Tinker, Tailor, Configure, Customize: The Articulation Work of Contextualizing an AI Fairness Checklist

MICHAEL A. MADAIO*, Microsoft Research, USA
JINGYA CHEN, Microsoft Research, USA
HANNA WALLACH, Microsoft Research, USA
JENNIFER WORTMAN VAUGHAN, Microsoft Research, USA

Many responsible AI resources, such as toolkits, playbooks, and checklists, have been developed to support AI practitioners in identifying, measuring, and mitigating potential fairness-related harms. These resources are often designed to be general purpose in order to be applicable to a variety of use cases, domains, and deployment contexts. However, this can lead to decontextualization, where such resources lack the level of relevance or specificity needed to use them. To understand how AI practitioners might contextualize one such resource, an AI fairness checklist, for their particular use cases, domains, and deployment contexts, we conducted a retrospective contextual inquiry with 13 AI practitioners from seven organizations. We identify how contextualizing this checklist introduces new forms of work for AI practitioners and other stakeholders, as well as opening up new sites for negotiation and contestation of values in AI. We also identify how the contextualization process may help AI practitioners develop a shared language around AI fairness, and we identify tensions related to ownership over this process that suggest larger issues of accountability in responsible AI work.

CCS Concepts: • **Computing methodologies → Artificial intelligence**; • **Human-centered computing → Empirical studies in HCI**; • **Social and professional topics → Socio-technical systems**.

Additional Key Words and Phrases: fairness, ethics, responsible AI, articulation work

ACM Reference Format:
Michael A. Madaio, Jingya Chen, Hanna Wallach, and Jennifer Wortman Vaughan. 2024. Tinker, Tailor, Configure, Customize: The Articulation Work of Contextualizing an AI Fairness Checklist. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 214 (April 2024), 20 pages. https://doi.org/10.1145/3653705

1 INTRODUCTION

As artificial intelligence (AI)[1] systems become ubiquitous across a variety of domains (e.g., finance, healthcare, education), there is growing recognition that they may cause fairness-related harms [18, 43, 57]. To help AI practitioners identify, measure, and mitigate these and other potential risks posed by AI systems during their design, development, and deployment, researchers in industry,

*Michael is now at Google Research, but this work was done while Michael was at Microsoft.

[1]Although AI is a contentious term [14, 66, 81, 84], we follow the OECD definition of AI: "*An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.*" [38].

Authors' addresses: Michael A. Madaio, Microsoft Research, New York City, NY, USA; Jingya Chen, Microsoft Research, Redmond, WA, USA; Hanna Wallach, Microsoft Research, New York City, NY, USA; Jennifer Wortman Vaughan, Microsoft Research, New York City, NY, USA.

ACM 2573-0142/2024/4-ART214
https://doi.org/10.1145/3653705

academia, and the public sector have developed toolkits and other resources for AI practitioners [e.g., 7, 9, 24, 37, 45, 49, 55, 92]—work that is sometimes referred to as responsible AI (RAI) [75].

However, as Wong et al. [92] discuss for RAI toolkits, general-purpose RAI resources may take a *"decontextualized approach to ethics"* that risks flattening critical nuance in what, e.g., RAI, ethics, societal inequity, and fairness may mean for the use cases, domains, and contexts in which AI systems are deployed [cf. 69]—and how those contexts may themselves shape AI systems' use and impacts [e.g., 26, 27, 50, 51, 67]. Thus, the general-purpose nature of many RAI resources may be at odds with AI practitioners' needs for specific RAI guidance for their particular use cases, domains, and deployment contexts [35, 48, 69]. Much like toolkits used in other domains [e.g., 52], general-purpose RAI resources may need to be customized. However, as Wong et al. [92] identify, this assumes that AI practitioners (or others involved in the design, development, and deployment of AI systems) will be able to adapt RAI resources for their purposes. Given that prior work has identified how *using* RAI resources requires navigating organizational cultures, work processes, and compliance requirements [cf. 24, 34, 49, 64], it is crucial to understand the work involved in contextualizing general-purpose RAI resources for AI teams' development workflows and their AI systems' particular use cases, domains, and deployment contexts. Thus, in this paper, we ask: **what new forms of work are introduced when AI teams contextualize a general-purpose AI fairness checklist for their development workflows and AI systems?**

To investigate this research question, we conducted a retrospective contextual inquiry with 13 AI practitioners from seven organizations who had experience identifying, measuring, and mitigating potential fairness-related harms in their AI design, development, and deployment processes. We explore how these AI practitioners contextualize a general-purpose, publicly available AI fairness checklist that we developed in our prior work [49]. We chose to focus on a checklist because, compared with other RAI resources [e.g., 7, 9], checklists are, broadly speaking, designed to reduce ambiguity by providing instructions for specific processes [22, 23, 32, 71], offering a fertile ground to explore tensions between general-purpose and context-specific RAI resources.

We draw on theories of *"articulation work"* to understand how the process of contextualizing a general-purpose AI fairness checklist introduces new forms of work for AI practitioners—work that is currently under-valued and under-recognized by their colleagues and employers. In other words, despite the focus of many available RAI resources on the *technical* aspects of RAI work [92], we find that the social and organizational work of contextualizing RAI resources is a crucial part of industry practices, but is not currently accounted for or valued as RAI work as such.

In addition, we find that this articulation work of contextualizing an AI fairness checklist *could* act as a sensitizing process to help AI teams align on RAI goals and priorities and develop a shared language around AI fairness, which is a key obstacle for current RAI work practices [25, 35, 48, 64]. Finally, we identify how the contextualization process opens up new sites for negotiation and contestation of values in AI, introducing tensions related to ownership and accountability in RAI work. By attending to AI practitioners' articulation work when contextualizing RAI resources, we can better understand how both users and designers of RAI resources might better navigate tensions between broadly applicable, general-purpose—yet decontextualized [92]—RAI resources, and more specific resources that are contextualized to AI teams' particular use cases, domains, and deployment contexts.

