Uncovering the Practices and Opportunities for Cross-functional Collaboration around AI Fairness in Industry Practice

Wesley Hanwen Deng
hanwend@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Nur Yildirim
yildirim@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Monica Chang
monicach@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Motahhare Eslami
meslami@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

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

Michael Madaio
madaiom@google.com
Google Research
New York, New York, USA

ABSTRACT

An emerging body of research indicates that ineffective cross-functional collaboration – the interdisciplinary work done by industry practitioners across roles – represents a major barrier to addressing issues of fairness in AI design and development. In this research, we sought to better understand practitioners' current practices and tactics to enact cross-functional collaboration for AI fairness, in order to identify opportunities to support more effective collaboration. We conducted a series of interviews and design workshops with 23 industry practitioners spanning various roles from 17 companies. We found that practitioners engaged in *bridging* work to overcome frictions in understanding, contextualization, and evaluation around AI fairness across roles. In addition, in organizational contexts with a lack of resources and incentives for fairness work, practitioners often *piggybacked* on existing requirements (e.g., for privacy assessments) and AI development norms (e.g., the use of quantitative evaluation metrics), although they worry that these tactics may be fundamentally compromised. Finally, we draw attention to the *invisible labor* that practitioners take on as part of this bridging and piggybacking work to enact interdisciplinary collaboration for fairness. We close by discussing opportunities for both FAccT researchers and AI practitioners to better support cross-functional collaboration for fairness in the design and development of AI systems.

KEYWORDS

fairness, collaboration, interdisciplinarity

ACM Reference Format:
Wesley Hanwen Deng, Nur Yildirim, Monica Chang, Motahhare Eslami, Kenneth Holstein, and Michael Madaio. 2023. Uncovering the Practices and Opportunities for Cross-functional Collaboration around AI Fairness in Industry Practice. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, US.* ACM, New York, NY, USA, 12 pages. https://doi.org/XXXXXXX.XXXXXXX

FAccT '23, June 12–15, 2023, Chicago, US

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Addressing unfairness in AI systems is a fundamentally socio-technical challenge, requiring the integration of skills and expertise across many different areas, including approaches from the social sciences to understand what (un)fairness means for particular communities and sociocultural contexts, approaches from user research and design to understand (un)fairness in particular use cases and domains, as well as technical skills to address unfairness in AI system design [29, 31, 33]. Yet a growing body of literature suggests that such integration of skills and expertise that comes in the format of *cross-functional collaboration* [35, 41, 67] — a term used in industry settings to describe collaboration among diverse roles with various disciplinary backgrounds [cf. 59] — is often absent or ineffective in industry AI fairness work [2, 24, 47, 52, 55, 56, 78, 79]. Prior work has identified several challenges around cross-functional collaboration for AI fairness such as differing awareness of AI fairness across roles [2, 52, 79], mismatched expectations for measuring fairness [23, 47, 56], and an absence of tools to support collaborative work across roles [2, 24, 76]. These challenges have been shown to hinder teams in effectively addressing fairness issues [2, 23, 47, 52, 79]. However, little is known regarding (1) whether and how industry practitioners navigate such challenges to overcome collaboration barriers, and (2) what opportunities exist to improve cross-functional collaboration around fairness in AI?

To investigate, we conducted a two-stage study with 23 industry AI practitioners spanning 17 companies and various roles (e.g., data scientists, UX practitioners, product managers, and subject matter experts), who have previously worked with other roles in their company to tackle fairness-related issues. In our study, we first conducted semi-structured interviews to understand practitioners' current practices and challenges around cross-functional collaboration. We then conducted workshops to bring together participants across multiple roles and companies, to better understand common practices and envision future opportunities for creating more effective cross-collaboration in AI fairness.

We found that practitioners go beyond their current job descriptions to undertake a range of "bridging" roles, aimed at fostering shared understandings around AI fairness, translating and contextualizing abstract fairness concepts for other roles, and aligning expectations around fairness evaluation across different roles (Section 4.1). In order to effectively collaborate in organizations constrained by a lack of resources or incentives for AI fairness work

[cf. 49, 59], practitioners also adopted "piggybacking"[1] tactics to facilitate collaboration around AI fairness, although they worried the use of such tactics might compromise their values in the long term (Section 4.2). Furthermore, participants shared frustration around the often invisible labor involved in supporting cross-functional collaborations (Section 4.3).

Building on an understanding of these bridging and piggybacking practices and the underlying collaboration challenges they are intended to address, we discuss opportunities for both researchers and industry AI teams to better support cross-functional collaboration for AI fairness work in industry practice (Section 6). In particular, we discuss tools and processes that could support "bridging" work and shed light on ways to make invisible labor in collaborations around AI fairness more visible to (and ideally valued by) teams and organizations.

Overall, this paper contributes an in-depth understanding of industry practitioners' current practices to facilitate cross-functional collaboration across roles and organizations, identifying "bridging" work and "piggybacking" as two major approaches. These practices serve as a starting point for practitioners to navigate cross-functional collaboration challenges in AI fairness work. In addition, we offer concrete design implications for FAccT researchers, practitioners, and organizations to better support cross-functional collaboration in AI fairness work.

2 BACKGROUND AND RELATED WORK

A growing body of research has empirically investigated industry practitioners' current practices and challenges in addressing issues of fairness *in practice*, during the design and development of AI systems (e.g., [21, 23, 24, 36, 42, 45, 47, 49, 55, 56, 59, 60, 66, 73]). Among other findings, this work has suggested that effectively tackling socio-technical challenges like AI fairness requires substantial interdisciplinary collaboration among multiple roles [24, 36, 49, 55, 59]. For example, based on fieldwork with a corporate data science team, Passi and Jackson highlighted that in order to build more responsible and trustworthy AI systems, data scientists must engage and negotiate with business and product teams throughout the AI developing lifecycle [56]. Rakova et al. suggested that addressing AI fairness issues requires AI developers to better understand the needs of stakeholders from different backgrounds (e.g., domain experts) [59].

However, despite the central role cross-functional collaboration plays in responsible AI development, prior research suggests cross-functional collaboration is often absent or ineffective in industry AI fairness work. To start with, prior work has surfaced that the metrics and methods used to evaluate and address AI fairness issues can vary, or might even be incommensurate, across disciplines. In practice, this can lead to communication breakdowns and ineffective collaboration around AI fairness [24, 38, 47, 55, 56]. For example, Passi and Barocas found that misalignments around problem formulation between data scientists and business teams can contribute to fundamental fairness issues from the early problem formulation phases of a project [55]. Madaio et al. found that practitioners across roles often defaulted to using their existing performance metrics for assessing the aggregate performance of their AI systems, when those metrics may not be best suited to identifying disparities in models' performance for *dis*aggregated subgroups [47]. Currently, there is a lack of processes for practitioners across disciplines to collectively decide on appropriate metrics to assess disparities in model performance.

