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

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

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

Jennifer Wortman Vaughan
Microsoft Research
New York, NY
jenn@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 industry practitioners and 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 industry practitioners' needs.

KEYWORDS

algorithmic bias, fair machine learning, product teams, needfinding, empirical study, UX of machine learning

1 INTRODUCTION

Machine learning (ML) systems increasingly influence every facet of our lives, including the quality of healthcare and education we receive [16, 31, 32, 49], which news or social media posts we see [4, 18, 84], who receives a job [46, 82], who is released from jail [8, 23], and who is subjected to increased policing [65, 71, 94]. With this growth, the potential of ML to amplify social inequities has received increasing 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 [40, 95], 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, 77].

Substantial effort in the rapidly growing research literature on fairness in machine learning [1, 2] has centered on the development of mathematical definitions of "fairness" [23, 29, 45, 76] and algorithmic methods to assess and mitigate undesirable biases in relation to these definitions [3, 45, 64]. As the field matures, integrated toolkits are being developed, with the aim of making these methods more widely accessible and usable [7, 28, 35, 42, 70]. While some fair ML toolkits are already being prototyped with practitioners, their initial design often appears to be driven more by the availability of algorithmic methods than by real-world needs for support (cf. [48, 99]). If such tools are to have a positive impact on industry practice, however, it is crucial that their design be informed by an understanding of practitioners' actual challenges in creating fair ML systems [99].

In this work, we investigate challenges faced by commercial ML product teams—whose products may affect the lives of millions of users [19, 89]—in monitoring for unfairness and taking appropriate action [89, 94]. Through semi-structured interviews with 35 individuals, 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 key themes surfaced in our interviews, we then conduct an anonymous survey of 267 industry ML practitioners, across a broader range of contexts.

The present work is the first, to our knowledge, to systematically investigate industry ML practitioners' challenges and needs around algorithmic fairness. Through our investigations, 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" algorithms, which view the training data as fixed [21, 55], most of our interviewees report that their teams consider data collection, rather than model training, 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.

Furthermore, while 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 model metrics [19, 24, 61], 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 their teams' own blind spots. For instance, teams often struggled to anticipate which subpopulations and forms of unfairness they need to consider when developing and auditing specific kinds of ML applications.

Based on these and other findings, we highlight several opportunities for the fair ML and HCI research communities to have a greater impact on industry practice.

2 BACKGROUND AND RELATED WORK

The design, prototyping, and maintenance of machine learning systems raises many unique challenges [27, 63, 78, 87] not commonly faced with other kinds of intelligent systems or computing systems more broadly [34, 66, 67]. 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 data, modeling, and system-design choices [27, 41, 47, 97]. In addition, a growing body of research focuses on the design and development of programmer tools that can support developers in debugging and effectively monitoring the predictive performance of complex ML systems [5, 22, 59, 63, 78].

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 machine learning to HCI to public policy, law, and ethics [11, 89, 94]. Specialized research communities and initiatives are forming with a focus on bias and fairness 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 [51], and the Partnership on AI [80]. Significant effort in the FAT/ML community has focused on the development of mathematical definitions of "fairness" [13, 23, 29, 76] and algorithmic methods to assess and mitigate bias in relation to these definitions [3, 15, 45, 64]. In contrast, HCI researchers have empirically studied users' expectations and perceptions related to fairness in algorithmic systems and found that these do not always align with existing mathematical definitions [14, 66, 67, 96]. Other work has focused on auditing widely-used ML products from the outside [8, 19, 26, 58], often coupled with high-level calls to action aimed at those responsible for developing and maintaining these systems or for regulating their use [33, 86, 88]. Crawford, Springer et al., and others have highlighted an urgent need for internal processes and tools to support companies in developing fairer systems in the first place [25, 89].

Despite this widespread attention on fairness and bias in ML, to the best of our knowledge, only one prior study, by Veale et al. [94], 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 [65, 71, 94] and child mistreatment detection [24], to understand the challenges they face in aligning the behavior of ML 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 FAT/ML literature.

In the same spirit as Veale et al., this work investigates ML practitioners' needs for support, with the aim of identifying fruitful opportunities for future research [90]. However, whereas Veale et al. studied 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. We see supporting industry ML practitioners as a critical, upstream step towards fairer algorithm-assisted decision making in the public sector, where systems are often built on top of ML products and APIs developed in industry [19, 38, 81, 91]. In comparison to the high-stakes public sector applications studied by Veale et al., industry practitioners also 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 are less clear-cut than in high-stakes public-sector contexts, and motivations and organizational priorities can differ considerably in industry.

3 METHODS

To better understand product teams' needs for support in developing fair ML systems, we conducted a series of semi-structured interviews with a total of 35 practitioners, across 25 ML product teams in 10 major companies. To investigate the prevalence and generality of themes that emerged from these interviews, we then conducted an anonymous survey with a broader sample of 267 industry ML practitioners. The study went through an ethical review and was IRB-approved. Semi-structured interview protocols and survey questions are provided in the supplementary materials.