2 RELATED WORK

2.1 Customizing Responsible AI Resources

Researchers in the field of responsible AI (RAI) have developed a variety of resources—such as toolkits [e.g., 7, 9, 45, 65, 92], playbooks [e.g., 37, 56], and checklists [e.g., 28, 49]—to help AI practitioners identify, measure, and mitigate potential fairness-related harms and other potential

risks in their AI design, development, and deployment processes. They have also studied how AI practitioners use such resources as part of their RAI work within different organizational contexts [e.g., 24, 35, 48, 49, 60]. RAI resources are often designed to be general purpose in order to be applicable to a variety of use cases, domains, and deployment contexts. However, there is a growing awareness from researchers, policymakers, and others that it is not possible to create RAI resources that are truly one-size-fits-all, due to differences in AI teams' development workflows and their AI systems' particular use cases, domains, and deployment contexts [69, 73, 82, 92].

Thus, designers of RAI resources often claim that users can (and should) customize them. However, although such resources often document in great detail the steps involved in *using* them, when it comes to customizing them, they often simply call for customization, without specifying how such work should take place, or by whom. For instance, Obermeyer et al. [56] write that "*bias-prevention practices need to be customized for your organization*," while Hong et al. [37] write that "*the platform is extensible, and that teams can adapt and extend the Playbook to accommodate such scenarios*."

In her historical analysis of the design metaphor of the toolkit (writ large; not specifically for RAI), Mattern argues that toolkits are designed to be technologies of abstraction, allowing "*skilled practitioners*" to take a general-purpose toolkit (e.g., first aid kit, traveling salesmen kit) and adapt its contents and procedures to fit the needs of particular contexts [41, 52]. Wong et al. [92] draw on this framework to analyze RAI toolkits, arguing that RAI toolkits are often designed to be decontextualized to foster greater adoption [e.g., 7, 9, 45], but this means that users must adapt them for their AI systems' particular use cases, domains, and deployment contexts. Moreover, this presumes the existence of a particular type of "*skilled practitioner*" who knows how a given RAI resource should be modified to fit the situation at hand and who is capable of and incentivized to carry out those modifications as part of their work. In this paper, we therefore investigate how AI practitioners might contextualize general-purpose RAI resources, and what this work entails.

Checklists, as one type of RAI resource designed to introduce proactive processes for RAI throughout the AI lifecycle, have a different historical lineage than toolkits [e.g., 30, 93]. As such, they implicitly convey different expectations for how their users will engage with them. Widely used in other high-stakes, safety-critical domains such as surgery, aviation, and structural engineering [30, 32], prior CSCW research has argued that checklists are designed to "*provide instructions to actors of possible or required next steps*" in order to better standardize and "*reduce local control*" over processes [71]. This design paradigm may implicitly suggest to users that checklists will provide them with a comprehensive and actionable set of tasks (e.g., for surgical safety [32] or aviation emergencies [23]). However, this suggestion may be at odds with the reality that, for complex sociotechnical systems such as AI systems [82], no checklist may be able to identify *a priori* the particular steps needed to identify, measure, and mitigate potential risks when those systems are deployed in social contexts and used in ways not anticipated by the checklist designers [22, 49, 73, 82]. Given the sociotechnical nature of AI systems—that both shape and are shaped by the social contexts in which they operate—the potential risks and necessary mitigations may only be able to be understood when AI systems are studied *in situ* [e.g., 26, 27, 50, 51, 67].

Some AI fairness checklists have been developed for particular use cases, domains, and deployment contexts, such as the checklist for algorithmic bias in music recommendation developed and used by researchers at Spotify [21]. Others, such as the Deon checklist from DrivenData, provide support for users to customize the default checklist for their particular circumstances; for example, by providing customizable templates using the Deon GitHub package. However, this requires users who want to customize their checklists to have sufficient technical knowledge to use Github or their command line tool, which may not always be the case, given the diversity of skills and backgrounds of AI practitioners, not to mention the other stakeholders involved in RAI work [44, 64, 92]. Meanwhile, the AI fairness checklist that we developed in our prior work [49]

was ostensibly designed to be general purpose, and in the preamble to the checklist, we similarly called for users to customize it for their needs, like many other RAI resources [e.g., 37, 56].

Without adapting RAI resources for particular use cases, domains, or deployment contexts, erroneous assumptions may be propagated. As one example, Pushkarna et al. [62] describe how a team had used a Data Card template to document a dataset that was used to train an AI model, and then modified that template for another dataset the team believed to be similar. However, the previous version of the Data Card template was missing crucial sections that turned out to be highly relevant for the subsequent dataset, but were nevertheless left out of that version [62].

Thus, practitioners who use general-purpose RAI resources may need to contextualize them for their development workflows and their AI systems' particular use cases, domains, and deployment contexts [e.g., 21, 24, 35, 49, 64, 92]. However, without attending to the work required to contextualize RAI resources—or indeed, without valuing the contextualization process itself as RAI work—we may fall victim to what Selbst et al. [74] refer to as the "*portability trap*" for sociotechnical systems: believing that a sociotechnical artifact (be it an AI system or a checklist) designed for a particular context will be able to be used in a new context without significant modifications. To help us understand the work involved in contextualizing general-purpose RAI resources, such as AI fairness checklists, for AI teams' development workflows and their AI systems' particular use cases, domains, and deployment contexts, we draw on theories of articulation work from CSCW.

2.2 Articulation work in information systems: "the work that makes work happen"

Articulation work is a theoretical lens developed and used in the CSCW community to describe how work happens—"*the specifics of putting together tasks [and] task sequences... and keeping them together*"—in response to changing circumstances [77]. This may include, for instance, planning and coordinating work practices, negotiating and persuading collaborators or organizational leadership of various courses of action, acquiring and deploying resources [77]—work that some researchers identify as being critical to RAI [25, 33, 49, 60, 86]. As Suchman argues, the work of design (whether of a technological artifact or a sociotechnical process) does not end prior to adoption by users, but continues in the "*in situ work of design in configuration*" [80], as designers may be unable to fully specify the set of requirements needed for every possible use case, domain, or context in which their artifacts or processes may be used—particularly as social contexts shape (and are shaped by) the use and impacts of algorithmic systems [cf. 26, 27, 50, 51, 67, 68]. That is, in order for sociotechnical systems to be useful (or used at all), Suchman argues that practitioners must "*take up the work of design*" to configure those artifacts or processes to better incorporate them into their social practices and material environments via, e.g., "*integration, configuration, customization, maintenance, and redesign*" [79].