Relatedly, existing tools and processes designed for tackling AI fairness issues (e.g., [1, 9, 16, 17, 28]) are largely designed to be used by technical roles working in isolation, and are not designed to support collaboration across roles [2, 24, 45, 76, 79]. For example, in their work studying how AI practitioners use fairness toolkits, Deng et al. identified that data scientists lack efficient tools and processes to translate and incorporate domain experts' knowledge into fairness analyses [24]. When studying how teams communicate about and evaluate the quality of ML models, Almahmoud et al. found that practitioners felt current tools made it challenging to have cross-functional conversations around fairness, as these tools are often tailored to evaluate and report model qualities like accuracy rather than broader sociotechnical concepts such as fairness [2].

Finally, prior literature has found that not all relevant roles in AI teams are incentivized or invited to collaborate around AI fairness [52, 79]. For example, Zhang et al. found that industry AI teams often treated fairness work primarily as a technical matter. As a consequence, roles with the most relevant domain, legal, policy, or lived expertise were often left out of the process [79]. While prior work has discussed challenges to collaboration around fairness, in the current work we seek to understand what industry AI practitioners currently do to support and enable cross-functional collaboration for AI fairness, with the goal of identifying opportunities to better support these efforts.

3 METHODS

To investigate our research questions, we conducted a two-stage study with 23 industry practitioners across 7 roles, involving semi-structured interviews followed by group workshops. We first conducted semi-structured interviews to understand participants' current practices and challenges around cross-functional collaboration in AI fairness work. In the next stage, we invited participants spanning different roles and organizations to workshops to collaboratively explore opportunities to better support cross-functional collaboration around AI fairness.

3.1 Participants

We adopted a purposive sampling approach [18], with the aim of recruiting industry AI practitioners from diverse roles (e.g., data scientists, UX researchers, product managers, etc) who have worked on addressing AI fairness in their role. During the study, all participants self-reported that they had experience collaborating with other roles in their work on AI fairness. Table 1 provides an overview of participants' product areas and roles. Throughout the paper, we use [Acronym of the role][Role number] to refer to our participants.

[1] Similar to prior work exploring supporting environmental sustainability and social justice work in industry settings [cf. 14, 15, 61], We use "piggybacking" to refer to the process of *"identifying potential allies with similar or overlapping interests, and utilize and 'piggyback' on existing organizational resources and programs as much as possible"* [14].

We recruited our participants through direct contacts at large technology companies, through recruitment posts on social media (e.g., Linkedin and Twitter), and snowball sampling from those participants. In total, 25 practitioners completed the recruitment screening form for our interview, of whom 18 met our recruitment criteria, responded to our interview study invitation, and participated in the study. In the end, 17 participants completed the interview study. We invited all interview participants to attend sessions in the first round of the workshops. six out of 17 interview participants responded to our invitation and joined the workshop (W1 [DS2, DS4, UX5], W2 [DS1, MLE1, DS5]). We sent another round of workshop recruitment through a purposive sampling approach similar to the previous round (e.g., direct contacts and social media) and received 42 responses. In the end, 19 people responded to our scheduling email, with seven of those able to participate in our second round of workshops (W3 [PM3, UX6], W4 [SME1, DS6], W5 [DS1, SWE2, SME2]). Similar to prior work studying AI fairness with industry practitioners [24, 45, 60], we encountered a large drop-out rate, potentially due to the sensitivity of the topic of AI fairness. In addition, scheduling the workshops with participants from different roles and companies was constrained by practitioners' joint availability, further adding challenges to participation in the study.

All participants were compensated at a rate of $35 per hour for their participation. In addition, for both interviews and workshops, we told participants that we would not ask them to reveal any confidential or personally identifying information about their colleagues and that we would anonymize all responses at the individual, team, and organization levels. Finally, participants were told that they were free to skip any questions they were uncomfortable answering, and were free to leave the workshop session at any time for any reason.

3.2 Study Design

3.2.1 Stage one: semi-structured interviews. To understand practitioners' current practices around cross-functional collaboration for fairness in AI, we conducted 17 individual semi-structured interviews with practitioners from diverse roles from 17 organizations, all of whom had some experience working on AI fairness. We adopted a directed storytelling approach [27, 32] in the interviews, each of which lasted roughly an hour. We first asked participants to describe their current practices around cross-functional collaboration in AI fairness, with a specific focus on artifacts, tools or resources their team used as part of that work. For example, we asked participants *"What tools have you been using to collaborate with other roles on your team around AI fairness?"* We probed deeper into challenges participants had encountered by asking follow-up questions like *"Were there disagreements or tensions between people from different disciplinary backgrounds?"* As participants shared specific collaboration challenges they had encountered, we invited them to share specific activities or strategies they adopted to address those, by asking questions like *"How did your team attempt to tackle these challenges?"* and *"How effective were your team's approaches?"*

3.2.2 Stage two: workshops. In the second stage, we conducted an iterative series of five workshops to create a space for multiple industry practitioners from *different roles, teams, or organizations* to share their experiences, discuss how their experiences related to other participants, and identify opportunities for improving cross-collaboration for AI fairness. 12 participants attended the workshops, including six participants from the interviews and six new participants. For the first two workshops (W1, W2), we recruited participants from the interviews in stage one and asked them to discuss preliminary findings from the interview study, to validate these preliminary results and share additional context. We then had participants conduct a speed boat activity [57]—a common workshop activity used in design workshops for facilitating discussion among multiple stakeholders. In the activity, we shared challenges to collaboration on AI fairness that we identified in our preliminary findings, and asked participants to collectively select one such challenge and imagine themselves to "be on the same boat" with each other to address the challenge. Drawing on the metaphor of a boat at sea, participants identified their goals, tailwinds and headwinds (e.g., enabling or hindering factors), and other aspects of their team or organization that might impact their ability to effectively collaborate on fairness with team members from other disciplines.

After running the first two workshop sessions, we observed that our participants particularly needed to explore the opportunities they saw for ideal cross-functional collaboration. We thus revised our workshop protocol to provide a space for brainstorming desired cross-collaboration opportunities via "journey mapping" [44] —a technique from user experience research that can be used to identify interactions between different stakeholders over time and support participants in moving from the current state to an ideal state. Before joining the workshops, participants watched a short tutorial video we prepared and filled out a journey map defining their teams' phases of AI development and the stakeholders involved in each phase, based on the nature of their team and organization. For each phase they listed, we asked participants to share more details around the work they did in this phase (related to fairness specifically); the other roles and stakeholders that are currently involved (and the roles they wished were more involved); as well as the artifacts or resources they currently used (and that they wished to have); what worked well and the pain points of each phase. We then compiled all participants' journey maps to a Miro board before each workshop session. During the workshop, we spent the first thirty minutes having each participant quickly present their journey maps so that they could learn more about each others' processes. We then used another thirty minutes to have the participants discuss and co-construct an "ideal" journey map using a journey map template we offered, focusing on envisioning the opportunities to better navigate cross-functional collaboration. Note that, similar to prior HCI research using journey maps in their study design (e.g., [44]), our goal for the workshop was not to produce a perfect journey map as an artifact-based contribution of the research, but instead, we used the process of discussing and producing journey maps to elicit participants' perceived challenges, needs, and opportunities, as well as to scaffold the workshop discussions with specific details from participants' roles, teams, and organizations.

