CONCLUSIONS AND FUTURE DIRECTIONS

Although industry practitioners are already grappling with biases and unfairness in ML systems, research on fair ML is rarely guided by an understanding of the challenges faced by practitioners [98, 106]. In this work, we conducted the first systematic investigation of industry teams' challenges and needs for support in developing fairer ML systems. Even when practitioners are motivated to improve fairness in their products, they often face various technical and organizational barriers. Below, we highlight just a few broad directions for future research to reduce some of these barriers:

- Although the fair ML literature has overwhelmingly focused on algorithmic "de-biasing," future research should also support practitioners in collecting and curating high-quality datasets in the first place, with an eye towards fairness in downstream ML models (cf. [23, 40, 59, 61, 73]).
- Because fairness can be context and application dependent [46, 71], domain-specific educational resources, metrics, processes, and tools (e.g., [13, 27]) are urgently needed.
- Although most fairness auditing methods assume access to individual-level demographics, many teams are only able to collect such information at coarser levels, if at all. Future research should explore ways to support fairness auditing with access to only coarse-grained demographic information (e.g., neighborhood- or school-level demographics).
- Another rich area for future research is the development of processes and tools for fairness-focused debugging [38, 45]. For instance, it can be highly challenging to determine whether specific issues (e.g., complaints from customers) are "one-offs" or symptomatic of broader, systemic problems that might require deeper investigation, let alone to identify the most effective strategies for addressing them.
- Finally, our findings point to needs for automated auditing tools and new approaches to prototyping ML systems (cf. [30]). Interviewees highlighted limitations of existing UX prototyping methods for surfacing fairness issues in complex, multi-component ML systems (e.g., chatbots), where fairness can be heavily context dependent [46, 71], and the space of potential contexts is often very large.

The rapidly growing area of fairness in ML presents many new challenges. ML systems are increasingly widespread, with demonstrated potential to amplify social inequities, or even to create new ones [8, 12, 17, 21, 93]. Therefore, as research in this area progresses, it is urgent that research agendas be aligned with the challenges and needs of those who affect and are affected by ML systems. We view the directions outlined in this paper as critical opportunities for the ML and HCI research communities to play more active, collaborative roles in mitigating unfairness in real-world ML systems.

RIGHTSLINK

ACKNOWLEDGMENTS

We thank all our interviewees and survey respondents for their participation. We also thank Mary Beth Kery, Bogdan Kulynych, Michael Madaio, Alexandra Olteanu, Rebekah Overdorf, Ronni Sadovsky, Joseph Seering, Ben Shneiderman, Michael Veale, and our anonymous CHI 2019 reviewers for their insightful feedback. This research was initiated while the first author was a summer intern at Microsoft Research.

REFERENCES

[1] ACM. 2018. ACM Conference on Fairness, Accountability, and Transparency (ACM FAT*). https://fatconference.org/ Accessed: 2018-06-15.

[2] ACM. 2018. FAT/ML. https://www.fatml.org.Accessed: 2018-06-15.

[3] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *Proceedings of the Thirty-fifth International Conference on Machine Learning (ICML 2018)*.

[4] Oscar Alvarado and Annika Waern. 2018. Towards algorithmic experience: Initial efforts for social media contexts. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI 2018)*. ACM, 286.

[5] Saleema Amershi, Max Chickering, Steven M Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. 2015. ModelTracker: Redesigning performance analysis tools for machine learning. In *Proceedings of the 2015 CHI Conference on Human Factors in Computing Systems (CHI 2015)*. ACM, 337–346.

[6] Evelin Amorim, Marcia Cançado, and Adriano Veloso. 2018. Automated essay scoring in the presence of biased ratings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Vol. 1. 229–237.

[7] Rico Angell, Brittany Johnson, Yuriy Brun, and Alexandra Meliou. 2018. Themis: Automatically testing software for discrimination. In *Proceedings of the Demonstrations Track at the 26th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE), Lake Buena Vista, FL, USA*.

[8] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals, and it's biased against blacks. *ProPublica* (2016).

[9] Josh Attenberg, Panagiotis G Ipeirotis, and Foster J Provost. 2011. Beat the machine: Challenging workers to find the unknown unknowns. *Human Computation* 11, 11 (2011), 2–7.