The CSCW community has long been concerned with how processes, workflows, and other organizational constructs, may shape—and be shaped by—practitioners' work practices [71]. This tension between standardization and flexibility in the interactions between formal procedures and practitioners' work practices is a recurring theme of much CSCW research, particularly research that draws on the lens of articulation work to understand how information infrastructures [4, 54], information standards [39], and agile development processes [83] are adapted to fit local needs. Jackson discusses "*the creative role of standards... [as] professionals fill in the gaps left behind by standards*" [39], describing the coordination, workarounds, and discretionary judgment that practitioners engage in to adapt standards for the contingencies of their circumstances.

In data science, prior work discusses the role of articulation work as the "*meta-project work*" that enables the work of data science to take place by determining what must by done, by whom, when, and how [59]. As Passi [59] argues, formal procedures for data science work break down—and must be reconstituted by data science practitioners—in the face of local demands, particularly as those practitioners negotiate priorities for fairness and other values in their work [60].

Participants ID	Role	Focus Area	Used a checklist?
P1	ML Engineer	Social networking	N
P2	Cloud Solutions Architect	Customer solutions	Y
P3	Solutions Sales Specialist	Customer solutions	Y
P4	Policy / Risk Management	Fairness/policy advising	N
P5	Research Scientist Manager	Public sector predictive analytics	Y
P6	Product Manager	Nonprofit partner research (Health)	Y
P7	Data scientist / ML Engineer	Health	Y
P8	Data scientist / ML Engineer	Social networking	N
P9	Data Scientist Manager	Data science consulting (Health)	Y
P10	Data and Applied Scientist	Text prediction	N
P11	Partner Development Manager	Machine translation	N
P12	Designer	Security, Compliance, and Management	N
P13	Program Manager	Responsible AI	N

Table 1. Details of participants' roles, focus areas, and whether they had used a checklist previously.

However, this crucial articulation work is often not accounted for or valued as work in the same way as the technical aspects of data science or software development, whether due to being left out of organizational procedures, as Papoutsi and Brown [58] write about for the articulation work of privacy; a lack of support for articulation work provided by technical systems [58] (including RAI toolkits [92]); or simply being undervalued by technical practitioners and their organizational leadership as part of a larger legacy of "*deleting the social*" in technical work generally [76] and in AI specifically [29]. Given the prevalence of technical values (e.g., accuracy, speed) in AI, and the "*hierarchy of knowledge*" [31] that privileges technical work over social (or sociotechnical) work in RAI [11, 20, 63], it is critical to understand the role of articulation work in RAI work practices. Thus, in this paper, we investigate the work involved in contextualizing RAI resources for AI teams' development workflows and their AI systems' particular use cases, domains, and deployment contexts.

3 METHODS

3.1 Participants

Our study included 13 AI practitioners from seven organizations. We recruited participants via a combination of purposive and snowball sampling. We used direct emails and posts on message boards related to AI fairness, and we asked each participant to send our recruitment email to other contacts. We sought to recruit participants in a variety of roles, working on a variety of AI systems (e.g., social networks, health algorithms, text prediction systems), from a variety of organizations, including both large and small technology companies, as well as public sector agencies. Participants included program managers, applied scientists, and data scientists, as well as people in RAI advisory roles and people working as consultants for third-party partners, in roles such as cloud solutions architect. Several participants reported that they had already used checklists in their development workflows; three had already used the AI fairness checklist that we developed in our prior work [49] [P5, P6, P7], one had used the Deon checklist from DrivenData [28] [P9], and two had used checklists about sales and solutions architecture unrelated to RAI [P2, P3]. See Table 1 for a summary of the 13 participants.

3.2 Study design

We chose to focus on a checklist because, compared with other RAI resources, checklists typically provide a well-specified set of procedures for their users [71], and there is prior work on checklist customization in other domains [e.g., 22, 23, 32]. Specifically, we used an AI fairness checklist that

Section	Example items
Envision	1.1.a Envision system and its role in society, considering: • System purpose, including key objectives and intended uses or applications • Sensitive, premature, dual, or adversarial uses or applications • Expected deployment contexts (e.g., geographic regions, time periods) • Expected stakeholders (e.g., people who will make decisions about system adoption, people who will use the system, people who will be directly or indirectly affected by the system, society), including relevant demographic groups (e.g., by race, gender, age, disability status, skin tone, and their intersections) • Expected benefits for each stakeholder group, including relevant demographic groups • Relevant regulations, standards, guidelines, policies, etc.
Define	2.2.a Define datasets needed to develop and test the system, considering: • Desired quantities and characteristics • Potential sources of data • Collection, aggregation, or curation processes • Relevant regulations, standards, guidelines, policies, etc. • Assumptions made when operationalizing system vision via datasets 2.2.b Scrutinize resulting definitions for potential fairness-related harms to stakeholder groups, considering: • Types of harm (e.g., allocation, quality of service, stereotyping, denigration, over- or underrepresentation) • Tradeoffs between expected benefits and potential harms for different groups
Prototype	3.4.a Undertake user testing with diverse stakeholders, analyzing results broken down by relevant stakeholder groups. This should be done even if the system satisfies the fairness criteria because the system may exhibit unanticipated fairness-related harms not covered by the fairness criteria. Consider conducting: • Online experiments • Ring testing or dogfooding • Field trials or pilots in deployment contexts 3.4.b Revise production system to mitigate any potential harms; if this is not possible, document why, along with future mitigation or contingency plans, etc., and consider aborting development
Build	4.5.a Solicit input on production system from diverse perspectives, including: • Members of stakeholder groups, including relevant demographic groups • Domain or subject-matter experts • Team members and other employees
Launch	5.2.a Establish processes for responding to or escalating stakeholder feedback, including: • Stakeholder comments or concerns • Third-party audits
Evolve	6.2.a Monitor fairness criteria for deviation from expectations, including: • Adversarial threats or attacks 6.2.b If system fails to satisfy fairness criteria, revise system accordingly; if this is not possible, document why, along with expected impacts on stakeholders, and consider rollback or shutdown.

Table 2. Example items from the AI fairness checklist [49].

we developed in prior work [49] because this checklist was designed to be general purpose, is publicly available, and was co-designed with AI practitioners to be aligned with the typical AI lifecycle. The checklist includes items that prompt AI practitioners to identify, measure, and mitigate potential fairness-related harms across six phases of the typical AI lifecycle, from envisioning

an AI system, to defining the datasets and system architecture, to prototyping and building the system, and continuing through the system's launch and ongoing monitoring and post-deployment revisions. See Table 2 for example checklist items for each one of the six phases.

