Opportunities for Machine Learning Research to Support Fairness in Industry Practice

Kenneth Holstein [1], Jennifer Wortman Vaughan [2], Hal Daumé III [2,3], Miroslav Dudík [2], Hanna Wallach [2]

[1] Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA

[2] Microsoft Research
New York, NY

[3] Departments of Computer Science & Language Science
University of Maryland
College Park, MD

1 Overview

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.

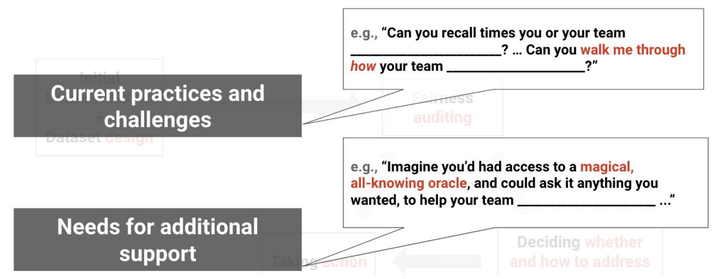
In this work, we conduct the first systematic investigation of industry practitioners' needs and challenges around fairness.

2 Methods

Semi-structured interviews

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 major technology companies.

- Interview participants were recruited through snowball sampling; we emailed direct contacts across over 30 major companies, as well as members of product teams whose products had received fairness/bias-related media coverage. 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 in prior work by Veale et al. (2018). For example, 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. We took various measures, beyond those typical of such studies, to preserve interviewee/team anonymity and promote trust.
- 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)...



We analyzed the interview data using two standard techniques from contextual design: interpretation sessions and affinity diagramming.

- 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.

Anonymous online survey

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.

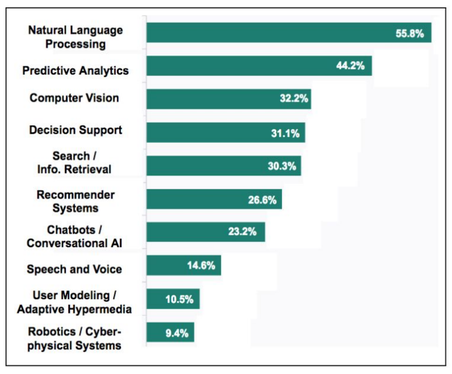
- We structured the survey as a quantitative supplement to our interview study: the survey's structure directly mirrored the interview, using branching logic.
- Participants were recruited via snowball sampling, conducted via social media, ML-focused online communities, and direct contacts at major companies.
- Both the interview procedure and the survey went through ethical review and were IRB-approved.

3 Results

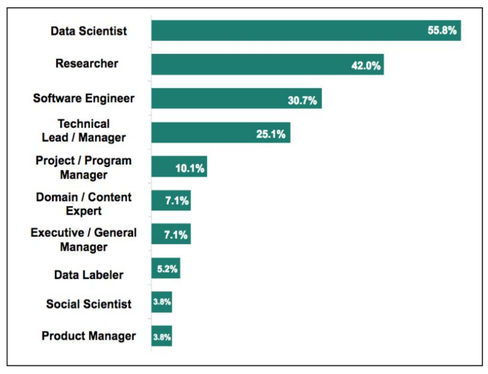
Interviewees' self-identified technology areas and team roles

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

Profiles of survey participants



Application domains



Team roles

Highlighted interview excerpts

"[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." - R32

"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." - R4

"If we had more people who we could throw at this... 'Can we leverage this fuzzy [coarse-grained] data to [audit]?' that would be great [...] It's a fairly intimidating research problem I think, for us." - R21

"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." - R19

Highlighted survey results

Survey respondents* indicated strong needs for better tools to...

Understand potential user experience (UX) side effects of a given fairness intervention
71% of 267

Support fairness testing without individual-level demographics
70% of 183

Reduce the influence of unfair human biases on their labeling/scoring processes
69% of 210

Scaffold fairness-aware test set design (e.g., rapid dataset annotation, sharing test cases)
66% of 187

Support teams in identifying relevant subpopulations in the first place, for specific kinds of ML applications
62% of 213

Actively guide data collection and curation processes, towards fairer outcomes
60% of 67

* Survey uses branching logic (e.g., respondents are only asked questions about addressing fairness issues if they report their teams have previously found issues), so some questions are only completed by a relevant subset of respondents.

4 Open problems

Our findings reveal disconnects between the kinds of problems commonly studied in fair ML research and those faced by practitioners.

Within these disconnects, we see fruitful directions for future research:

- **While the existing fair ML literature has overwhelmingly focused on algorithmic "de-biasing," future research should support practitioners in fairness-aware data collection & curation.**
  for example... by developing methods to actively guide data collection processes, by designing tools and processes to support more effective communication between data collectors and modelers, or by developing methods to better understand (and correct for) biases in human/crowd labeling processes
- **Given that "fairness" can be highly context- and application-specific, there is urgent need for domain-specific educational resources, workflows, measures, and tools.**
  teams working on applications involving richer, more complex interactions between the user and the system—such as chatbots, web search, and adaptive tutoring/mentoring technologies often reported struggling to apply existing FAT/ML methods in their contexts.
- **Future research should explore methods to support effective fairness auditing given only partial information about individual demographics (e.g., neighborhood or school level statistics).**
  whereas most fairness auditing methods proposed in the literature assume access to protected attributes at an individual level, such information is frequently available only at coarser levels in practice, if at all
- **Another rich area for future research is the development of tools for "fairness debugging".**
  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;
  participants also highlighted needs for support in efficiently diagnosing the cause(s) of particular unfair behaviors, to help their teams decide whether to focus efforts on the data vs the model, or on specific components in multi-component ML systems.
- **Beyond automated monitoring tools, our findings indicate needs for new prototyping & testing approaches to effectively surface unfair behaviors before ML systems are deployed in the real world.**
  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, and the space of possible contexts is often very large