[10] Ryan SJd Baker, Albert T Corbett, Ido Roll, and Kenneth R Koedinger. 2008. Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction* 18, 3 (2008), 287–314.

[11] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Cal. L. Rev.* 104 (2016), 671.

[12] BBC. 2013. Google searches expose racial bias, says study of names. *BBC News* (Feb 2013). https://www.bbc.com/news/technology-21322183. Accessed: 2018-09-03.

[13] Emily Bender and Batya Friedman. 2018. Data statements for NLP: Toward mitigating system bias and enabling better science. *OpenReview Preprint*.

[14] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in criminal justice risk assessments: The State of the Art. *Sociological Methods & Research* (2018).

[15] Reuben Binns. 2018. Fairness in machine learning: Lessons from political philosophy. *Proceedings of Machine Learning Research* 81, 149–159.

[16] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. "It's reducing a human being to a percentage": Perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI 2018)*. ACM, 377.

[17] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems (NeurIPS 2016)*. 4349–4357.

[18] Nigel Bosch, Sidney K D'Mello, Ryan S Baker, Jaclyn Ocumpaugh, Valerie Shute, Matthew Ventura, Lubin Wang, and Weinan Zhao. 2016. Detecting student emotions in computer-enabled classrooms. In *Proceedings of the 2016 International Joint Conference on Artificial Intelligence (IJCAI 2016)*. 4125–4129.

[19] Justin Bozonier. 2015. *Test-driven machine learning*. Packt Publishing Ltd.

[20] Taina Bucher. 2017. The algorithmic imaginary: Exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society* 20, 1 (2017), 30–44.

[21] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 2018 Conference on Fairness, Accountability and Transparency (FAT* 2018)*. 77–91.

[22] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI 2017)*. ACM, 2334–2346.

[23] Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why is my classifier discriminatory? *Advances in Neural Information Processing Systems (NeurIPS 2018)*.

[24] Nan-Chen Chen, Jina Suh, Johan Verwey, Gonzalo Ramos, Steven Drucker, and Patrice Simard. 2018. AnchorViz: Facilitating classifier error discovery through interactive semantic data exploration. In *23rd International Conference on Intelligent User Interfaces (IUI 2018)*. ACM, 269–280.

[25] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2 (2017), 153–163.

[26] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *ACM Conference on Fairness, Accountability, and Transparency (FAT* 2018)*. 134–148.

[27] Henriette Cramer, Jean Garcia-Gathright, Sravana Reddy, Aaron Springer, and Romain Takeo. In press. Translation, tracks and data: Algorithmic bias in practice. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA 2019)*.

[28] Kate Crawford. 2017. Artificial intelligence with very real biases. http://www.wsj.com/articles/artificial-intelligencewith-very-real-biases-1508252717. Accessed: 2018-06-15.

[29] Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI 2018)*. ACM, 412.

[30] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX design innovation: Challenges for working with machine learning as a design material. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI 2017)*. ACM, 278–288.

[31] DSSG. 2018. Aequitas: Bias and fairness audit toolkit. http://aequitas.dssg.io. Accessed: 2018-08-29.

RIGHTSLINK

[32] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the Third Innovations in Theoretical Computer Science Conference (ITCS 2012)*. ACM, 214–226.

[33] Cynthia Dwork and Christina Ilvento. 2018. Fairness under composition. CoRR arXiv:1806.06122.

[34] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 7639 (2017), 115.

[35] Kadija Ferryman and Mikaela Pitcan. 2018. Fairness in precision medicine. *Data & Society* (2018).

[36] World Wide Web Foundation. 2017. Algorithmic accountability. *World Wide Web Foundation* (2017).

[37] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)* 14, 3 (1996), 330–347.

[38] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. 2017. Fairness testing: Testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering (FSE 2017)*. ACM, 498–510.

[39] Timnit Gebru, Jonathan Krause, Jia Deng, and Li Fei-Fei. 2017. Scalable annotation of fine-grained categories without experts. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI 2017)*. ACM, 1877–1881.

[40] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. CoRR arXiv:1803.09010.

[41] Dave Gershgorn. 2018. America's biggest body-camera company says facial recognition isn't accurate enough for police. https://qz.com/1351519/facial-recognition-isnt-yet-accurate-enough-for-policing-decisions/. Accessed: 2018-08-30.