To investigate our research question, we conducted a retrospective contextual inquiry [42], a method from HCI that uses semi-structured interviews and artifacts to probe on work processes that unfold over longer time periods or that may be difficult or impossible to observe using a more traditional contextual inquiry [36] for other reasons. During the interviews, we asked participants to reflect on their typical AI fairness work practices (by describing a recent relevant project), presented participants with the AI fairness checklist, and asked participants to customize it as needed for their particular use cases, domains, and deployment contexts, prompting them to think aloud.

We also asked participants to draw storyboards of how they *had* customized (for participants who had already used an AI fairness checklist) or how they *would* customize (for other participants) the checklist to contextualize it for their teams' development workflows and AI systems' particular use cases, domains, and deployment contexts. Throughout this portion of the study, we prompted participants to think aloud by reflecting on the processes they sketched out in the storyboards and how these processes might occur within their teams and organizations. We also prompted participants to reflect on the storyboards that other participants had created.[2] Each interviews was around 60 minutes long. Participants were compensated, and the study was approved by our institution's IRB.

3.3 Data analysis

To understand the most salient themes in the transcripts from the interviews and the storyboard think-alouds, we adopted an inductive thematic analysis approach, following Braun and Clarke [16, 17]. One of the authors coded the transcripts using Atlas.ti and clustered the codes into themes using virtual whiteboard software, discussing these themes with the other authors, combining codes, and iterating on the themes until consensus was reached. For example, the codes include "modifying the checklist structure" and "assigning tasks to team members," while the themes include "ambiguity in ownership over customization" and "configuration as a sensitizing process." In the remaining sections, we report the themes we identified, and we highlight some implications of our findings.

4 FINDINGS

4.1 Customizing the checklist

First, participants described how contextualizing the checklist introduces new forms of work, in which they would need to customize the checklist items and structure for their AI systems' use cases, domains, and deployment contexts, involving team members and external domain experts in the process of deciding which parts of the checklist to keep, change, or remove.

One participant, who worked on a machine learning model in healthcare and had already used an AI fairness checklist, told us that a general-purpose checklist is *"what's off the shelf and then that sort of gets updated to one that has domain-specific concerns added... like, not sharing HIPAA*[3] *protected information"* [P9]. Other participants similarly reflected on how they needed to incorporate policy requirements for their AI systems' domains or add new checklist items to attend to domain-specific aspects of AI fairness. To do this, some participants described holding a meeting at the start of the project to review the checklist and discuss the changes needed for their AI systems' domains: *"we had the bones of the checklist already. We actually usually start with a meeting where we're talking*

[2]The reflections prompted by the storyboards are more salient to our research than the storyboards themselves, so we do not include example storyboards.

[3]HIPAA, or the Health Insurance Portability and Accountability Act, is a U.S. federal law that created national standards to protect patients' healthcare information.

about that. So it's either on the agenda for a half day kickoff meeting, where one of the sessions we do is on the ethics checklist, or it's the case that we schedule a separate meeting to go over that" [P9]. However, decisions around which parts of the checklist to keep, change, or remove were not always straightforward and were often contentious. We further discuss this in Section 4.4.

Part of the work of customizing the checklist involved teams' decisions about whether and how to change the checklist structure to align with their development workflows. Participants voiced concerns about the extent to which they would be expected to follow the existing checklist structure or adapt it to fit their needs: *"is there a way to bypass certain stages if they're not relevant to you?"* [P6]. Other participants said that they wanted to ungroup the checklist items from their grouping into six phases of the typical AI lifecycle, instead wanting to group them into related tasks or themes: *"I think the first thing I would do would be to kind of adapt it into pieces. Like, treat it as a sort of modular system. So it's like, ideation playbook, fairness testing, affected groups, thematically clustered... but some of these things might come up in different chunks at different times"* [P4].

Not all participants felt that they knew how to—or would have the autonomy to—make major changes to the checklist items and structure. The primary users of many checklists [e.g., 28, 49] and toolkits [e.g., 9, 24] are often assumed to be data scientists, but participants primarily saw the contextualization process as the responsibility of PMs (i.e., project managers, program managers, or product managers, each of which have slightly different roles and responsibilities), whom they described as having ownership over other compliance processes. As one participant told us: *"on our team, the PM is responsible for completing the compliance and legal aspect... initially the PM would go in, scope out the project, complete all the compliance and legal requirements"* [P6]. These decisions also involved input from other team members because *"the PM would have to work with our data scientists to figure out what parts of the checklist apply"* [P6]. Other participants (albeit significantly fewer) described involving external domain experts in the contextualization process, *"to make sure that the whole lifecycle is covered from each perspective... from fairness research or from machine learning research, and from clinicians, from nurses, from the administrative standpoint, the person who has to do the billing in the end, for example. So, the more the merrier"* [P7].

In other words, customizing the checklist was seen as a new part of project scoping or adherence with compliance procedures, albeit one for which PMs would need to draw on the expertise of other team members and external domain experts. However, many AI teams lack the meaningful collaboration structures or engagement opportunities that would enable their PMs to do this [cf. 60]. We further discuss tensions in ownership over the contextualization process in Section 4.4.

4.2 Integrating the checklist into AI teams' development workflows

4.2.1 Integrating the checklist into existing processes and tools. Participants described how they would need to integrate the checklist into their teams' existing processes and tools. Although our prior work argues that AI teams would need to integrate the checklist into their development workflows [49], we did not specify *how* such integration might be done. Here, we discuss the ways that AI teams might integrate the checklist into existing processes for project scoping or adherence with compliance procedures, or their AI development processes more generally. We also discuss challenges introduced by integrating the checklist into existing tools for project management.

As one example, integrating the checklist into AI teams' privacy review processes was seen as a way to avoid "*review fatigue*." For instance, one participant told us:

> *"My instinct has always been to shoehorn related topics into the existing review processes while making it seem minimally different than what's currently required because of review fatigue and the fact that there is so much overlap as well... there's sort of already a jurisdiction to think about that, and so how do we encourage the creation of the muscle."* [P4]

However, this can lead to concerns about fairness being sublimated into privacy [cf. 25] or potential jurisdictional conflicts about who should own different focus areas in review processes [cf. 1].