Role or Job Titles	Company Employee Number	Gender	Location
Data Scientist (DS) (6)	25,000 and more (9)	Female (7)	US (15)
UX Practitioners (UX) (6)	5,000 - 24,999 (2)	Male (15)	India (2)
Product Manager (PM) (3)	1,000 - 4,999 (3)	Non-binary (1)	Sweden (2)
ML Engineer (MLE) (2)	250 - 999 (0)		Australia (1)
Software Engineer (SWE) (2)	50 - 249 (1)		France (1)
Research Scientist (RS) (2)	10 - 49 (2)		Netherlands (1)
Subject Matter Expert (SME) (2)			Mexico (1)

Table 1: Overall Demographics and Background of our 23 study participants. Next to each demographic information, we include the number of the participants within that demographic group in parenthesis.

3.3 Data Analysis

Our study sessions yielded approximately 23 hours of audio that we transcribed. To analyze our interview and workshop transcripts, we used inductive thematic analysis, a common qualitative data analysis method in HCI [13]. Six of the authors conducted an open coding of a subset of the transcripts, then discussed their codes with the entire research team. After the team reached consensus on the format and granularity of the codes, two authors independently coded all transcripts and reconvened to resolve any major disagreements through discussion. For example, codes include *"MLE1 decided to use their spare time to develop shareable documentations and materials to help other team members learn more about AI fairness."* Then, four of the authors iteratively grouped the codes to identify higher-level themes. For example, one lower-level theme around practitioners' current practices was *"Participants leverage numerical, measurable metrics to translate the impact of fairness to AI developer teams"*, which was later grouped into the higher-level theme of *"Piggybacking on the quantification culture of AI development."* We present key themes from our data in the following Findings sections.

4 FINDINGS

4.1 Bridging Gaps in the Understanding, Contextualization, and Evaluation around AI Fairness to Improve Cross-Functional Collaboration

We found that participants spanning a range of formal roles took on a critical bridging role by identifying and creating opportunities to foster their team's learning about AI fairness, contextualizing abstract concepts and guidelines to make AI fairness work concrete, and aligning mismatched goals and metrics for fairness evaluations. Throughout this section, we highlight how these bridging activities attempt to overcome barriers to cross-functional collaboration around AI fairness.

4.1.1 Bridging the gaps in incompatible disciplinary evaluations around AI fairness. We find that our participants take on bridging work to overcome tensions in the methods or metrics that various disciplines use to assess or measure fairness in AI systems, metrics which may be incompatible with each other [cf. 2, 47]. For example, participants in our study reported that technical roles on their teams (DS, MLE) tend to evaluate their models by *"mainly focusing on the output of the models they built without thinking about how [these models] interact with real customers"* (SWE2). In contrast, user- and product- facing roles (e.g., UX designers and PMs) often have a better understanding of *"shareholders and customers' concerns"* (PM2) but may lack an understanding of how to translate their understanding into effective evaluations of fairness (e.g., in ways that respond to customers' concerns).[2] As a consequence, multiple participants mentioned that in order to facilitate communication about fairness among team members with diverse backgrounds, they needed to bridge the goals for fairness assessments between technical roles focused primarily on the model outputs (i.e., model-focused evaluation) and user- and product- facing roles who tend to focus on biases and harms perceived by users (i.e., user-focused evaluation).

Participants shared that they bridged these evaluation gaps by initiating and organizing meetings for cross-functional teams to align model-focused and user-focused fairness evaluations. For example, PM1 repurposed some of their regular team-level all-hands meetings, which were originally designed for team members to report on their work progress and goals, to *"co-evaluation meetings"* (PM1) for fairness issues. In particular, PM1 spent extra effort designing activities to scaffold technical roles and user-facing roles in understanding and aligning each others' perspectives on evaluating fairness issues: *"I will have the entire team together and have the failure mode effect analysis, we see what's happening with the false positive rate and false negative rate of the model for our use cases and making sure that we always align on this [...] for every single step [in building the model], our team makes sure that there is a fairness requirement from the product team and there is a specific guideline of implementation from the engineers."* (PM1)

UX5 shared that they organized similar *"co-evaluation sessions"* (UX5) in their company, with the purpose to create spaces to understand the differences and similarities between the evaluation metrics for AI fairness that different roles employed. *"In these sessions, I come up with hierarchies and frameworks whilst everybody else was talking talking talking [...] and sharing these artifacts in the moment [...] and showing what are the connections between different evaluations people were just talking about [...] and then people made the connections between the AI [models] and the product and they suddenly started to collaborate because they finally understood each other's goals [...] and started to use similar terminologies."* (UX5)

[2] See Madaio et al. [47] for a discussion of the risks of prioritizing shareholders, customers, or other business-oriented stakeholder groups over marginalized communities who may be most impacted by algorithmic systems.

However, participants shared that, while they attempted to bridge the gaps between model-facing and user-facing evaluations of AI fairness, the culture of AI development (broadly speaking) often prioritizes quantitative over qualitative evaluation approaches [11, 31], making this bridging work around fairness evaluation less effective. In Section 4.2.2, we further expand on practitioners' strategies on navigating the mismatches between quantitative and qualitative evaluation approaches in organizations that disincentivize fairness work.

4.1.2 Creating collaborative processes and spaces to bridge gaps in understanding about AI fairness. We found that the incompatibility between different teams was not limited to fairness evaluation methods and metrics. In fact, most participants shared that there were crucial differences in their understanding of AI fairness in general, which introduced challenges for effective collaboration. Without a common grounding of what AI fairness entails for the specific domain they were working in, *"disagreements in terms of definitions of bias and fairness will affect how people use the [fairness] toolkits and all the downstream collaborations"* (RS1). During the workshop (W1), UX5 told us that the conversations around fairness among their team members *"stayed at a superficial level and went nowhere"* when team members failed to align their own understandings of what "fairness" means with what it means to their users — in the context of their specific application. In another workshop (W5), SME2 shared in their "journey map" activity that they realized some of their coworkers were *"not aware of representational harms* [3] *caused by AI systems at all."* This made them realize the importance of *"check[ing] each others' understanding of [fairness] problems before just diving into the conversation around fixing the problems."*

To bridge these gaps and inconsistencies in AI practitioners' understanding of AI fairness, participants designed various collaborative activities that were intended to scaffold conversations among cross-functional team members. These activities range from small team design workshops to company-wide hackathons. For instance, as a UX designer working on image captioning and product recommendation applications, UX1 held design workshops *"similar to those workshops us designers often conduct with external clients and users"* with their team members working on AI fairness. In these workshops, UX1 led the conversations with their team members working on different aspects of the products to explore what *"a fair product recommendation system mean[s]"* (UX1).

However, participants mentioned that team members, especially those who were new to the topic of AI fairness, often struggled with what and how to discuss. To overcome these communication barriers, participants drew on toolkits to scaffold the design process. For example, UX4 incorporated a toolkit they often used in user research sessions called *"ideation cards"*—a deck of cards including a hundred questions concerning the value and ethics around product design—to facilitate cross-role design workshops and better engage team members in asking questions probing how their AI products might cause fairness issues under different scenarios. UX4 shared that ideation cards helped their team to have constructive debates around the meaning of being fair to different stakeholders who might be affected by the AI service.