3.1 Interview Study

Our interview study proceeded in two rounds. First, to get a broad sense of challenges, we conducted six formative interviews. Building on initial themes that emerged, we then conducted more in-depth interviews with a broader sample.

In the first round, to get a broad sense of current practices and challenges, we conducted a series of six semi-structured

Table 1: Interview participants' self-identified technology areas and team roles. Where multiple participants were interviewed from the same product team, participant identifiers are grouped within square brackets.

Technology Area	Roles of Participants	Participants
Adaptive Tutoring & Mentoring	Chief Data Scientist, CTO, Data Scientist, Research Scientist	R10, [R13, R14], R30
Chatbots / Conversational AI	CEO, Product Manager, UX Researcher	[R17, R18], R35
Computer Vision & Multimodal Sensing	CTO, ML Engineer, Product Manager, Software Engineer	[R2, R3, R4], R6, R7, R9, R26
General-purpose ML Tools (e.g., APIs)	Chief Architect, Director of ML, Product Manager	R25, R32, R34
Natural Language Processing (e.g., Automated Writing Evaluation, NLU, Speech, Translation)	Data Collection Manager, Data Collector, Domain Expert, ML Engineer, Product Manager, Research Software Engineer, Technical Manager, UX Designer	R1, [R15, R16, R19, R20, R21, R22], R24, [R27, R29], R28, R31
Recommender Systems (e.g., Hiring, Personalized Retail)	Chief Data Scientist, Data Scientist, Head of Diversity Analytics	R8, R12, R23, R33
Web Search	Product Manager	R5, R11

interviews with product managers (PMs), each in a different technology area. Each of these initial interviews lasted 30 minutes and was conducted by teleconference since 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. Participants were then asked whether fairness is something their team regularly discusses or incorporates into their workflow. The meaning of "fairness" was intentionally left open: we were interested in hearing PMs' notions about what it might mean for their ML products to be "fair." However, if a PM requested clarification at any point, the interviewer 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. [94]). 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. Interview participants included 29 individuals, across 10 major technology companies and 19 ML product teams. As shown in Table 1, we interviewed individuals across a range of technology areas and team roles. Whenever possible, we tried to interview multiple roles on the same team to hear (potentially) different perspectives on the team's current practices and challenges around ML fairness. Interview participants were recruited using a snowball sampling approach. We searched for popular press articles related to algorithmic bias and fairness and contacted members of product teams whose products had previously received relevant media coverage (i.e., news stories about unfair behavior observed in these products). In addition, we emailed direct contacts across over 30 major companies. In both cases, we asked email recipients to share our interview invitation with any colleagues working on ML products (in any team role) at their own company or others. At the end of each interview, participants 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. [94]. For instance, given a recent trend of negative media pieces calling out algorithmic bias and unfairness in widely-used ML systems (e.g., [12, 91, 95, 100]), our contacts often expressed strong fears that their company or team'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' 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 participants' 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 participants that we would allow them to review any (de-identified) direct quotes before including them in any research publications. All 29 participants in the main study consented to be audio recorded.

All interviews in the main study lasted between 40 and 60 minutes. In each interview, participants were first reminded

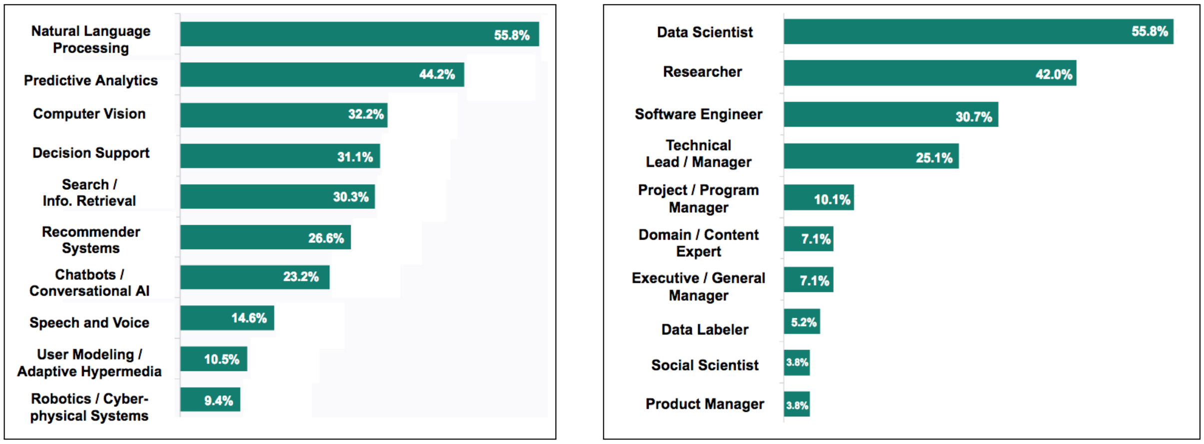

Figure 1: Survey demographics: the top 10 reported application domains (left) and the top 10 reported team roles (right).