Other participants described how integrating the checklist into their AI development processes might leverage the familiarity of those processes to promote adoption of the checklist. As one participant put it, *"it feels like a resource for the actors that are trying to inject this thinking into workflows where they're not naturally injected rather than saying, "Here's an off-the-shelf tool that you can apply"* [P4]. For this participant, integrating the checklist into their AI development processes would "*inject this thinking*" or foster the "*creation of the muscle*" [P4] about AI fairness in ways that an "*off-the-shelf tool*" might not. They suggested that making time for contextualizing the checklist might create opportunities for discussing AI fairness—opportunities for which their team might not otherwise make time. We further discuss this potential benefit in Section 4.3.

The process of integrating the checklist into AI teams' AI development processes served both a *rhetorical* role—to convince team members and other stakeholders to adopt the checklist—as well as an *educational* role to make it easier for AI practitioners to learn about AI fairness. As one participant, a solutions architect who consulted for multiple third-party partners, told us:

> *"[clients] love the kind of templates and blueprints we can give them that can be attached to the actual blueprint they already do. So, for example, they already work with CRISP-DM [Cross-industry standard process for data mining], KDD [knowledge discovery and data mining], or something like that, or their own methodology... so, they want to have this type of asset like a checklist about how to start these type of projects."* [P2]

For many participants, integrating the checklist into their teams' development workflows involved integrating it into their AI development *tools* as well, which introduced new challenges. We repeatedly heard about the value of integrating the checklist into DevOps tools. DevOps (development operations) describes both a philosophy, as well as a set of processes and tools, for integrating the development and deployment of new software [46]. As one participant told us:

> *"[We use] CICD [continuous integration and continuous delivery], boards, planning, this kind of stuff. There are experimentation platforms as well. And we have machine learning services. If we integrate [the fairness checklist] and we have examples with the platform that customers and partners are using during experimentation this will be easier because these involves a lot of experimentation."* [P2]

However, this conflation of AI development processes and tools is complicated by the fact that many organizations use proprietary tools that may not allow teams to easily integrate new artifacts like an AI fairness checklist. Thus, many participants reported needing to create *ad hoc* processes and tools for their teams to better integrate with their existing tooling infrastructure.

4.2.2 Translating checklist items into actionable tasks. In addition to incorporating the checklist into their teams' existing processes and tools, participants talked about translating checklist items into actionable tasks, assigning those tasks to relevant team members, and defining the tasks' timelines and priorities. However, for some participants, translating checklist items into actionable tasks required significant effort, especially for some checklist items. As one participant told us:

> *"I love the question about 'considering who the system will give power or take power from', but that's one where the follow-up questions from whoever [on their team] is being asked that question will be, "Well, what do I do about that? What's something actionable?"* [P4]

This kind of translation work is a crucial step in transforming a checklist from a set of guiding prompts into a set of actionable tasks, but participants often saw it as being difficult for *any* team member to do, either due to gaps in experience or knowledge [cf. 92] or due to the difficulty of making time (or finding the resources needed [cf. 48]). As one participant reported, *"The cognitive*

load of translating the high-level thing, I think would be unexpected. And not because the content is not important. [...] The question is how do you inject the different pieces into each part of the process?" [P4].

Moreover, AI teams must make decisions about the priorities of different checklist items, given their other priorities. Indeed, prior work suggests that these prioritization decisions may be an additional source of inequity, as prioritizing identifying, measuring, and mitigating potential fairness-related harms for, e.g., the largest numbers of users or the highest-paying markets may inadvertently compound societal inequities caused by AI systems [48]. As one participant reported, their team *"would want additional guidance about [...] what is something that I must do right now versus what I must do in [a longer] period of time."* [P4]. However, many participants felt unsure about how to make these prioritization decisions, due to a lack of organizational clarity about who would have the decision-making power to determine the priorities of different checklist items.

4.3 Contextualization as a *sensitizing process* of reaching alignment

Participants also saw contextualizing the checklist as a key step toward helping their teams reach alignment on the scope of fairness work for their AI systems, as well as helping their team members develop their awareness of RAI issues or "*ethical sensitivity*" [cf. 15]. In other words, contextualizing the checklist was seen as a *sensitizing process* that could provide an opportunity for AI teams to align on the necessary steps for undertaking AI fairness work, as well as enabling team members to learn more about identifying, measuring, and mitigating potential fairness-related harms.

This alignment was often framed in terms of getting buy-in on the scope of work involved in identifying, measuring, and mitigating potential fairness-related harms from team members, organizational leadership, and other stakeholders. In contrast with many RAI resources, like toolkits, that are envisioned as having primarily technical users [92], most of our participants described expansive sets of stakeholders who are—or should be, but are not—involved in the work of using the checklist to identify, measure, and mitigate potential fairness-related harms. These stakeholders include PMs who assign tasks to team members, data scientists who conduct fairness assessments, organizational leaders who might incentivize or otherwise enable their teams to engage in AI fairness work, as well as customers, clients, external auditors, and other stakeholders.

However, as nearly all of our participants told us, the stakeholders involved in AI fairness work may not have equivalent levels of experience with or knowledge of either AI fairness or AI more generally. As one participant said: *"I know with responsible AI, not everyone has equal knowledge in the space"* [P6]. Or, as another participant put it: *"It's a big topic. If people haven't thought about that before, it can be difficult to know where to start"* [P4]. In particular, several participants pointed out how AI fairness work *"can be a steep learning curve for a lot of engineers"* [P1], and wondered aloud whether *"our data scientists have the training to think about [fairness] productively"* [P6]. As prior work identifies [e.g., 48, 63, 92], the skills required for AI fairness work (and RAI work more generally) may involve more *socio*technical skills [cf. 26, 50, 51] than many AI practitioners have encountered in their training. For instance, AI fairness work may involve understanding what marginalization or fairness means in the particular sociocultural contexts in which AI systems are deployed [e.g., 8, 69]; understanding the public's expectations' for fairness in algorithmic pricing models [e.g., 27] or employee management algorithms [e.g., 67], as well as understanding how to collect demographic data for use in measuring system performance disparities [e.g., 3, 6, 48].