To foster participation and engagement in conversations around AI fairness, some participants tailored collaborative activities to the skills and knowledge of specific roles. For example, UX2 and PM2 mentioned hosting company-wide hackathons, a familiar and engaging format for technical roles like engineers, around building more responsible AI applications to better engage engineers in critically examining their understanding of AI fairness. These hackathons often happened *"at the problem formulation stage when the team starts to work on algorithmic fairness or transparency relevant topics"* (UX2), serving as a chance for team members to begin the conversation around fairness issues of their AI products and learn more about each others' perspectives on fairness.

4.1.3 Developing educational resources and documentations to support understanding about fairness. Complementing the efforts for building a common ground between team members about AI fairness, participants also reported developing documentation (in addition to hosting workshops) that aimed to increase their team's knowledge about AI fairness and facilitate conversations about this topic in collaborations. In particular, participants with strong technical backgrounds (DS1, MLE1, RS2, DS6) who report being familiar with the state of the art of technical fair AI research literature shared that they created accessible educational materials to help their team members learn about fundamental AI fairness concepts (e.g., fairness metrics, bias mitigation algorithms, and their limitations). For example, DS6 created a guidebook covering *"common fairness metrics, rationales for some bias mitigation algorithms, and also different types of harms that could be caused by algorithm."* MLE1 worked in a small start-up company offering consulting services for building responsible AI and self-reported being the most knowledgeable team member around AI fairness in their growing engineering team. After repeatedly getting similar questions from colleagues around *"AI fairness concepts and confusions when reading some paper they saw from FAccT,"* MLE1 decided to use their spare time[4] to develop *"shareable documentations and materials"* to provide their team members with a *"free crash course[s]"* that explained basic AI fairness concepts using accessible language. This documentation then became standard onboarding materials for new employees in their company to learn more about fairness in AI.

4.1.4 Translating and contextualizing abstract AI fairness concepts in concrete terms. Participants in our study repeatedly mentioned that publicly available AI fairness guidelines and tutorials (and the AI fairness concepts presented in them) are *"usually too abstract"* (DS4) for tackling the AI fairness issues in their organizations. As a result, when attempting to follow these AI fairness guidelines and tutorials, technical roles often found them *"not always relevant to the task at hand"* (MLE2) and struggled to *"understand which fairness metrics or techniques to use for the specific application"* (DS6). To this end, we observed translation and contextualization work between technical roles and user-facing roles to better understand what fairness means for their specific domain and use case or user populations.

[3] Representational harms are fairness-related harms that involve how groups of people are represented by algorithmic systems, including stereotyping, demeaning, or erasure of particular groups entirely [e.g., 12, 22, 72]

[4] In section 4.3, we discuss the consequences of AI team members needing to use their spare time to develop resources to bridge disciplinary gaps.

Participants in technical roles (DS, SWE) — who are often responsible for directly conducting algorithmic fairness analyses — often proactively initiated collaboration with user-facing roles (PM, customer services) to better contextualize their analysis of the AI models. For instance, as a data scientist, DS5 worked closely with UX researchers to *"build a glossary to contextualize the abstract concept of fairness metrics"* such as equalized odds and demographic parity using real-world scenarios in their AI services for healthcare. Similarly, SWE1 reported that they went beyond assessing model fairness through fairness metrics, trying to contextualize the analysis through conversations with customer service manager. Participants shared that the contextualizing process sensitized data scientists about *"how their model might interact with users in an unintended harmful way"* (DS5), helping technical roles learn more about other roles' perspectives to better inform their fairness analysis.

Furthermore, we found that multiple user-facing roles (UX, PM) often spent extra effort annotating and documenting how to contextualize abstract AI fairness guidelines to their actual practices for other team members, in order to *"get the entire teams on the same page around how the models might cause fairness issue when they are interacting with users"* (PM1). In particular, PM1 annotated the AI fairness guidelines developed by their company (a technology company with over 25,000 employees) with concrete examples and prompts to explore, for instance, *"what does this [guide]line entail for the sentiment analysis product [they] were developing."* During W1, UX5 shared with DS2 and DS4 that they developed *"modules that characterize how different guidelines could be represented back in concrete examples... to help colleagues understand what these [AI fairness] guidelines would mean in an actual solution."* For example, based on their user experience research, UX5 documented the stakeholders who directly interacted with their system — and those who might not directly interact with the systems but who might still be affected — in a shareable document that was available across multiple AI teams in their organization. In the workshop, DS2 and DS4 both agreed with UX5 that these modules they created were extremely valuable to facilitate communications among team members in ways that attend to the real-world context for fairness issues.

4.2 Piggybacking as a Compromised Tactic for Collaboration under Organizational Constraints

In parallel to "bridging," we find that participants employed "piggybacking" as a tactic to push fairness work forward in organizations that might not otherwise provide resources or incentives for fairness work. The "piggybacking" observed in our study took multiple forms: a) some participants piggybacked on related institutional processes, such as privacy impact assessments, to get buy-in for fairness work, b) some also positioned their work within the "quantification" culture of AI development to better communicate with technical roles on their teams. However, when sharing how these piggybacking tactics may have enabled them to conduct fairness work on cross-functional teams, participants also expressed their concerns around the limitations and compromised nature of these tactics.

4.2.1 Piggybacking on institutionalized procedures. To address challenges in AI fairness work, some participants (DS3, PM1, UX4) shared their strategies around piggybacking on organization-wide mandatory privacy-related procedures (e.g., questionnaires, checklists) in order to raise awareness about AI fairness and put AI fairness efforts into practice. For example, DS3 developed a set of checklists to help the AI product teams reflect and assess if their AI features contained potential racial biases against their users, but ran into challenges to incorporate these artifacts into the current day to day AI work. However, since "*team members and leadership are extremely cognizant about privacy*" at DS3's company, the privacy team in the company had already developed a questionnaire called a *"privacy impact assessment,"* containing 10 privacy-related questions for all product teams to complete before launching any AI features or products. DS3 shared stories about how they built allyship with the privacy team and later piggybacked on the *"privacy impact assessment"* to promote their AI fairness effort: *"After multiple times getting rejected by our company to implement our fairness checklists, my teammate had this brilliant idea which was: What if we piggyback onto an existing process that already exists?... all we did was add an extra set of questions specifically concerning machine learning fairness to the privacy impact assessment... the beauty of that is that we don't have to persuade people to fill out this form since they already have to fill out the larger privacy impact assessment in order to launch their products or features."* (DS3)

Furthermore, DS3 told us that adding fairness-related questions to the privacy impact assessment helped to bring the awareness of AI fairness to DS3's entire organization, as there were increasing amount of *"people from other teams reaching out and saying that they want to work on fairness too."* By doing so, participants were also able to build a coalition of team members and develop a network of allies within the organization who were committed to advancing fairness in the company's AI systems. Similarly, during the interview, PM1 brought up the concept of *"change management"* – approaches to prepare and support organizational changes. PM1 told us that *"change management is very, very difficult for all of software practice, but especially with responsible AI, when we need to go the extra mile and explain why this is so important."* Therefore, PM1 would *"always start by adding little things to [their] privacy practice instead of creating an entirely new fairness thing... we don't have to evangelize about how important privacy is."*

However, both DS3 and PM1 shared their desire to implement stand-alone fairness assessment procedures *"eventually when there is more buy-in for AI fairness"* (PM1). DS3 shared that their current efforts around piggybacking on the privacy team might not be sustainable, and they wanted higher-level leadership to allocate more resources for AI fairness work. Furthermore, DS3 brought up the need for further exploring the trade-offs and complementarity between privacy and fairness work instead of smuggling fairness in with privacy assessments.