of the overall purpose of the study, 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), developing an ML product, to detecting and potentially addressing fairness issues in that product. For each of these stages, participants 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 development, a core sequence of four follow-ups was used across all stages. First, participants 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, participants were instructed to imagine they could return to these critical episodes, but this time with access to a magical oracle which they could ask any questions to help them in the moment [48, 59]. We asked the question in this way to encourage participants to speak freely about their current challenges and needs without feeling constrained to those for which they believed a solution was currently possible [48, 59]. After participants 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 (e.g., *"Is the issue caused by the training set we're using, or is it the model?"*). This follow-up often led participants to reflect on gaps between their current and ideal processes. Finally, participants 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 two standard techniques from contextual design: interpretation sessions and affinity diagramming [44, 50]. Following a bottom-up, affinity diagramming approach (using MURAL [75]), we iteratively generated and grouped 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*.

3.2 Survey

To validate our interview findings on a broader population, we next conducted an anonymous online survey (using Qualtrics [83]). Participants were recruited via snowball sampling. We emailed the survey to direct contacts at over 40 companies that develop ML products, and invited them to pass the survey on to others (either within or outside of their companies) who are part of a team developing ML products (in any role). We also announced the survey on social media (e.g., Twitter) and several online communities related to ML and AI. These venues included the Kaggle forums and special interest groups on LinkedIn, Facebook, Reddit, and Slack.

We structured the survey to act as a quantitative supplement to the interview. As such, the high-level structure of the survey mirrored that of the main interview study. Based on our interview findings, we developed survey questions to investigate the prevalence and generality of emerging themes. First, we asked a set of demographics questions to understand our participants' backgrounds, including their role(s) on their product team, and their team's technology area(s). In a branching sequence of survey sections, participants 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 model development pipeline. For each survey question, closed-ended response options were provided based on themes that emerged through affinity diagramming, in addition to free-response options that allowed respondents to elaborate on their responses.

A total of 287 individuals started the survey. However, not all respondents answered all of the questions, so here we analyze only the 267 respondents who completed at least one section beyond the demographics. The most common technology areas and team roles are shown in Figure 1, and additional demographics are available in the supplementary materials.

4 RESULTS AND DISCUSSION

Although participants spanned a diverse range of companies, team roles, and applications, we observed many commonalities. In the following, we discuss product teams' current challenges and needs around fairness in ML, organized by top-level themes that emerged through our affinity diagramming. These include needs for support in collecting/curating more representative datasets, anticipating and overcoming teams' blind spots, implementing more proactive auditing processes, monitoring fairness in complex ML systems, deciding how best to address particular instances of unfairness, and assessing and mitigating 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 literature thus far. We supplement these interview findings with corresponding survey results. Interviewees are identified with an "R," and survey responses are identified with percentages. Given that the survey uses branching logic (e.g., respondents are only asked questions about addressing fairness issues if they report that their team has previously detected such issues in their products), some questions are completed by a subset of survey respondents. In such cases, question-specific response totals are provided in addition to percentages. To illustrate general themes, we share direct quotes in cases where we have received explicit permission from interviewees, in accordance with our study's IRB-approval and consent form. As such, while the presented themes were drawn from a range of domains, the set represented by direct quotes may be narrower.

4.1 Fairness-aware Data Collection

While popular press articles on fairness in ML systems often use a "bias in, bias out" framing, emphasizing the central role of dataset quality [39, 95], the fair ML research literature has overwhelmingly focused on the development of algorithmic methods that attempt to correct for such biases, assuming a fixed dataset [21, 55]. 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 174 survey respondents whose teams have some control over data collection/curation, 58% reported that they currently consider fairness at these stages. Furthermore, out of 55 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."

4.1.1 Needs for support in collecting more representative datasets. Most interviewees reported that their teams do not currently have processes in place to support the collection of 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 those responsible for developing ML models. On a 5-point Likert scale from "Not at all" to "Extremely" useful, 52% of 212 respondents for whom this questions applied indicated that tools to facilitate communication between modelers and data collectors would be "Very" or "Extremely" useful.

In addition, interviewees 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 they collect, 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."*

Out of 67 survey respondents for whom this question applied, 60% marked such active guidance as at least "Very" useful. By contrast, in contexts where training data is collected via regular, "in-the-wild" use 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 effective strategies to incentivize app usage within specific populations, such as targeted gamification (cf. [54]).

4.1.2 Scaffolding fairness-aware test-set design. Several interviewees stressed the central importance of careful test set design in detecting potential fairness issues. For example, R4's team would discuss possible fairness 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 captioning system (e.g., images of female doctors were frequently mislabeled as nurses) which they ultimately traced to imbalances in their training data. An ML engineer working on a gesture recognition system (R6) noted it would be helpful to have better tooling 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."*

Out of 187 survey respondents, 66% indicated that such tooling would be "Very" or "Extremely" useful.