[42] Dave Gershgorn. 2018. If AI is going to be the world's doctor, it needs better textbooks. https://qz.com/1367177/if-ai-is-going-to-be-the-worlds-doctor-it-needs-better-textbooks. Accessed: 2018-09-16.

[43] Vivian Giang. 2018. The potential hidden bias in automated hiring systems. https://www.fastcompany.com/40566971/the-potential-hidden-bias-in-automated-hiring-systems. Accessed: 2018-09-03.

[44] Google. 2018. The UX of AI - Library. https://design.google/library/ux-ai/. Accessed: 2018-08-28.

[45] Google. 2018. The What-If Tool: Code-free probing of machine learning models. https://ai.googleblog.com/2018/09/the-what-if-tool-code-free-probing-of.html. Accessed: 2018-09-18.

[46] Ben Green and Lily Hu. 2018. The myth in the methodology: Towards a recontextualization of fairness in machine learning. In *the ICML 2018 Debates Workshop*.

[47] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. 2018. Gender recognition or gender reductionism?: The social implications of embedded gender recognition systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI 2018)*. ACM, 8.

[48] Bruce Hanington and Bella Martin. 2012. *Universal methods of design: 100 ways to research complex problems, develop innovative ideas, and design effective solutions*. Rockport Publishers.

[49] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS 2016)*. 3315–3323.

[50] HireVue.com. 2018. Video interview software for recruiting & hiring. https://www.hirevue.com/. Accessed: 2018-08-28.

[51] Kenneth Holstein, Gena Hong, Mera Tegene, Bruce M McLaren, and Vincent Aleven. 2018. The classroom as a dashboard: Co-designing wearable cognitive augmentation for K-12 teachers. In *Proceedings of the Eighth International Conference on Learning Analytics and Knowledge (LAK 2018)*. ACM, 79–88.

[52] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. 2017. Intelligent tutors as teachers' aides: Exploring teacher needs for real-time analytics in blended classrooms. In *Proceedings of the Seventh International Learning Analytics and Knowledge Conference (LAK 2017)*. ACM, 257–266.

[53] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. 2018. Student learning benefits of a mixed-reality teacher awareness tool in AI-enhanced classrooms. In *Proceedings of the International Conference on Artificial Intelligence in Education (AIED 2018)*. Springer, 154–168.

[54] Karen Holtzblatt and Sandra Jones. 1993. Contextual inquiry: A participatory technique for system design. *Participatory design: Principles and practices* (1993), 177–210.

[55] AI Now Institute. 2018. AI Now Institute. https://ainowinstitute.org. Accessed: 2018-08-03.

[56] Alankar Jain, Florian Pecune, Yoichi Matsuyama, and Justine Cassell. 2018. A user simulator architecture for socially-aware conversational agents. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents (IVA 2018)*. ACM, 133–140.

[57] David Janzen and Hossein Saiedian. 2005. Test-driven development concepts, taxonomy, and future direction. *Computer* 38, 9 (2005), 43–50.

[58] Yuan Jia, Bin Xu, Yamini Karanam, and Stephen Voida. 2016. Personality-targeted gamification: A survey study on personality traits and motivational affordances. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI 2016)*. ACM, 2001–2013.

[59] Nathan Kallus and Angela Zhou. 2018. Residual unfairness in fair machine learning from prejudiced data. CoRR arXiv:1806.02887.

[60] Ece Kamar. 2016. Directions in hybrid intelligence: Complementing AI systems with human intelligence.. In *Proceedings of the 2016 International Joint Conference on Artificial Intelligence (IJCAI 2016)*. 4070–4073.

[61] Ece Kamar, Ashish Kapoor, and Eric Horvitz. 2015. Identifying and accounting for task-dependent bias in crowdsourcing. In *Third AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2015)*.

[62] Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 2015 CHI Conference on Human Factors in Computing Systems (CHI 2015)*. ACM, 3819–3828.

[63] Mary Beth Kery, Marissa Radensky, Mahima Arya, Bonnie E John, and Brad A Myers. 2018. The story in the notebook: Exploratory data science using a literate programming tool. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI 2018)*. ACM, 174.

[64] Niki Kilbertus, Adrià Gascón, Matt J Kusner, Michael Veale, Krishna P Gummadi, and Adrian Weller. 2018. Blind justice: Fairness with encrypted sensitive attributes. *Proceedings of the Thirty-Fifth International Conference on Machine Learning (ICML 2018)*.