4.2.2 Piggybacking on the quantification culture of AI development. As mentioned in Section 4.1.1, various disciplines may value different evaluation goals and methods for AI fairness. Current AI development culture still largely favors quantitative over qualitative methodologies and forms of evidence [5, 11, 31]. As a result, many participants reported drawing on quantitative approaches in

order to overcome communication barriers with team members in technical roles. For example, PM1 shared strategies around using a quantitative score to fit their AI fairness work into existing AI work metric systems and the *"numerical culture of AI modeling work"*: *"It is always hard to convince the teams across organizations to accept something new, unless it is something that they are already familiar with [...] so we started using a scale of 1 to 10. If your [model's] feature is not even showing any explainability or analysis around fairness then the score will be low. Engineers, they just don't like low scores"* (PM1). In other words, PM1 used these scores to make the value of working on AI fairness directly relevant to the AI developers on the team. During the workshops, another product manager, PM3, told us that when they are communicating with data scientists, *"using [a fairness toolkit] to calculate numbers like demographic disparity made it much more straightforward for communicating with other data scientists about the impact of working on fairness."* Within the same workshop, UX6 concurred with PM3 by sharing that *"having quantifiable fairness related scores,"* was a commonly used strategy that works well when collaborating with data scientists, and shared with us that they often change their communication strategy to include significant amount of quantitative data when working with data scientists on AI fairness.

However, many participants were aware of the potential pitfalls of using "scores" and "percentages" in AI fairness work. For example, PM2 told us that even though they would always *"make sure to show something quantitative"* when communicating with engineers and data scientists as a *"trick to communicate,"* they don't believe quantitative data is *"necessarily the best way to represent the concept of fairness."* During the workshop (W3), while PM3 shared how relying on quantification dramatically helped them with communicating with the AI development team, they also acknowledged that the numbers produced using fairness toolkits served as an (often inaccurate) proxy of the actual fairness harms that might be caused by their AI products, *"essentially losing a lot of nuances of fairness"* (PM3). As a result, while creating their "ideal journey map," PM3 and others in their workshop shared the desire for better processes to help navigate the current AI development culture around prioritizing quantification to appropriately address socio-technical challenges that may require drawing on and integrating both quantitative and qualitative evidence of (un)fairness.

4.3 Invisible Labor and Burden in AI Fairness Cross-functional Collaboration

While working towards more effective cross-functional collaboration, participants shared that their efforts during the collaborations were often invisible to their team members or leadership. Worse, sometimes other team members would undermine their efforts around AI fairness when collaborating, creating additional burden and frustration for participants.

Participants reported that one reason their "bridging" and "piggybacking" efforts encountered difficulties was that other team members did not always understand what AI fairness work actually involves. This contributed to under-recognition and unrealistic expectations. For example, during the workshop (W2), DS1, MLE1 and DS5, all from different organizations, had discussed how creating educational documentation or accessible visualizations to bridge the knowledge gaps among participants required substantial effort. MLE1 mentioned that updating the educational documentation to keep up with the state of art technical AI fairness research was also extremely time consuming, sometimes requiring several rounds of major iteration even within a week. The value of such extra efforts, however, were usually overlooked. As DS5 concurred with DS1 and MLE1 during the workshop session, *"teams don't understand the amount of work that goes into doing AI fairness work. It's not quite as simple as pulling from a table and doing a statistical analysis. You have to put thoughts into making intentional analysis choices and think broadly about the socio-technical nature of AI fairness.* (DS5)

Furthermore, participants shared that other team members often held unrealistic expectations that collaboration around fairness would be a one-off effort, rather than an iterative, thoughtful process requiring long-term engagement. For example, UX5 shared that some colleagues expected design workshops around AI fairness (see 4.1.2) to be a *"one time thing"* rather than a recurring activity, asking *"did[n't] we already do that last time and get the answer?"* In addition, since AI fairness efforts were not reflected in most organizations' KPIs, individuals who were motivated to do fairness work often felt responsible for *"holding the entire team accountable for the fairness work"* (MLE2), rather than having that accountability located in, for instance, "ethics owners" [50] or other organizational leadership. Similarly, MLE1 shared that they felt *"frustrated by the attitude of [others] dodging responsibility for fairness work."* However, as a woman of color, MLE1 reported that they often felt uncomfortable *"giv[ing] everyone a lecture on why fairness should be prioritized."* As a result, "bridging" work often falls onto a small number of team members who are self-motivated to do the work, yet who are not incentivized or supported in sustaining their efforts long term. As RS2 shared, *"engineers and data scientists just kind of leave the conversation behind, walking out of the meetings [...], and I have to spend more time to do most of the work to make sure the team can crystallize certain principles or certain best practices."*

As a result, many participants desired better strategies to help team members better understand the iterative nature of AI fairness work, as well as to better advocate for long-term collaboration across roles. When creating the "ideal journey map" during the workshop (W5), SWE2 and SME2 wanted team members to continuously address AI fairness across the lifecycle of AI development, instead of *"discussing fairness almost like an empty gesture at the beginning of each season"* (SME2). Furthermore, participants repeatedly brought up the importance of changing the *"education and training team members received"* (PM3) to set up a common grounding among members of their organization about the iterative, social-technical nature of AI fairness practice, rather than leaving that to individuals to create and deliver.

5 LIMITATIONS

As discussed in Section 3, similar to prior FAccT work studying AI fairness practices in industry (e.g., [24, 36, 49, 59]), we recruited participants using a purposive sampling approach [18]. Although interview-based qualitative research does not necessarily make claims to generalizability, the participants we were able to recruit may have been limited by our positionality and the reach of our professional network. In addition, all of our participants self-identified

as having already worked on addressing AI fairness issues in their AI products and services, most of our participants worked at large technology companies, and the majority of our participants were located in the US or Europe. Future research should explore perspectives from people contributing to AI systems from outside technology companies, including non-profits, government agencies, and members of the public more generally; as well as perspectives from stakeholders outside of the U.S. and Europe.

6 DISCUSSION

Prior work has indicated the importance and challenges of cross functional collaboration for AI fairness. Through a series of interviews and workshops with industry practitioners, we have identified existing practices that practitioners have developed to support cross-functional collaboration. Below, we highlight key findings and discuss their implications to inform FAccT communities, AI practitioners, and organizations about how to better support cross-functional collaboration in AI fairness work.

6.1 Supporting Bridging work in Cross-Functional Collaboration

Our findings surfaced critical bridging activities that practitioners use to overcome barriers to cross-functional collaboration (Section 4.1). As practitioners work to bridge gaps in the evaluation, understanding, and contextualization around AI fairness for their teams, how might researchers better support these bridging activities?