4.2 Challenges Due to Blind Spots

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

4.2.1 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 a particular ML application, to ensure they collect sufficient data from these groups 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 (see [43]), 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."* Out of 213 surveyed, 62% indicated it would be at least "Very" useful to have additional support in identifying relevant subpopulations for specific kinds of ML applications.

4.2.2 Detecting unexpected issues before the public does it for us.

Interviewees often reported that their teams do not discover serious fairness issues until receiving customer complaints about deployed products (or worse yet, by reading negative press articles 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 unfairness pre-deployment, even in cases where they may not have anticipated all relevant subpopulations or kinds of unfairness. Despite their efforts running user studies and human evaluations, 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).

4.2.3 Team biases and limitations.

Several of the teams we interviewed currently have a practice of getting together and trying to imagine everything that could go wrong with a particular system, so that they can make sure to proactively monitor for those issues. A few teams even reported including fairness-focused quizzes in their interview process, with the aim of hiring employees who would be good at spotting undesirable biases in training data or model outputs. 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 pitfalls in specific domains across teams with different backgrounds, who may have complementary knowledge and blind spots. Out of 189 surveyed, 67% indicated 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' cultural blind spots. A developer working on image captioning (R4) recalled cases in which customers had complained that a globally deployed system performed well for celebrities from some countries, but routinely misidentified major celebrities from others:

> *"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."*

4.3 Needs for More Proactive Auditing Processes

Once data is collected, ensuring fairness presents many unique auditing challenges. Interviewees often described their teams' current practices around fairness auditing as "reactive"—with efforts tightly focused around specific customer complaints—in contrast to these teams' proactive approaches for dealing with security vulnerabilities. As a PM working on 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."*

Interviewees revealed a range of needs for support in implementing more proactive, systematic, and comprehensive fairness auditing processes. In the following we highlight needs that surfaced across diverse technology areas and application domains. These include domain-specific auditing processes, metrics, and tools; methods to effectively monitor fairness in scenarios where individual-level demographic data is unavailable; more scalable auditing approaches; and ways to determine if an individual case is part of a systemic problem.

4.3.1 Needs for (domain-aware) 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 the teams we interviewed did not currently have fairness metrics against which they could monitor performance and progress. Likewise, out of 70 survey respondents whose teams had previously found fairness issues in their products, only 23% marked the statement *"We have [fairness] metrics / key performance indicators (KPIs)"* and only 20% marked *"We run automated tests [for fairness]"* as at least "Very" accurate in describing their current auditing processes. 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 consulted the published literature on fair ML for existing metrics. However, they often failed to find measures of "fairness" that readily applied in their specific application domains. For example, while much of the fair ML literature has focused on developing precise metrics for "allocative harms" (relating to the distribution of limited resources), many of the fairness issues that arise in domains such as web search, conversational AI, or image captioning are "representational harms" (e.g., where systems perpetuate harmful stereotypes) [25].

Some interviewees shared that they had tried holding meetings with other teams and subteams within their companies, 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 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."*

Most interviewees noted that they are not generally rewarded for their efforts around fairness. On a 5-point Likert scale from "Not at all" to "A great deal," only 21% of those surveyed reported that their team prioritizes fairness "A lot" or "A great deal," and 36% indicated "Not at all". Interviewees often engaged in ML fairness efforts on their own time and initiative, and thus 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.'"* Others suggested it would be extremely helpful to have access to tools and resources that can help their team anticipate what kinds of issues can arise in their specific application domain, together with domain-specific frameworks that can help them navigate associated complexities.

4.3.2 Fairness auditing without individual-level demographics. Although most existing auditing methods in the fair ML research literature assume access to sensitive characteristics such as gender or race at an individual level [93], many of the teams we interviewed are only able to collect such demographics at coarser levels, if at all (cf. [11, 60, 93]). For example, companies working with K-12 student populations in the US are typically restricted from collecting such demographics under school/district policies and FERPA laws [79]. Of 183 surveyed, 70% indicated that having access to tools that could support fairness auditing without individual-level demographics would be at least "Very" useful.

A small set of teams reported having attempted to use the coarse-grained demographic data they have available (e.g., region- or organization-level demographics) for fairness auditing. However, each of these teams reported quickly abandoning 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 information for 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 know, it turns out [our product] is only used by the gifted [or remedial] students, and you may not have means [to check]."*

Some teams shared that they had experimented with developing ML models to infer sensitive demographics for individuals based on available proxies, so that they could then use these models to audit their main ML products. However, interviewees worried that the use of proxies may in itself encode undesirable biases, introducing a need to audit the auditing tool. A data scientist working on automated hiring/recruiting (R23) recounted a time their team had developed one such 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 out there [to] do this automatically and with a trusted data source... that would be super useful"*

Interviewees often commented that, ultimately, it would be ideal to simply 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 [60, 93]. But interviewees emphasized that, while such technical solutions are an important prerequisite, at least half of the battle would lie in convincing stakeholders and policymakers that these mechanisms are truly secure, and that the benefits would 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 demographics to improve fairness in an ML system without ever revealing these demographics externally.