Although some AI practitioners may lack sociotechnical skills, others may have these skills, but face difficulties in using them due to disciplinary boundaries [cf. 25]. For many participants, the varying types of expertise among the stakeholders involved in AI fairness work posed challenges to having productive conversations about identifying, measuring, and mitigating potential fairness-related harms. As one participant described, there may be expertise gaps from *"non-[technical] reviewers who have to encounter the technical teams and understand whether and to what extent*

they've followed any sort of [fairness] guidance, or identify issues that [the team] would need to follow up on" [P4]. In such cases, some participants felt that using the checklist may help *"bring people in in the middle [of the process of fairness assessments] and make them confident that some level of work has already happened and some level of due diligence and thoughtfulness has happened"* [P4].

Despite these challenges, participants felt that engaging multiple stakeholders—even stakeholders with varying types of expertise—in the contextualization process might help bridge some gaps:

> *"It lets us sort of shorthand some things later in the process when we get to the other questions like, did you include race or gender in the model? And then we say, 'Hey, remember, we talked about that and it depends on the goal of our model whether or not we want to include those things. And, you know, then in this case, where we're just measuring the effect, based on those classes then we need to include it in the model, otherwise we don't know what the disparate effect is and so that's why it's in there in this case.' And so having that kind of shorthand where you could point back to it and talk to people, I think it's one of the key valuable things for us in our process."* [P9]

Given differences in their team members' knowledge of AI fairness, many participants described the value of the checklist—and the process of involving their team in customizing it—as *"a way to sensitize people that [fairness] is important to think about"* [P7]. That is, going through the contextualization process might help their team members develop their awareness of potential fairness-related harms (or, what Boyd and Shilton [15] refer to as "*ethical sensitivity*"):

> *"We can't discuss all the possibilities [for potential fairness-related harms] and what we need to create is more like an awareness. I believe for all these checklists, at least when I see like checklists being really followed, the main result of that is not really the creating the template and using that, but the awareness that it's creating in the professionals who are responsible for creating the scope. After doing this template a few times, they have the awareness."* [P2]

To sum up, identifying, measuring, and mitigating potential fairness-related harms is deeply interdisciplinary, sociotechnical work [cf. 26, 51], yet many RAI resources are envisioned as having primarily technical users [92]. Moreover, the aspects of this work that prove most challenging to AI teams are those that require sociotechnical skills, such as awareness of societal inequities, marginalization, and harms [48]. Given these experience and knowledge gaps, AI teams that work together to contextualize a general-purpose AI fairness checklist for their particular use cases, domains, and deployment contexts may benefit from using the contextualization process as a way to develop or strengthen those skills—i.e., to sensitize team members about RAI work.

4.4 Tensions in ownership over the contextualization process

Finally, we find that contextualizing the checklist opens up new sites for negotiation and contestation of values, such as fairness, in AI. Across nearly all of our participants, we encountered tensions related to ownership and ambiguity over who should be responsible for doing the work of contextualizing the checklist—despite believing that that work is important—as well as for making decisions about whether and how to proceed once a checklist item is completed (or indeed, defining what it means for a checklist item to be completed [cf. 82]). As one participant told us:

> *"the role of creating the [checklist] template is a gray zone. [...] at the end of the day, we don't have a clear owner of this. And this is one thing that hurts, because if nobody is the owner, sometimes we don't see the evolution of that being created."* [P2]

As we described in Section 4.1, customizing the checklist was seen as the responsibility of the PMs, but for many participants, once the checklist was customized and integrated into their development

workflows, they were unsure about who should be responsible for deciding whether checklist items had been sufficiently completed. Several participants wondered aloud whether there should be some "approver" to review and approve the corresponding tasks, but most felt they did not have team members (or even others in their organizations) with sufficient RAI expertise to do this:

> *"What would be helpful is to have every project fill out this [checklist] to make sure our data scientists and our PMs are thinking about it. And I think what will be helpful is, when we finish, to have an RAI advisor specifically on the team that has the training that can then look at that documentation and be like, 'Okay, you thought about this, but you didn't think about this. How are you going to address this in the model? I don't know if we should go forward.' Someone with a higher [authority] that could kind of approve that the task has been completed properly or productively."* [P6]

This desire to offload responsibility for decisions related to the checklist also extended to participants wanting an external authority to make "go or no-go" decisions about whether and how to proceed with the design, development, and deployment of their AI systems following the completion of various checklist items. For instance, after conducting a fairness assessment and measuring system performance disparities, participants were unclear *"who has the authority to make the ultimate decision on code, like 'Go' versus 'No-go.' Is it going to be a senior level person? Is it someone who needs to have certain expertise in RAI or fairness?"* [P6]. In addition, when situating decision-making responsibility for the completion of checklist items with RAI advisors or other consultants, participants raised concerns about the checklist being seen as simply another compliance task, rather than a "grassroots" process that AI practitioners might buy into and use to inform their individual work and their conversations with others, including organizational leadership:

> *"A lot of those data scientists aren't going to be the ultimate decision-makers on these kinds of things, but we give them some framing, and some tools to help force that conversation to the people that are decision-makers. So thinking of it like a little more as a grassroots thing, and less like a compliance thing, where you've got a compliance officer that is like "Okay, every project has to do the ethics checklist," and people just do it as busywork."* [P9]

Here, we can see how integrating the checklist into AI teams' development workflows introduces ambiguity about who should be responsible for owning decisions related to AI fairness. As this participant points out, data scientists may not have the authority to decide whether and how to proceed with the design, development, and deployment of their AI systems. Indeed, prior work suggests that individual AI practitioners may be disincentivized to raise concerns about potential fairness-related harms by the culture and organizational incentives of their teams [2, 49, 64, 85].

Participants also identified tensions across the sets of stakeholders involved in the work of using the checklist to identify, measure, and mitigate fairness-related harms, some as a result of differences in their expertise, their values, or their relative power. One participant, a cloud solutions architect who worked with multiple third-party partners to help them develop AI systems, discussed tensions in priorities for RAI and fairness across multiple levels of their partner organizations:

> *"Customers, they have a preoccupation with responsible AI and fairness at the executive level. But at the mid-manager level, they are more focused on the having the best performance and having the best metrics. And I see that there is a misalignment there. And the data scientists at the bottom, if they are more mature data scientists they care about responsible AI. But the young data scientists, mainly nowadays, they have this Kaggle culture. They only want to have the best performance and be the top one."* [P2]

As this participant describes, the values of the actors at different organizational levels may not be well-aligned, from data scientists focusing on achieving state-of-the-art system performance (what

they refer to as "Kaggle"[4] culture, which emphasizes performance on narrowly scoped metrics for predefined AI tasks on publicly visible leaderboards), to managers trying to optimize for the metrics that matter most for their own promotions, to executives who may care about AI fairness (perhaps due to reputation effects [cf. 48]) but who are furthest removed from operationalizing it.