In our study, we found that practitioners spent a large amount of effort on navigating mismatched goals for evaluations of AI fairness across different disciplines (Section 4.1.1). However, AI fairness methods and toolkits developed by researchers [e.g., 1, 9, 10, 17] often focus primarily on supporting quantitative analysis from more technical roles [8, 24, 76]. Hence, future research should explore better mechanisms to support mixed-methods approaches to evaluate AI fairness. For example, designers and researchers could explore the development of new tools and processes to aid cross-functional teams in combining fairness-related insights from quantitative analyses (e.g., statistical disparities across subgroups) and qualitative data from user research (e.g., insights into perceptions of algorithmic harms among users or other groups impacted by a system). Such resources could support the "co-evaluation meetings" brought up by practitioners in Section 4.1.1.

In addition, future research should explore ways to better support cross-functional teams in *translating* evaluation results into actionable steps that are both technically feasible and appropriately account for the socio-technical complexities involved in a given use case or context. In our study, we found that participants from both technical roles (DS) and user-facing roles (UX) engaged in substantial work to translate and contextualize abstract AI fairness concepts in the documentation they created for those in other roles (Section 4.1.4). We see opportunities to reduce the amount of translation work that falls on practitioners by developing more context- and usage-specific AI fairness guidelines and frameworks that highlight unique considerations for specific real-world applications, especially since resources and guidelines from industry and academia often serve as a benchmark for practitioners [4, 65, 71]. For instance, to better support AI teams working on building chatbot services for health care, FAccT researchers and practitioners could design AI fairness guidelines covering how to assess and mitigate potential biases (e.g., around sex, gender, and race) that might occur in health care applications [19, 54]. As regulations like *"Blueprint for an AI Bill of Rights"* [37] and *"AI Risk Management Framework"* [53] being recently released, future AI fairness guidelines should also update their relevant principles and strategies to reflect these regional governances and policies.

In addition, practitioners bridge their understanding and contextualize AI fairness concepts through discussion and deliberation across roles (Section 4.1.2). While prior work suggested implications around creating space and boundary objects for data scientists and UX practitioners discussing the integration between AI models and AI products [69, 78], our finding highlighted that the sociotechnical, contested nature of AI fairness [7, 33] added layers of complexity to the cross-functional communication and collaboration (Section 4.1.2 and 4.1.4).To this end, future research could draw inspiration from toolkits like Timelines [77] and Value Cards [63] to design processes for practitioners across roles to exchange perspectives and deliberate around AI fairness issues. In particular, these processes should scaffold practitioners negotiating about the problem formation of fairness issues [55], discussing the trade-offs among particular fairness metrics [30, 43], aligning disagreements around how to operationalize fairness in specific use cases [54, 58], and collectively contextualizing their AI products and services in end users' day to day experience interacting with the AI systems [25, 64].

6.2 Thriving, not just surviving: From piggybacking to re-shaping organizational culture

In our study, we observed that, faced with a lack of resources and organizational buy-in around fairness work, practitioners often "piggybacked" on existing initiatives with buy-in and support (Section 4.2). This tactic has previously been documented in other settings beyond AI fairness work, including environmental sustainability efforts and social justice work in industry [14, 61]. Researchers have argued that "piggybacking" effectively helped the ideas around environmental sustainability "survive" under conditions with limited organizational resources, but fundamentally "reshaping" the organizational culture to value the"long-term growth and resilience led by prioritizing sustainablity" was much needed to help the sustainability efforts to "thrive" [26]. Relatedly, our findings point to the value and relevance of such "piggybacking" tactics for enacting collaboration in AI fairness work *as a starting point*. At the same time, we believe there are still opportunities for practitioners working on AI fairness to go beyond "piggybacking" and enable broader "reshaping" [cf. 74] of the organizational culture around AI fairness. In particular, drawing from work by Nafus and Sherman [51], Wong described "tactics of soft resistance" that UX practitioners employ to make the value of their work relevant to other roles, and to *re-shape* their organization's culture, by rooting their resistance "within a broader logic of the role of the market or the usefulness of technology" [74].

To this end, future research should explore the missing opportunities around re-shaping the organizational culture, to help practitioners working on AI fairness "thriv[e], not just surviv[e]" [26] when navigating the collaboration around AI fairness under the organizational constraint. For example, resonating with DS3's call to explore the complementarity between privacy and fairness (Section 4.2.1), future research could explore more in-depth synergies and interplay between privacy and fairness, especially when responsible AI guidelines list AI fairness and privacy side-by-side as part of the key components to enable ethical and responsible AI systems [39]. For example, future FAccT research could conduct empirical study to understand how privacy work in industry practice come a long way, to offer implications for institutionalizing fairness work. Furthermore, future research and practices are much needed to empower practitioners to re-shape the quantification culture around AI development that runs the risk of compromising the socio-technical, contestable nature of AI fairness [31, 62]. In particular, tying closely to the point around mixed-methods evaluation approach discussed in the previous section (See section 6.1), instead of simply replying on "scores" and "percentages" to make fairness issue relevant to technical roles, future FAccT research and practices could design better structures and processes to help incorporate more nuanced notions of fairness into the current AI development pipeline. For example, future FAccT research and practice could explore extending the current model-building platforms (e.g., [20, 40]) with contextual messages that bring in nuanced understandings around AI fairness from direct stakeholders, to integrate mixed-method, socio-technical AI fairness analysis into technical roles' existing AI working pipelines.

6.3 Making Invisible Labor Visible and Valuable

Throughout our findings, we see that carrying out the efforts of bridging and piggybacking often requires practitioners to go beyond their traditional job descriptions, devote additional time and effort, and take on emotional burdens (Section 4.3) — what Star and Strauss described as "invisible work" [68]. This finding also draws a parallel with the recent journalism about the burnout problem in industry responsible AI work – due to the lack of appropriate recognition of their invisible work from colleagues and organizations, practitioners reported"feeling undervalued, which can affect their mental health and lead to burnout."[34] More recently, Wang et al. also surfaced the "hidden work" and "emotional labor" of UX practitioners when building more responsible AI, and how this work is often not recognized or valued by UX practitioners' managers or organizations, or leads others to view UX practitioners as a "blocker" of the design and development process [73]. We extend this prior work by highlighting how practitioners across roles in AI teams (including UX practitioners) brought in their relevant expertise and perspectives to go beyond their job descriptions to enact the collaboration in AI fairness.