4.3.3 Needs for greater scalability and comprehensiveness.

Of 169 respondents, 49% reported that their team has previously found potential fairness issues in their products. Of those who responded that their team had not found any issues, 80% suspected there might be issues they have not yet detected, with a majority (55%) reporting that undetected issues "Probably" or "Definitely" exist in their products.

Interviewees often complained about limitations of their current testing strategies, given the enormous space of possible 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 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, if it's possible."* Drawing parallels to existing, automated tools that scan natural language datasets for potentially sensitive terms, R1 suggested that 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 procedures would be a challenging research problem, given that "fairness" can be so context dependent. R5 emphasized that domains like image captioning or web search are particularly difficult, because "fairness" can depend jointly on the system's output and the user-provided input (e.g., a query).

4.3.4 Diagnosing systemic issues from isolated cases.

Interviewees also highlighted challenges in diagnosing whether concerning isolated observations (e.g., specific complaints from user studies) are symptomatic of a broader, systemic issue, rather than being "one offs." A product and data manager for a machine translation system (R1) suggested,

> *"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."*

Of 187 respondents, 62% indicated that tools to help their teams find other instances of a potential issue, after observing a small set of cases, would be at least "Very" useful.

4.4 Needs for More Holistic Auditing Methods

Much of the existing fair ML literature has focused on applications such as recidivism prediction and automated hiring, where it can make sense to understand "fairness" in terms of well-defined metrics of ML models (e.g., between-group parity in error rates or decisions [19, 24, 61]). However, teams working on applications involving richer, more complex interactions between the user and the system—such as chatbots, automated writing evaluation, adaptive tutoring and mentoring, and web search—brought up various needs for more holistic, system-level auditing methods.

4.4.1 Fairness as a system-level property.

Many interviewees noted disconnects between the way they tend to think about fairness in their application domains 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 auditing individual ML models for bias, but instead evaluating the real-world impacts of ML systems: *"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 complex, multi-component systems, there is not always a clean mapping between performance metrics of an ML model and the system's utility for users. Machine learning components may interact with one another [30, 78], and with other (non-ML) aspects of a system's design, in ways that can be difficult to predict absent an empirical study with actual users or use contexts (cf. [34]). Furthermore, in certain domains, it may not be straightforward to even define "fair" system behavior without first understanding users' expectations and beliefs about the system (cf. [66, 67]). For example, a PM working on web search (R11) shared that their team had previously experimented with correcting a bias in image search results (a search for the term "CEO" yielded mostly 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 act 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, so that we can try to get something for everyone."'*

4.4.2 Needs for simulation-based approaches in complex domains.

In applications involving rich sequences of interaction between the user and system, fairness can be heavily contextual. As a PM for a chatbot service (R17) noted, *"contextual kinds of responses are harder to [...] be able to predict all the outcomes, when you have... such a huge possibility space."* R17 suggested it would be valuable to have ways to prototype conversational agents more rapidly, including methods to simulate conversational trajectories (cf. [52, 101]), *"and then find ways to automate the identification of risky conversation patterns that emerge."* Similarly, a data scientist working on adaptive mentoring software (R10) suggested that since their product involves a long-term feedback loop between users and the system, it would be ideal to be able to run the system against a population of "simulated mentees" (cf. [73]), to predict whether certain forms of personalized adaptation might have negative effects with respect to equity.

4.5 Addressing Detected Issues

Interviewees revealed a range of challenges and needs around the debugging and remediation of detected fairness issues. These included, among others, needs for support in identifying the cheapest, most effective strategies to address particular issues; effective methods to determine how much more training data to collect for a particular subpopulation to address gaps in model performance; processes to anticipate potential trade-offs between specific notions of fairness and other goals for an ML system (not limited to predictive accuracy); and frameworks to help navigate complex ethical decisions (the fairness of fairness interventions).

4.5.1 Needs for support in strategy selection. Interviewees reported that their teams often struggle to isolate the causes of unexpected fairness-related issues, especially when working with ML models the team consider to be "black boxes." As such, it is often difficult for teams to decide where to focus their efforts: switching to a different model class, augmenting the training data in some way, collecting more or different kinds of data, post-processing model outputs, changing the objective function, and so on (cf. [21]). Of the survey respondents who are on teams that had previously tried to address detected issues, 54% indicated it would be at least "Very" useful to have better support in comparing among specific strategies for addressing particular fairness issues. Depending on a team's specific context, different costs may be associated with different strategies. 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. By 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. [21, 35, 42]) 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 (cf. [78]), and in deciding whether to focus their efforts on their training data or on their models to address a particular issue. Of those surveyed, 63% indicated that tools to aid in these decisions (cf. [21, 42]), would be at least "Very" useful.