This misalignment in values, and hence priorities, has implications for whether and how the checklist might be used. One participant pointed out the consequences of this misalignment, saying:

> *"This checklist is designed for people who are engaging in good faith, and discussion of data science ethics. If you can get people to engage in good faith, then it's a useful tool. But I don't have that next level of 'How do you get people, how do you get organizations, how do you get companies to engage in this discussion in good faith?' That is a question of building internal political will within organizations that I don't think is easy to do."* [P9]

To summarize, the contextualization process opens up new sites for negotiation and contestation of values in AI, introducing tensions related to ownership and ambiguity over who should be responsible for contextualizing the checklist, as well as decisions about whether and how to proceed with design, development, and deployment once a checklist item is completed.

5 DISCUSSION

In the previous section, we described the new forms of work that are introduced when AI teams contextualize a general-purpose fairness checklist for their development workflows and AI systems. We found that although AI teams may lack the skills, resources, or capacity to do this work, it could act as a sensitizing process to help AI teams align on RAI goals and priorities and develop a shared language around AI fairness. Despite these benefits, we also identified tensions in ownership over the customization process that suggest larger issues of accountability in RAI work. In this section, we discuss some implications of these findings for the design of RAI resources and for RAI work itself.

5.1 Designing for positive ambiguity in RAI resources

Prior work identifies how toolkits—in order to travel widely and be usable (and scalable) across contexts—may inadvertently act as "*devices for decontextualization*" [41, 52], which, for RAI toolkits (as one type of RAI resource), may flatten crucial differences in what AI fairness and RAI mean, and how to operationalize these concepts within particular political and sociocultural contexts [69, 92]. Here, we shed light on how AI practitioners might *contextualize* one such general-purpose RAI resource, an AI fairness checklist, for their particular use cases, domains, and deployment contexts.

Our findings suggest that designers of RAI resources should consider how to support "*positive ambiguity*" [cf. 5, 64], in turn supporting AI practitioners in contextualizing RAI resources. Bamberger and Mulligan [5] identify how broad legislation can shift accountability to companies to operationalize that legislation in productive ways, potentially creating a larger role for employees within those companies who must reflect on how to interpret the legislation [5, 64].

For RAI, this might mean creating resources that are inherently under-specified to allow for their use in a variety of use cases, domains, and deployment contexts [cf. 72]. However, checklists are often intended to *reduce* local control and standardize procedures [71]. Navigating this tension between the minimum information or functionality that AI practitioners need from RAI resources and the positive ambiguity needed to support AI practitioners in contextualizing those resources is a critical direction for future work to explore. For instance, how much adaptation from an RAI resource is possible before it no longer reflects the processes (or even values) its designers intended?

[4] https://www.kaggle.com/

In addition, we identify a tension that arises when conflating AI development processes and tools to help AI practitioners instantiate those processes. This tension manifested in our study via participants who wanted to incorporate the AI fairness checklist into their teams' existing tools for project management and DevOps. Participants recognized that AI fairness work is an iterative process that requires reflection and sustained engagement with stakeholders. However, this work may not fit neatly within the design of their AI development tools, which are often designed for tasks that are well scoped, well specified, and able to be completed a single time before the project can move on, rather than tasks that need to be revisited and iterated on throughout a system's design, development, and deployment (like many RAI tasks). It is thus worth asking how RAI resources might be designed to unsettle existing development workflows toward more ongoing responsible design, development, and deployment, rather than to fit neatly within existing development workflows.

The structure of the AI fairness checklist that we developed in our prior work [49] was divided into six phases of the typical AI lifecycle, from envisioning to launching and evolving models after deployment—a structure that we co-designed with AI practitioners [49]. However, AI teams may begin their work in different phases of this AI lifecycle, so they may not be able to start from the beginning of the checklist and work their way down—an issue that may be exacerbated by modern AI development paradigms, such as AI APIs, AI services, or large-scale pretrained models (e.g., large language models) that are fine-tuned for downstream tasks [13, 40, 47]. As Wang et al. [86] identify, many modern AI development paradigms involve taking large models that are pretrained and building AI-powered applications with those pretrained models, posing new potential risks. In such paradigms, it is not clear to what extent AI teams will be able to follow the structure of a general-purpose AI fairness checklist—which may require individual AI practitioners to decide for themselves which checklist items to complete at what points in time, which checklist items are not relevant or possible for their AI systems, or which checklist items should instead be completed by another team—all of which may raise new issues of accountability over such decisions [cf. 79, 90]. Future work might explore how RAI resources might account for (or be designed specifically for) these development paradigms, including further support for understanding the potential risks introduced by the large-scale datasets used to pretrain large language models [e.g., 10, 12].

More generally, because large-scale pretrained models are (or are claimed to be [13]) general purpose and task agnostic, they place more burden on downstream developers, users, auditors, and other stakeholders to identify the potential risks posed by these models for *their* particular use cases, domains, and deployment contexts. As such, the articulation work described in this paper may become far more widespread, as, for example, taxonomies of risks for large language models [e.g., 89] manifest in particular ways when those models are fine-tuned and integrated into applications in domains such as finance, healthcare, and education. Downstream developers, users, auditors, and other stakeholders will therefore need to take on more of this articulation work [cf. 86, 87]—although they may encounter issues in doing so due to a lack of information about how pretrained models are developed, including which datasets were used to pretrain them [91].

5.2 Implications for RAI work

We find that contextualizing general-purpose RAI resources for teams' development workflows and their AI systems' particular use cases, domains, and deployment contexts requires sociotechnical skills [cf. 26] that many AI practitioners lack. In her analysis of toolkits, Mattern [52] describes how their design summons a particular type of "*skilled practitioner*" to do the last-mile work of translation and adaptation as toolkits are used across contexts. However, it is worth asking to what extent AI practitioners using RAI toolkits and other resources have the skills required to contextualize those resources for their particular use cases, domains, and deployment contexts—or, failing this, whether they are able to do the relational work needed to bridge disciplinary boundaries [25]. Despite many

RAI toolkits largely providing support for technical practitioners doing technical work [92], the work of contextualizing RAI resources (e.g., toolkits, playbooks, checklists) may require involvement from domain experts and other stakeholders with backgrounds beyond AI [cf. 63]. Moreover, RAI resources that rely on their users having knowledge of programming languages or other technical skills in order to customize or use them [e.g., 28] may further exacerbate these disciplinary divisions.