To better support more sustained collaboration efforts around AI fairness, future research needs to explore processes and tools that help team members and organizations better recognize and value the efforts that go behind fairness work. For instance, abstraction has been highlighted as an important skill for collaborating and communicating in software engineering and data analysis in cross-functional teams [3, 46, 52]—although with the risk of losing the nuance of particular contexts [cf. 62]. However, our findings surfaced that, because of the abstraction that was intended to facilitate conversations across roles, other team members were not able to fully understand and appreciate the labor hidden behind the efforts individuals invested in enabling the collaboration in AI fairness 4.3. To this end, instead of hiding the currently invisible labor behind the abstraction, perhaps researchers could design interactive interfaces or visualizations offering options for practitioners across roles to see the detail of often laborious work that goes into the complex considerations and tradeoffs involved in fairness work, such as deciding on appropriate fairness metrics or larger analysis decisions when conducting disaggregated evaluations of fairness [6, 47], or the arduous processes that enable the clear presentation of an accessible visualization from data scientists (Section 4.3). An important step after recognizing the "invisible work" would be rewarding them. However, as repeatedly discussed in prior work, the efforts and the impact of "invisible work" are often hard to make clear or quantify [68, 73, 74] We suggest that FAccT researchers could draw inspiration from prior work attempting to quantify the invisible labour in crowdwork [cf. 70] as well as the tools to track and report "work tasks related to raising discussion and address[ing] values at work places" that are often invisible to other team members and organizations [cf. 75]. It is our hope that these new structures and tools could to some extent ameliorate practitioners' burnout induced by taking on under-recognized work.

More broadly, we urge companies to proactively recognize and reward the critical invisible labor that enables cross-functional collaboration in AI fairness. One concrete way to implement this is to include the time and efforts devoted to these invisible labor as part of companies' performance indicators (e.g., Key Performance Indicators, or KPIs) in order to incentivize collaborative fairness efforts—although adapting KPIs for fairness or responsible as a whole brings with it a host of tensions and contradictions around quantifying unobservable phenomena such as fairness [48]. Extending the implications from Wong and Wang et al. around better supporting UX designers' value work to assigning them roles like "Responsible AI expert," [73, 74] organizations could also explore formalizing more teams and roles for fairness work to empower practitioners across roles who currently go beyond their traditional job descriptions to facilitate cross-functional collaboration for building more fair AI systems.

7 CONCLUSION

In this research, we sought to better understand current strategies and challenges for cross-functional collaboration around fairness in AI, in order to identify opportunities to support more effective collaboration. Through a series of interviews and workshops with industry practitioners across a range of roles and companies, we found that practitioners engaged in critical, yet under-recognized *bridging* work to help teams overcome key barriers to cross-functional collaboration around AI fairness. In addition, given organizational constraints, practitioners often *piggybacked* on existing initiatives and corporate rhetorics to enable fairness efforts. Overall, we hope

that this work can (1) increase awareness among the FAccT community around the strategies and tactics that industry practitioners currently employ to facilitate collaborative AI fairness work; and (2) offer directions for future FAccT research and practice to better support cross-functional collaboration.

REFERENCES

[1] IBM Resaerch Trusted AI. 2021. AIF360 API. (2021). https://aif360.mybluemix.net/

[2] Jumana Almahmoud, Robert DeLine, and Steven M Drucker. 2021. How Teams Communicate about the Quality of ML Models: A Case Study at an International Technology Company. *Proceedings of the ACM on Human-Computer Interaction* 5, GROUP (2021), 1–24.

[3] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. 2019. Software engineering for machine learning: A case study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE, 291–300.

[4] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[5] Jacqui Ayling and Adriane Chapman. 2022. Putting AI ethics to work: are the tools fit for purpose? *AI and Ethics* 2, 3 (2022), 405–429.

[6] Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn M. Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, Duncan Wadsworth, and Hanna M. Wallach. 2021. Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (2021).

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

[8] Alex Bäuerle, Ángel Alexander Cabrera, Fred Hohman, Megan Maher, David Koski, Xavier Suau, Titus Barik, and Dominik Moritz. 2022. Symphony: Composing interactive interfaces for machine learning. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–14.

[9] Sarah Bird. 2020. Fairlearn API. https://fairlearn.github.io/v0.5.0/api_reference/fairlearn.datasets.html

[10] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. *Fairlearn: A toolkit for assessing and improving fairness in AI.* Technical Report MSR-TR-2020-32. Microsoft. https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

[11] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 173–184.

[12] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050* (2020).

[13] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qual. Res. Psychol.* 3, 2 (Jan. 2006), 77–101.

[14] Howard Brown and Timothy J. Larson. 1998. Making business integration work: A survival strategy for EHS managers. *Environmental Quality Management* 7 (1998), 1–8.

[15] Helen Burnie. 2016. Piggybacking on sustainability. *Practical Literacy: The Early and Primary Years* 21, 3 (2016), 35–38.

[16] Ángel Alexander Cabrera, Abraham J Druck, Jason I Hong, and Adam Perer. 2021. Discovering and validating ai errors with crowdsourced failure reports. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–22.

[17] Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. 2019. FairVis: Visual analytics for discovering intersectional bias in machine learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 46–56.

[18] Steve Campbell, Melanie Greenwood, Sarah Prior, Toniele Shearer, Kerrie Walkem, Sarah Young, Danielle Bywaters, and Kim Walker. 2020. Purposive sampling: complex or simple? Research case examples. *Journal of research in Nursing* 25, 8 (2020), 652–661.

[19] Davide Cirillo, Silvina Catuara-Solarz, Czuee Morey, Emre Guney, Laia Subirats, Simona Mellino, Annalisa Gigante, Alfonso Valencia, María José Rementeria, Antonella Santuccione Chadha, et al. 2020. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ digital medicine* 3, 1 (2020), 81.

[20] Colab. 2020. Welcome To Colaboratory. https://colab.research.google.com/

[21] Henriette Cramer, Jean Garcia-Gathright, Aaron Springer, and Sravana Reddy. 2018. Assessing and addressing algorithmic bias in practice. *Interactions* 25, 6 (2018), 58–63.

[22] Kate Crawford. 2017. The trouble with bias. keynote at neurips. (2017).

[23] Wesley Hanwen Deng, Bill Boyuan Guo, Alicia Devos, Hong Shen, Motahhare Eslami, and Kenneth Holstein. 2022. Understanding Practices, Challenges, and Opportunities for User-Driven Algorithm Auditing in Industry Practice. *arXiv preprint arXiv:2210.03709* (2022).

[24] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring how machine learning practitioners (try to) use fairness toolkits. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 473–484.

[25] Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. 2022. Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.

[26] Andrea di Miceli, Birgit Hagen, Maria Pia Riccardi, Francesco Sotti, and Davide Settembre-Blundo. 2021. Thriving, Not Just Surviving in Changing Times: How Sustainability, Agility and Digitalization Intertwine with Organizational Resilience. *Sustainability* (2021).

[27] Shelley Evenson. 2006. Directed storytelling: Interpreting experience for design. *Design Studies: Theory and research in graphic design* (2006), 231–240.

[28] EY. 2021. EY's Trust Score: Defining business priorities for long-term value. (2021). https://www.ey.com/en_us/consulting/trusted-ai-platform

[29] Diana Forsythe. 2001. *Studying those who study us: An anthropologist in the world of artificial intelligence.* Stanford University Press.

[30] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Commun. ACM* 64, 4 (2021), 136–143.

[31] Timnit Gebru. 2021. Hierarchy of Knowledge in Machine Learning Related Fields and Its Consequences. https://www.youtube.com/watch?v=OL3DowBM9uc

[32] Dave Gray, Sunni Brown, and James Macanufo. 2010. *Gamestorming: A playbook for innovators, rulebreakers, and changemakers.* " O'Reilly Media, Inc.".