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

> *"Even if your [model 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."*

Of survey respondents, 71% indicated that it would be at least "Very" useful to have better tooling to help their teams understand what UX side effects a particular "fix" for a fairness issue might have. 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 band-aids, such as censoring specific model outputs or responses to certain user inputs, have sometimes resulted in other kinds of fairness issues (e.g., where the censoring itself negatively impacts certain user populations).

4.5.3 How much more data would we need to collect? In cases in which a team had considered addressing an issue (e.g., a between-group gap in model accuracy) through additional data collection, interviewees often shared desires for support in determining the minimum number of samples per subpopulation that they would need to collect to address the issue (66% of those surveyed indicated they would find such support at least "Very" useful). Most teams we interviewed currently rely on developers' intuitions for these estimates. 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, this much data to make an impact on this type of model to change things this much.' "*

A developer on the same team (R4) noted that their initial guesses 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."*

4.5.4 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 ethics of targeting specific subpopulations for additional data collection, even if this data collection serves to improve fairness in their ML products (cf. [85]):

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

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

4.5.5 Changes to the broader system design.

The fair ML research literature has tended to focus heavily on the development of algorithmic methods to improve the fairness of individual ML components (e.g., classifiers). However, interviewees emphasized that many fairness issues that arise in real-world ML systems may be most effectively addressed through modifications to the broader system design. For example, R3 recalled a case in which 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 labels "nurse" and "doctor" with the more generic "healthcare professional." Several interviewees described "fail soft" design strategies their team has employed to try to ensure that the worst-case harm of model errors is minimized (cf. [10, 69]). As a PM for web search (R5) put it, *"Sometimes, you start with what you know won't cause more harm, and [then] iterate."* In their educational software, R14 noted that when their team designs the actions (e.g., personalized messages) taken in response to specific classifier outputs, they try to imagine the impacts these actions might have in specific false-positive and false-negative scenarios. Of survey respondents, 40% reported having tried such fail-soft strategies to mitigate fairness issues.

4.6 Biases in the Humans in the Loop

Finally, several interviewees stressed the importance of explicitly considering biases that may come from the humans embedded at various stages of the ML development and maintenance pipeline, such as crowdworkers who annotate training data or user-study participants tasked with surfacing undesirable biases in ML systems [9, 47, 56, 78, 92]. For example, a UX designer working on automated essay scoring (R20) noted that their models are trained by hiring human scorers to evaluate essays according to a detailed rubric. However, their team suspects that irrelevant factors may influence scorers' judgments [6, 74]. An ML engineer on the team (R19) suggested it would be valuable to have support in auditing their human scoring process for biases, so that they could improve the process if need be. For example, R19 proposed testing scorers by injecting artificially-generated essays into the scoring pool, aided by 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?' "*

Of those surveyed, 68% marked tools to simulate counterfactuals (cf. [42, 64]), as in the above example, as at least "Very" useful. More broadly, 69% (out of 210) marked tools to reduce the influence of human biases on their labeling/scoring processes (cf. [57]) at least "Very" useful; and 69% (out of 85 whose teams currently consider fairness during data collection/curation) reported that their teams already actively try to mitigate bias in human labelers/scorers at least "Sometimes."

5 CONCLUSION AND FUTURE DIRECTIONS

Although machine learning practitioners are already grappling with algorithmic fairness in industry contexts, academic research on fair ML is rarely guided by an understanding of daily challenges faced by practitioners [89, 94]. In this work, we conducted the first systematic investigation of industry product teams' challenges and needs for support in creating fair ML systems. We uncovered a range of opportunities for future research that could have substantial impact on industry practice. Below, we highlight just a few broad challenges and opportunities, noting both contrasts and alignments with ongoing efforts in the fair ML and HCI research communities.

While the existing fair ML literature has overwhelmingly focused on algorithmic "de-biasing" methods, future research should support practitioners in collecting and/or curating representative datasets in the first place (cf. [21, 37, 55])—for example, by developing methods to actively guide data collection processes as they occur (cf. [68]), by designing tools and processes to support more effective bi-directional communication between data collectors and modelers (cf. [94]), or by developing methods to better understand (and correct for) biases in human/crowd labeling and scoring processes (see [57, 72, 92]). 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., [20, 22, 36, 62]). 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, or augmentation.

Given that "fairness" can be highly context- and application-specific [43, 66], there is urgent need for domain-specific educational resources, workflows, measures, and tools for fair ML that can help practitioners navigate the unique challenges that can arise in their specific technology and application areas. Although study participants expressed strong desires for more proactive processes around fairness in ML, they also reported struggling to apply existing, domain-general auditing and debiasing methods to their specific contexts (e.g., [23, 35, 45, 64]). Relatedly, knowing which subpopulations to consider when developing particular kinds of ML application, or what forms of unfairness to watch out for, can depend on nuanced cultural and domain knowledge that no single team is likely to have—especially in cases where ML products are deployed globally. As such, participants highlighted needs for ways to effectively pool knowledge across teams/companies with complementary knowledge and blind spots, for example via shared test sets or case studies from other teams who have worked on similar applications. A promising direction for future research within this space is the design of processes and tools to scaffold "fairness-aware test-set design": the construction and use of test sets that can effectively surface unfair system behaviors (cf. [17, 53, 98]).

