Opportunities for Machine Learning Research to Support Fairness in Industry Practice

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

A surge of recent research has focused on the development of algorithmic tools to assess and mitigate unfairness in ML systems. If such tools are to have a positive impact on industry practice, it is critical that their design be informed by an understanding of ML practitioners' actual needs. In this work, we conduct the first systematic investigation of needs and challenges of industry practitioners around fair ML. Through semi-structured interviews and a survey, we identify disconnects between the problems commonly studied in fair ML research and those faced by industry practitioners, and identify fruitful directions for future research.

1 Introduction

Substantial effort in the growing literature on fairness in machine learning has been devoted to the development of mathematical definitions of "fairness" [4, 6, 16, 24, 26, 30] and algorithmic tools to assess and mitigate undesirable biases in relation to these definitions [1, 8, 10, 13, 16]. If research in fairness is to have a positive impact on industry practice, however, it is critical that research agendas are aligned with ML practitioners' actual needs [18, 35, 38]. Despite widespread attention on fairness and bias in ML, to the best of our knowledge, only one prior study, by Veale et al. [35], has investigated actual decision makers' challenges and needs around fairness, focusing on algorithm-assisted decision-making in the public sector. In this work, we conduct the first systematic investigation of challenges faced by industry product teams in creating fair ML systems. Through a series of semi-structured interviews and an anonymous survey, we identify several disconnects between the actual challenges faced by industry practitioners and those addressed in the research literature on fair machine learning. We identify a range of open problems for the fair ML research community to better support fairness in practice.

2 Methods

To better understand ML product teams' needs for support around fairness, we conducted a series of semi-structured interviews with 35 industry ML practitioners across 25 product teams from 10 companies (Table 1). Interview participants were recruited through snowball sampling. We began by emailing direct contacts across over 30 major companies as well as members of product

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

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

Technology Area	Roles of Participants	Participant IDs
Adaptive Tutoring & Mentoring	Chief Data Scientist, CTO, Data Scientist, Research Scientist	R10, [R13, R14], R30
Chatbots	CEO, Product Manager, UX Researcher	[R17, R18], R35
Vision & Multimodal Sensing	CTO, ML Engineer, Product Manager, Software Engineer	[R2, R3, R4], R6, R7, R9, R26
General-purpose ML (e.g., APIs)	Chief Architect, Director of ML, Product Manager	R25, R32, R34
NLP (e.g., Speech, Translation)	Data Manager, Data Collector, Domain Expert, ML Engineer, PM, 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 Manager	R5, R11

teams whose products had received relevant media coverage related to fairness. In both cases, we encouraged contacts to share our invitation with colleagues working on ML products (in any role).We encountered several challenges in recruiting similar to those detailed by Veale et al. [35]. For instance, our contacts often expressed a general distrust of researchers, citing cases where researchers benefited by publicly critiquing companies' products instead of engaging with them to improve their products (e.g., [2, 32, 36, 39]). We assured contacts that our goal was to understand teams' needs around fair ML, and that we would not link findings to specific companies. Furthermore, we noted that audio recordings of interviews would be destroyed following transcription and that interviewees would have the chance to review any (de-identified) quotes from their interviews as a pre-condition to inclusion in any publications.

Each interview lasted 30–60 minutes. Participants were asked a series of questions about fairness at each stage of their teams' ML development pipelines—collecting data, designing datasets (e.g., training and test), developing an ML product, and detecting and potentially addressing fairness issues in that product. For each stage, participants were asked about critical episodes their teams had encountered (e.g., specific times their teams had found fairness issues). First, they were asked to walk though how their teams navigated these episodes. Then they 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 they wanted, to help them in the moment (cf. [18, 22]). This format was meant to encourage participants to speak freely about their challenges and needs, without feeling constrained to those for which they believed a (technical) solution was currently possible [15, 18, 22]. Once participants had generated questions for the oracle, they were asked, for each question, how their team currently goes about trying to answer these questions in the absence of such an oracle.

To analyze the interview data, we worked through transcriptions of approximately 25 hours of audio recordings to synthesize findings using two standard techniques from contextual design: interpretation sessions and affinity diagramming [15, 19]. Using a bottom-up, affinity diagramming approach, we iteratively generated topics encoding the statements provided by interviewees, and grouped these codes into successively higher-level themes.

To investigate the prevalence and generality of themes surfaced in our interview study, we then conducted an anonymous online survey with a broader sample of 267 industry ML practitioners (Figure 1). We structured the survey as a quantitative supplement to our interview study: the survey's structure directly mirrored the interview. Participants were recruited via snowball sampling, conducted via social media, ML-focused online communities, and direct contacts at major companies. Both the interview procedure and the survey went through ethical review and were IRB-approved.