[65] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *Proceedings of the Eighth Innovations in Theoretical Computer Science Conference (ITCS 2017)* (2016).

[66] Todd Kulesza, Saleema Amershi, Rich Caruana, Danyel Fisher, and Denis Charles. 2014. Structured labeling for facilitating concept evolution in machine learning. In *Proceedings of the 2014 CHI Conference on Human Factors in Computing Systems (CHI 2014)*. ACM, 3075–3084.

RIGHTSLINK

[67] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI 2015)*. ACM, 126–137.

[68] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems (NeurIPS 2017)*. 4066–4076.

[69] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Eric Horvitz. 2017. Identifying Unknown Unknowns in the Open World: Representations and Policies for Guided Exploration.. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2017)*.

[70] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)* 9 (2016).

[71] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684.

[72] Min Kyung Lee and Su Baykal. 2017. Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work (CSCW 2017)*. 1035–1048.

[73] Anqi Liu, Lev Reyzin, and Brian D Ziebart. 2015. Shift-pessimistic active learning using robust bias-aware prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2015)*. 2764–2770.

[74] Hugo Liu and Push Singh. 2002. MAKEBELIEVE: Using commonsense knowledge to generate stories. In *Proceedings of the Fourteenth Innovative Applications of Artificial Intelligence Conference (IAAI 2002)*. 957–958.

[75] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed impact of fair machine learning. *Proceedings of the Thirty-fifth International Conference on Machine Learning (ICML 2018)* (2018).

[76] Natasha Lomas. 2018. Accenture wants to beat unfair AI with a professional toolkit. https://techcrun ch.com/2018/06/09/accenture-wants-to-beat-unfair-ai-with-a-professional-toolkit/. Accessed: 2018-06-14.

[77] Natasha Lomas. 2018. IBM launches cloud tool to detect AI bias and explain automated decisions. https://techcrunch.com/2018/09/19/ibm-launches-cloud-tool-to-detect-ai-bias-and-explain-automated-decisions. Accessed: 2018-09-20.

[78] Kristian Lum and William Isaac. 2016. To predict and serve? *Significance* 13, 5 (2016), 14–19.

[79] Lingyu Lyu, Mehmed Kantardzic, and Tegjyot Singh Sethi. 2018. Sloppiness mitigation in crowdsourcing: detecting and correcting bias for crowd scoring tasks. *International Journal of Data Science and Analytics* (2018), 1–21.

[80] Christopher J Maclellan, Erik Harpstead, Rony Patel, and Kenneth R Koedinger. 2016. The Apprentice Learner architecture: Closing the loop between learning theory and educational data.. In *Proceedings of the 2016 International Conference on Educational Data Mining (EDM 2016)*. 151–158.

[81] John M Malouff and Einar B Thorsteinsson. 2016. Bias in grading: A meta-analysis of experimental research findings. *Australian Journal of Education* 60, 3 (2016), 245–256.

[82] MURAL. 2018. MURAL - Make remote design work. https://mural.co/. Accessed: 2018-08-02.

[83] Arvind Narayanan. 2018. 21 fairness definitions and their politics. *FAT* 2018 tutorial* (2018).

[84] Safiya Umoja Noble. 2018. *Algorithms of oppression: How search engines reinforce racism.* NYU Press.

[85] Besmira Nushi, Ece Kamar, Eric Horvitz, and Donald Kossmann. 2017. On human intellect and machine failures: Troubleshooting integrative machine learning systems. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2017)*. 1017–1025.

[86] US Department of Education (ED). 2018. Family Educational Rights and Privacy Act (FERPA). https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html. Accessed: 2018-09-04.

[87] Partnership on AI. 2018. The Partnership on AI. https://www.partnershiponai.org. Accessed: 2018-09-03.

[88] Julia Powles and Hal Hodson. 2017. Google DeepMind and healthcare in an age of algorithms. *Health and Technology* 7, 4 (2017), 351–367.

[89] pymetrics. 2018. matching talent to opportunity. https://www.pymetrics.com/. Accessed: 2018-08-28.

[90] Qualtrics. 2013. Qualtrics. *Provo, UT, USA* (2013).