In addition, future research should explore methods to support effective fairness auditing given only partial information about individual demographics (e.g., neighborhood or school level statistics). Recent work [60, 93] has begun to explore the design of encryption mechanisms to ensure that any collected individual-level demographic information can only be used for fairness auditing. However, participants suggested that in particularly sensitive contexts, stakeholders may be unwilling to reveal individual demographics, even with such privacy-preserving mechanisms in place.

A rich area for future research is the development of workflows and tools for "fairness debugging" [35, 42]. For example, it can be challenging to determine whether isolated observations of unfairness are "one-offs" that can be effectively addressed with a "band-aid" solution or indicative of systemic problems that might require deeper investigation. In addition, participants highlighted needs for support in efficiently diagnosing the cause(s) of particular unfair behaviors (cf. [63]), to help their teams decide whether to focus their efforts on the data versus the model [21, 42], or on specific model components in multi-component ML systems [63, 78].

Finally, participants highlighted limitations of existing UX prototyping methods for surfacing fairness issues in complex data-driven systems (e.g., chatbots and adaptive tutoring software), where "fairness" may be highly context dependent [43, 66], and the space of possible contexts is often very large. These observations point not only to a need for automated monitoring tools, but also for new forms of prototyping (cf. [27]) that can effectively surface unfair behaviors before ML systems are deployed in the real world.

The rapidly growing area of fairness in machine learning presents many new challenges. ML applications are increasingly widespread, with demonstrated potential to amplify social inequities, or even to create new ones [8, 12, 15, 19, 86]. Even when practitioners are motivated to improve fairness, they face various technical and organizational barriers. Thus, as research in this area progresses, it is urgent that research agendas are aligned with actual needs. We view the directions outlined above as critical opportunities for the fair ML and HCI communities to play more active, collaborative roles in addressing real-world algorithmic bias.

ACKNOWLEDGMENTS

This research was initiated while the first author was an intern at Microsoft Research. We thank all interviewees and survey respondents for their participation. In addition, we thank Michael Veale, Ben Shneiderman, and our anonymous reviewers for their insightful feedback.

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. *arXiv preprint arXiv:1803.02453* (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*. 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 33rd Annual ACM Conference on Human Factors in Computing Systems*. 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] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2017. Fairness in criminal justice risk assessments: The state of the art. *arXiv preprint arXiv:1703.09207* (2017).

[14] 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*. ACM, 377.

[15] 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*. 4349–4357.

[16] 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 *IJCAI*. 4125–4129.

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

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

[19] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*. 77–91.

[20] 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. ACM, 2334–2346.
[21] Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why is my classifier discriminatory? *arXiv preprint arXiv:1805.12002* (2018).
[22] 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*. ACM, 269–280.
[23] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.
[24] 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 *Conference on Fairness, Accountability and Transparency*. 134–148.
[25] Kate Crawford. 2017. Artificial intelligence with very real biases. http://www.wsj.com/articles/artificial-intelligencewith-very-real-biases-1508252717. Accessed: 2018-06-15.
[26] 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*. ACM, 412.
[27] 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*. ACM, 278–288.
[28] DSSG. 2018. Aequitas: Bias and fairness audit toolkit. http://aequitas.dssg.io. Accessed: 2018-08-29.
[29] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. ACM, 214–226.
[30] Cynthia Dwork and Christina Ilvento. 2018. Group fairness under composition.
[31] 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.
[32] Kadija Ferryman and Mikaela Pitcan. 2018. Fairness in precision medicine.
[33] World Wide Web Foundation. 2017. Algorithmic accountability. *World Wide Web Foundation* (2017).
[34] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)* 14, 3 (1996), 330–347.
[35] 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*. ACM, 498–510.
[36] 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*. ACM, 1877–1881.
[37] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010* (2018).
[38] 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.
[39] 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.
[40] 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.
[41] Google. 2018. The UX of AI - Library. https://design.google/library/ux-ai/. Accessed: 2018-08-28.
[42] 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.
[43] Ben Green and Lily Hu. 2018. The myth in the methodology: Towards a recontextualization of fairness in machine learning. *2018 International Conference on Machine Learning* (2018).
[44] 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.
[45] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.
[46] HireVue.com. 2018. Video interview software for recruiting & hiring. https://www.hirevue.com/. Accessed: 2018-08-28.
[47] 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 8th International Conference on Learning Analytics and Knowledge*. ACM, 79–88.
[48] 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 & Knowledge Conference*. ACM, 257–266.
[49] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. 2018. Student learning benefits of a mixed-reality teacher awareness tool in AI-enhanced classrooms. In *International Conference on Artificial Intelligence in Education*. Springer, 154–168.
[50] Karen Holtzblatt and Sandra Jones. 1993. Contextual inquiry: A participatory technique for system design. *Participatory design: Principles and practices* (1993), 177–210.
[51] AI Now Institute. 2018. AI Now Institute. https://ainowinstitute.org. Accessed: 2018-08-03.
[52] Alankar Jain, Florian Pecune, Yoichi Matsuyama, and Justine Cassell. (in press). A user simulator architecture for socially-aware conversational agents. In *Proceedings of ACM International Conference on Intelligent Virtual Agents*. ACM.
[53] David Janzen and Hossein Saiedian. 2005. Test-driven development concepts, taxonomy, and future direction. *Computer* 38, 9 (2005), 43–50.
[54] 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*. ACM, 2001–2013.
[55] Nathan Kallus and Angela Zhou. 2018. Residual unfairness in fair machine learning from prejudiced data. *arXiv preprint arXiv:1806.02887* (2018).
[56] Ece Kamar. 2016. Directions in hybrid intelligence: Complementing AI systems with human intelligence.. In *IJCAI*. 4070–4073.
[57] 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*.
[58] Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 3819–3828.
[59] 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*. ACM, 174.
[60] Niki Kilbertus, Adrià Gascón, Matt J Kusner, Michael Veale, Krishna P Gummadi, and Adrian Weller. 2018. Blind justice: Fairness with encrypted sensitive attributes. *arXiv preprint arXiv:1806.03281* (2018).
[61] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).
[62] 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*. ACM, 3075–3084.
[63] 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*. ACM, 126–137.