3 Results

In the following, we provide examples of themes that emerged through our affinity diagramming, highlighting research and design opportunities that have received relatively little attention in the fair ML research literature thus far. The themes are illustrated with direct quotes from interviews (where we received explicit permissions) and supplemented with corresponding survey results.[1]

[1]Our survey uses branching logic (e.g., respondents are only asked questions about addressing fairness issues if they report their teams have previously found such issues), so some questions are only completed by a relevant subset of respondents. In such cases, question-specific response totals are provided in addition to percentages.

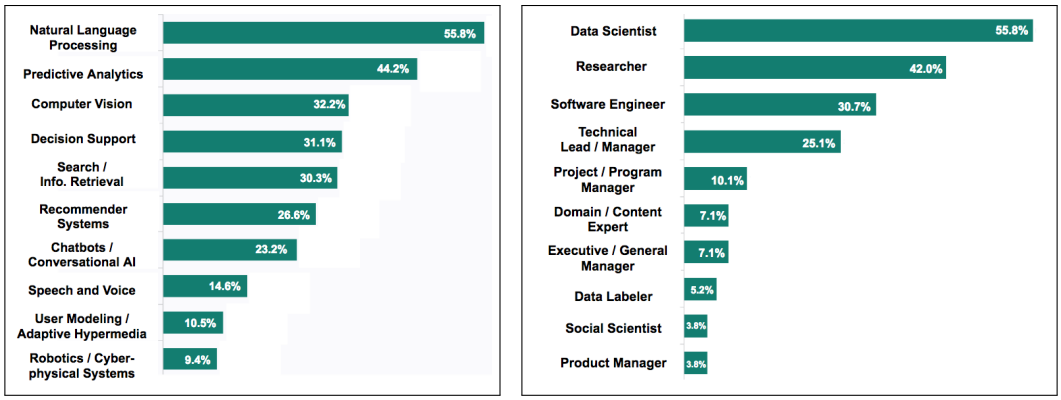

Figure 1: Profiles of survey participants: the top 10 reported application domains (left) and team roles (right). Participants were allowed to check multiple options, so percentages do not add to 100%.

3.1 Fairness-aware Data Collection and Curation

While popular press articles on fairness in ML often use a "bias in, bias out" framing, emphasizing the central role of dataset quality, the fair ML research literature has overwhelmingly focused on the development of algorithmic methods that attempt to correct for such biases, assuming little or no agency over data collection and curation. However, several of our interviewees and survey respondents reported that their teams typically look to their training datasets, not their ML models, as the most important place to intervene. 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, of 55 respondents whose teams' had previously tried to address fairness issues in their products, the most commonly attempted strategy (73%) was "collecting more training data." Interviewees highlighted needs for tools and processes that can guide data collection to help teams mitigate detected fairness issues, or even avoid such issues in the first place [11, 28]. For example, an ML engineer (R19; see Table 1) working on automated essay scoring noted that

> *"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? [...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% indicated that such active guidance would be "Very" or "Extremely" useful (on a five-point Likert scale from "Not at all" to "Extremely" useful). In addition, several interviewees stressed the importance of careful test set design in detecting potential fairness issues, raising needs for better tooling, such as tools to facilitate rapid dataset annotation and use of annotations for fairness auditing, or mechanisms to share test cases that encode nuanced cultural/domain knowledge across teams and companies [37]. Of 187 surveyed, 66% indicated that such tooling would be at least "Very" useful.

3.2 Challenges Due to Blind Spots

Many of interviewees' concerns revolved around their team's potential blind spots. Several interviewees highlighted needs for support in identifying which subpopulations their team needs to consider. 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 domain. R32 emphasized:

> *"[although people tend to] start thinking about attributes like [ethnicity and gender], the biggest problem I found is that these [subpopulations] should be defined based on the domain and problem. For example, for [automated writing evaluation] maybe [the subpopulations] should be defined based on [...whether the person is] a native speaker."*

(See also [14].) 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. A few interviewees also shared experiences in which efforts to obtain additional training data were hampered

by their teams' cultural blind spots. R4 recalled cases in which customers had complained that a globally deployed image captioning system performed well for celebrities from only some countries:

> *"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."*

3.3 Limitations of Existing ML Fairness Methods

Much of the existing fair ML literature has focused on applications such as recidivism prediction, automated hiring, and face recognition where "fairness" can be understood, at least partially, in terms of well-defined parity metrics. However, teams working on applications involving richer, more complex interactions between the user and the system—such as chatbots, web search, and adaptive tutoring—often reported struggling to apply existing ML fairness methods in their contexts. These interviewees brought up needs for more holistic, system-level auditing methods, such as ones that simulate the causal impacts of a system's use in the real world [7, 9, 31]. In addition, several interviewees shared experiences where, after making changes to models or datasets to improve some aspect of fairness, their system changed in subtle, unexpected ways that harmed the user experience (UX). Absent tools and testing methods to help teams anticipate such side effects, many teams reported erring on the side of caution: implementing highly specialized band-aid fixes such as censoring/suppressing specific model outputs. However, such "band-aids" can be very brittle, and can sometimes even create *new* fairness issues. Of those surveyed, 71% indicated that it would be at least "Very" useful to have better tools to understand potential UX side effects of a given fairness intervention. Together, these observations point to a need for new testing and prototyping methods for complex ML systems [5, 12, 17] that can effectively surface unfair behaviors pre-deployment.

Furthermore, whereas most fairness auditing methods proposed in the literature assume access to sensitive attributes (such as race, age, or gender) at an individual level, in practice such information is available only at coarse-grained levels, if at all [23, 34]. Thus, our interviewees often desired methods to reliably audit for unfairness given only coarse information on sensitive attributes. For example, 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."* Out 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.

3.4 Biases in the Humans in the Loop

Several interviewees stressed the importance of explicitly considering biases that may be present in the *humans* embedded at various stages of the ML development and maintenance pipeline, such as crowdworkers who annotate training data or study participants involved in surfacing undesirable biases in ML systems [33]. Of 210 surveyed, 69% indicated that tools to reduce the influence of unfair human biases on their labeling/scoring processes (cf. [21]) would be at least "Very" useful.

4 Open Problems

Even when practitioners are motivated to improve fairness, they face barriers. It is urgent that research agendas are aligned with real-world needs. We conclude with several directions for future research:

- While the existing fair ML literature has overwhelmingly focused on algorithmic "de-biasing," future research should support practitioners in collecting and/or curating representative datasets in the first place (cf. [3, 20, 11])—for example, by developing methods to actively guide data collection processes (cf. [28]), by designing tools and processes to support more effective communication between data collectors and modelers (cf. [35]), or by developing methods to better understand (and correct for) biases in human/crowd labeling processes (see [21, 29, 33]).
- Given that "fairness" can be highly context- and application-specific [14, 27], there is urgent need for domain-specific educational resources, workflows, measures, and tools.
- Future research should explore methods to support effective fairness auditing given only partial information about individual demographics (e.g., neighborhood or school level statistics).

- Another rich area for future research is the development of tools for "fairness debugging" [10, 13]. For example, it can be challenging to determine whether isolated observations of unfairness are "one-offs" or indicative of systemic problems that might require deeper investigation.
- Finally, participants highlighted needs for support in efficiently diagnosing the cause(s) of particular unfair behaviors (cf. [25]), to help their teams decide whether to focus their efforts on the data versus the model [3, 13], or on specific model components in multi-component systems [31, 25].

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453* (2018).

[2] 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.

[3] Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why is my classifier discriminatory? *arXiv preprint arXiv:1805.12002* (2018).

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

[5] 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.

[6] 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.

[7] Cynthia Dwork and Christina Ilvento. 2018. Group fairness under composition. (2018).

[8] Data Science for Social Good. 2018. Aequitas: Bias and fairness audit toolkit. (2018). http://aequitas.dssg.io. Accessed: 2018-08-29.

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

[10] 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.

[11] 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).

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

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

[14] 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).

[15] 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.

[16] Moritz Hardt, Eric Price, Nati Srebro, and others. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[17] 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.

[18] 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.

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

[20] Nathan Kallus and Angela Zhou. 2018. Residual unfairness in fair machine learning from prejudiced data. *arXiv preprint arXiv:1806.02887* (2018).

[21] 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*.

[22] 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.

[23] 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).

[24] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[25] 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.

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

[27] 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.

[28] Anqi Liu, Lev Reyzin, and Brian D Ziebart. 2015. Shift-pessimistic active learning using robust bias-Aware prediction. In *AAAI*. 2764–2770.

[29] 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.

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

[31] 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.

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

[33] Jennifer Wortman Vaughan. 2017. *Making better use of the crowd*. Technical Report. Working paper.

[34] 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.

[35] 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.

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

[37] 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.

[38] 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.

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