[33] Ben Green. 2021. The contestation of tech ethics: A sociotechnical approach to technology ethics in practice. *Journal of Social Computing* 2, 3 (2021), 209–225.

[34] Melissa Heikkilä. 2022. Responsible AI has a burnout problem. https://www.technologyreview.com/2022/10/28/1062332/responsible-ai-has-a-burnout-problem/

[35] John W. Henke, A. Richard Krachenberg, and Thomas F. Lyons. 1993. Cross-Functional Teams: Good Concept, Poor Implementation! *Journal of Product Innovation Management* 10 (1993), 216–229.

[36] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–16.

[37] White House. 2022. Blueprint for an AI Bill of Rights. https://www.whitehouse.gov/ostp/ai-bill-of-rights/

[38] Benton C. Hutchinson, Negar Rostamzadeh, Christina Greer, Katherine Heller, and Vinodkumar Prabhakaran. 2022. Evaluation Gaps in Machine Learning Practice. *2022 ACM Conference on Fairness, Accountability, and Transparency* (2022).

[39] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.

[40] Jupyter. 2020. Jypyter: Free software, open standards, and web services for interactive computing across all programming languages. https://jupyter.org/

[41] Kenneth B Kahn. 1996. Interdepartmental integration: a definition with implications for product development performance. *Journal of product innovation management* 13, 2 (1996), 137–151.

[42] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna M. Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020).

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

[44] Jung-Joo Lee, Christine Ee Ling Yap, and Virpi Roto. 2022. How HCI Adopts Service Design: Unpacking current perceptions and scopes of service design in HCI and identifying future opportunities. In *CHI Conference on Human Factors in Computing Systems*. 1–14.

[45] Michelle Seng Ah Lee and Jat Singh. 2021. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–13.

[46] Barbara Liskov. 1987. Keynote address-data abstraction and hierarchy. In *Addendum to the proceedings on Object-oriented programming systems, languages and applications (Addendum)*. 17–34.

[47] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–26.

[48] Michael Madaio, Luke Stark, Wortman Vaughan Jennifer, and Hanna Wallach. 2020. Need for Organizational Performance Metrics to Support Fairness in AI. *Fair Responsible AI Workshop at the 2020 CHI Conference on Human Factors in Computing Systems* (2020), 1–14.

[49] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[50] Jacob Metcalf, Emanuel Moss, et al. 2019. Owning ethics: Corporate logics, silicon valley, and the institutionalization of ethics. *Social Research: An International*

Quarterly 86, 2 (2019), 449–476.
[51] Dawn Nafus and Jamie Sherman. 2014. This One Does Not Go Up to 11: The Quantified Self Movement as an Alternative Big Data Practice.
[52] Nadia Nahar, Shurui Zhou, Grace Lewis, and Christian Kästner. 2021. Collaboration Challenges in Building ML-Enabled Systems: Communication, Documentation, Engineering, and Process. *arXiv preprint arXiv:2110.10234* (2021).
[53] NIST. 2023. Artificial Intelligence Risk Management Framework. https://www.nist.gov/itl/ai-risk-management-framework
[54] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.
[55] Samir Passi and Solon Barocas. 2019. Problem formulation and fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 39–48.
[56] Samir Passi and Steven J Jackson. 2018. Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–28.
[57] Katrina Pavelin, Sangya Pundir, and Jennifer A Cham. 2014. Ten simple rules for running interactive workshops. *PLoS computational biology* 10, 2 (2014), e1003485.
[58] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 145–151.
[59] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23.
[60] Brianna Richardson, Jean Garcia-Gathright, Samuel F Way, Jennifer Thom, and Henriette Cramer. 2021. Towards Fairness in Practice: A Practitioner-Oriented Rubric for Evaluating Fair ML Toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.
[61] Thomas N Robinson. 2010. Save the world, prevent obesity: piggybacking on existing social and ideological movements. *Obesity* 18, n1s (2010), S17.
[62] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*. 59–68.
[63] Hong Shen, Wesley H Deng, Aditi Chattopadhyay, Zhiwei Steven Wu, Xu Wang, and Haiyi Zhu. 2021. Value cards: An educational toolkit for teaching social impacts of machine learning through deliberation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 850–861.
[64] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–29.
[65] Ben Shneiderman. 2020. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction* 36 (2020), 495 – 504.
[66] Jessie J Smith, Saleema Amershi, Solon Barocas, Hanna Wallach, and Jennifer Wortman Vaughan. 2022. REAL ML: Recognizing, Exploring, and Articulating Limitations of Machine Learning Research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 587–597.
[67] X Michael Song and Mark E Parry. 1997. A cross-national comparative study of new product development processes: Japan and the United States. *Journal of marketing* 61, 2 (1997), 1–18.
[68] Susan Leigh Star and Anselm Strauss. 1999. Layers of silence, arenas of voice: The ecology of visible and invisible work. *Computer supported cooperative work* 8, 1-2 (1999), 9–30.
[69] Hariharan Subramonyam, Jane Im, Colleen Seifert, and Eytan Adar. 2022. Solving Separation-of-Concerns Problems in Collaborative Design of Human-AI Systems through Leaky Abstractions. In *CHI Conference on Human Factors in Computing Systems*. 1–21.
[70] Carlos Toxtli, Siddharth Suri, and Saiph Savage. 2021. Quantifying the Invisible Labor in Crowd Work. *Proceedings of the ACM on Human-Computer Interaction* 5 (2021), 1 – 26.
[71] Kush R Varshney. 2019. Trustworthy machine learning and artificial intelligence. *XRDS: Crossroads, The ACM Magazine for Students* 25, 3 (2019), 26–29.
[72] Angelina Wang, Solon Barocas, Kristen Laird, and Hanna Wallach. 2022. Measuring representational harms in image captioning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 324–335.
[73] Qiaosi Wang, Michael Adam Madaio, Shivani Kapania, Shaun Kane, Michael Terry, Lauren Wilcox, et al. 2023. Designing Responsible AI: Adaptations of UX Practice to Meet Responsible AI Challenges. (2023).
[74] Richmond Y Wong. 2021. Tactics of Soft Resistance in User Experience Professionals' Values Work. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–28.
[75] Richmond Y Wong. 2021. Using Design Fiction Memos to Analyze UX Professionals' Values Work Practices: A Case Study Bridging Ethnographic and Design Futuring Methods. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.
[76] Richmond Y Wong, Michael A Madaio, and Nick Merrill. 2022. Seeing Like a Toolkit: How Toolkits Envision the Work of AI Ethics. *arXiv preprint arXiv:2202.08792* (2022).
[77] Richmond Y Wong and Tonya Nguyen. 2021. Timelines: A world-building activity for values advocacy. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.
[78] Nur Yildirim, Alex Kass, Teresa Tung, Connor Upton, Donnacha Costello, Robert Giusti, Sinem Lacin, Sara Lovic, James M O'Neill, Rudi O'Reilly Meehan, et al. 2022. How Experienced Designers of Enterprise Applications Engage AI as a Design Material. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–13.
[79] Amy X Zhang, Michael Muller, and Dakuo Wang. 2020. How do data science workers collaborate? roles, workflows, and tools. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–23.