[91] Emilee Rader and Rebecca Gray. 2015. Understanding user beliefs about algorithmic curation in the Facebook news feed. In *Proceedings of the 2015 CHI Conference on Human Factors in Computing Systems (CHI 2015)*. ACM, 173–182.

[92] Manish Raghavan, Aleksandrs Slivkins, Jennifer Wortman Vaughan, and Zhiwei Steven Wu. 2018. The externalities of exploration and how data diversity helps exploitation. In *Proceedings of the Thirty-first Annual Conference on Learning Theory (COLT 2018)*.

[93] Dillon Reisman, Jason Schultz, K Crawford, and M Whittaker. 2018. Algorithmic impact assessments: A practical framework for public agency accountability. *AI Now Institute* (2018).

[94] Ari Schlesinger, Kenton P O'Hara, and Alex S Taylor. 2018. Let's talk about race: Identity, chatbots, and AI. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI 2018)*. ACM, 315.

[95] D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. 2015. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems (NeurIPS 2015)*. 2503–2511.

[96] Andrew D Selbst, danah boyd, Sorelle Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *ACM Conference on Fairness, Accountability, and Transparency (FAT* 2018)*.

[97] Data & Society. 2018. Algorithmic accountability: A primer. *Data & Society* (2018).

[98] Aaron Springer, J. Garcia-Gathright, and Henriette Cramer. 2018. Assessing and addressing algorithmic bias—But before we get there. In *Proceedings of the AAAI 2018 Spring Symposium: Designing the User Experience of Artificial Intelligence*.

[99] Donald E Stokes. 1997. *Pasteur's quadrant: Basic science and technological innovation.* Brookings Institution Press.

[100] Sarah Tan, Julius Adebayo, Kori Inkpen, and Ece Kamar. 2018. Investigating Human+Machine Complementarity for Recidivism Predictions. CoRR arXiv:1808.09123.

[101] Rob Thubron. 2018. IBM secretly used NYPD CCTV footage to train its facial recognition systems. https://www.techspot.com/news/76323-ibm-secretly-used-nypd-cctv-footage-train-facial.html. Accessed: 2018-09-16.

[102] Kentaro Toyama. 2018. From needs to aspirations in information technology for development. *Information Technology for Development* 24, 1, 15–36.

[103] Melissa A Valentine, Daniela Retelny, Alexandra To, Negar Rahmati, Tulsee Doshi, and Michael S Bernstein. 2017. Flash organizations: Crowdsourcing complex work by structuring crowds as organizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI 2017)*. ACM, 3523–3537.

RIGHTSLINK

[104] Jennifer Wortman Vaughan. 2018. Making better use of the crowd. *Journal of Machine Learning Research* 18, 193 (2018), 1–46.

[105] Michael Veale and Reuben Binns. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society* 4, 2 (2017), 2053951717743530.

[106] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI 2018)*. ACM, 440.

[107] Sara Wachter-Boettcher. 2017. AI recruiting tools do not eliminate bias. http://time.com/4993431/ai-recruiting-tools-do-not-eliminate-bias. Accessed: 2018-09-01.

[108] Allison Woodruff, Sarah E Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI 2018)*. ACM, 656.

[109] Qian Yang. 2018. Machine learning as a UX design material: How can we imagine beyond automation, recommenders, and reminders?. In *Proceedings of the AAAI 2018 Spring Symposium: Designing the User Experience of Artificial Intelligence*.

[110] Qian Yang, Jina Suh, Nan-Chen Chen, and Gonzalo Ramos. 2018. Grounding interactive machine learning tool design in how non-experts actually build models. In *Proceedings of the 2018 Conference on Designing Interactive Systems (DIS 2018)*. ACM, 573–584.

[111] Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey, and James F Antaki. 2016. Investigating the heart pump implant decision process: Opportunities for decision support tools to help. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI 2016)*. ACM, 4477–4488.

[112] Maggie Zhang. 2015. Google photos tags two African-Americans as gorillas through facial recognition software. https://tinyurl.com/Forbes-2015-07-01. Accessed: 2018-07-12.

[113] Zian Zhao, Michael Madaio, Florian Pecune, Yoichi Matsuyama, and Justine Cassell. 2018. Socially-conditioned task reasoning for a virtual tutoring agent. In *Proceedings of the Seventeenth International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018)*. 2265–2267.

RIGHTSLINK