[64] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*. 4066–4076.
[65] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)* 9 (2016).
[66] 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.
[67] Min Kyung Lee and Su Baykal. 2017. Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *CSCW*. 1035–1048.
[68] Anqi Liu, Lev Reyzin, and Brian D Ziebart. 2015. Shift-pessimistic active learning using robust bias-Aware prediction. In *AAAI*. 2764–2770.
[69] Hugo Liu and Push Singh. 2002. MAKEBELIEVE: Using commonsense knowledge to generate stories. In *AAAI/IAAI*. 957–958.
[70] Natasha Lomas. 2018. Accenture wants to beat unfair AI with a professional toolkit. https://techcrunch.com/2018/06/09/accenture-wants-to-beat-unfair-ai-with-a-professional-toolkit/. Accessed: 2018-06-14.
[71] Kristian Lum and William Isaac. 2016. To predict and serve? *Significance* 13, 5 (2016), 14–19.
[72] 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.
[73] 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 *EDM*. 151–158.
[74] 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.
[75] MURAL. 2018. MURAL - Make remote design work. https://mural.co/. Accessed: 2018-08-02.
[76] Arvind Narayanan. 2018. 21 fairness definitions and their politics. *FAT* 2018* (2018).
[77] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How search engines reinforce racism.* NYU Press.
[78] Besmira Nushi, Ece Kamar, Eric Horvitz, and Donald Kossmann. 2017. On human intellect and machine failures: Troubleshooting integrative machine learning systems. In *AAAI*. 1017–1025.
[79] 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.
[80] Partnership on AI. 2018. The Partnership on AI. https://www.partnershiponai.org. Accessed: 2018-09-03.
[81] Julia Powles and Hal Hodson. 2017. Google DeepMind and healthcare in an age of algorithms. *Health and technology* 7, 4 (2017), 351–367.
[82] pymetrics. 2018. matching talent to opportunity. https://www.pymetrics.com/. Accessed: 2018-08-28.
[83] Qualtrics. 2013. Qualtrics. *Provo, UT, USA* (2013).
[84] Emilee Rader and Rebecca Gray. 2015. Understanding user beliefs about algorithmic curation in the Facebook news feed. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM, 173–182.
[85] Manish Raghavan, Aleksandrs Slivkins, Jennifer Wortman Vaughan, and Zhiwei Steven Wu. 2018. The externalities of exploration and how data diversity helps exploitation. *arXiv preprint arXiv:1806.00543* (2018).
[86] Dillon Reisman, Jason Schultz, K Crawford, and M Whittaker. 2018. Algorithmic impact assessments: A practical framework for public agency accountability.
[87] 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*. 2503–2511.
[88] Data & Society. 2018. Algorithmic Accountability: A primer. *Data & Society* (2018).
[89] 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*.
[90] Donald E Stokes. 1997. *Pasteur's quadrant: Basic science and technological innovation.* Brookings Institution Press.
[91] 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.
[92] Jennifer Wortman Vaughan. 2017. *Making better use of the crowd.* Technical Report. Working paper.
[93] 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.
[94] 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*. ACM, 440.
[95] 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.
[96] 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*. ACM, 656.
[97] Qian Yang. 2018. Machine learning as a UX design material: How can we imagine beyond automation, recommenders, and reminders?. In *AAAI Spring Symposium Series*.
[98] 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 on Designing Interactive Systems Conference 2018*. ACM, 573–584.
[99] 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*. ACM, 4477–4488.
[100] 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.
[101] 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 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2265–2267.