In addition, we find that the work of contextualizing RAI resources can be a generative process—one that may help *sensitize* team members about RAI work, as well as helping them develop their awareness of RAI issues. Building on Boyd and Shilton [15]'s work on *"ethical sensitivity"* among data scientists, our findings suggest that the contextualization process itself may help AI practitioners develop that sensitivity. That is, by collectively taking time as a team to review an AI fairness checklist and decide which parts to keep, change, or remove, team members may develop more proactive awareness of potential fairness-related harms. In addition, given the well-documented challenges of interdisciplinary communication about RAI [e.g., 35, 37, 60, 61, 64, 78], going through this contextualization process as a team may help develop a shared language around AI fairness, a crucial need identified in prior research on cross-functional, interdisciplinary RAI work [25].

Our findings also raise questions about ownership and accountability in RAI work. Specifically, we find tensions related to ownership and ambiguity over who should be responsible for doing the work of contextualizing the checklist, including the initial work of customizing the checklist and integrating it into existing processes and tools, as well as the ongoing communicative work—the articulation work—involved in translating checklist items into actionable tasks, assigning those tasks to relevant team members, defining the tasks' timelines and priorities, and reviewing and approving the completed work. As Strauss [77] argues, articulation work does not only happen at the beginning of a project, but continues throughout its duration in the frequent, often-invisible, efforts to keep work happening despite contingencies and changes to local circumstances.

However, for RAI work, Schiff et al. [70] and others [e.g., 53, 90] discuss the "*many-hands problem*" of challenges to determining accountability when many teams are involved in designing, developing, and deploying AI systems, each with a potentially different reporting or accountability structure—issues that may be compounded by the "*dislocated accountabilities*" of the modularity inherent in many AI development paradigms [90]. And yet, as Wong et al. [92] point out, many RAI resources are designed for individual users, rather than the more complex collaborative reality of AI development and use [cf. 27, 50, 67]. Our findings suggest that existing discussions about ownership over RAI processes would benefit from a wider lens—including the various types of work involved in contextualizing general-purpose RAI resources, as well as larger sets of stakeholders with varying types of expertise. Moreover, these findings suggest that additional work is needed to understand the organizational incentives that might foster more collaborative, interdisciplinary RAI work, including AI development, evaluation, as well as the use of—and the work involved in contextualizing—RAI resources.

5.3 Limitations

Our study used semi-structured interviews and storyboards in a retrospective contextual inquiry to probe on the work involved in contextualizing a general-purpose AI fairness checklist. This approach may have limited our findings to what participants were able to (or chose to) recall from their previous experiences. To complement this approach, future work should conduct observational studies to observe how AI teams contextualize RAI resources in their work practices *in situ* and over time. In addition, participants were recruited via a combination of purposive and snowball sampling, which limited our participants to people in our (and other participants') networks. The study involved 13 participants, which, although within the norm for sample sizes for qualitative research in HCI [19], limits us from being able to make claims about the representativeness of our sample. Furthermore, the participants were primarily from large technology companies (with some participants

from small technology companies and one public sector agency); future work should include more participants from smaller companies, public sector agencies, and civil society organizations. More generally, our study focused on how AI practitioners contextualized an AI fairness checklist, but many RAI resources are designed for members of the public to use [e.g., 44]; as such, future work should consider how members of the public might contextualize RAI resources for particular AI use cases, domains, or deployment contexts. Finally, our study focused on how teams contextualize one specific RAI resource, an AI fairness checklist that we developed in our prior work [49], which we chose because checklists provide well-specified guidelines for practitioners and this particular checklist is publicly available. It remains an open question, however, to what extent the articulation work required for this checklist would arise for other RAI resources, such as fairness toolkits [e.g. 88, 92].

5.4 Conclusion

Many RAI resources, such as toolkits, playbooks, and checklists, have been developed to support AI practitioners in identifying, measuring, and mitigating potential fairness-related harms. These resources are often designed to be general purpose in order to foster greater adoption, but as a result, they are decontextualized, meaning that they lack the level of relevance or specificity needed to use them. In this paper, we investigate how AI practitioners might contextualize one such general-purpose RAI resource, an AI fairness checklist, for their particular use cases, domains, and deployment contexts, and what this work entails. Through a retrospective contextual inquiry with 13 AI practitioners from seven organizations, we identify how this contextualization process introduces new forms of work for AI practitioners and other stakeholders, as well as opening up new sites for negotiation and contestation of values in AI. We also identify how the contextualization process may help AI practitioners develop a shared language around AI fairness, and we identify tensions related to ownership over this process that suggest larger issues of accountability in RAI work.

REFERENCES

[1] Andrew Abbott. 1986. Jurisdictional conflicts: A new approach to the development of the legal professions. American Bar Foundation Research Journal 11, 2 (1986), 187–224.

[2] Sanna J Ali, Angèle Christin, Andrew Smart, and Riitta Katila. 2023. Walking the Walk of AI Ethics: Organizational Challenges and the Individualization of Risk among Ethics Entrepreneurs. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. 217–226.

[3] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2021. What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 249–260.

[4] Karen S Baker and Florence Millerand. 2007. Articulation work supporting information infrastructure design: Coordination, categorization, and assessment in practice. In 2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07). IEEE, 242a–242a.

[5] Kenneth A Bamberger and Deirdre K Mulligan. 2015. Privacy on the ground: driving corporate behavior in the United States and Europe. MIT Press.

[6] Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, W Duncan Wadsworth, and Hanna Wallach. 2021. Designing disaggregated evaluations of ai systems: Choices, considerations, and tradeoffs. (2021), 368–378.

[7] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. https://arxiv.org/abs/1810.01943

[8] Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. 2022. Cultural Re-contextualization of Fairness Research in Language Technologies in India. arXiv preprint arXiv:2211.11206 (2022).

[9] 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. Microsoft, Tech. Rep. MSR-TR-2020-32 (2020).

[10] Abeba Birhane, Sanghyun Han, Vishnu Boddeti, Sasha Luccioni, et al. 2024. Into the LAION's Den: Investigating hate in multimodal datasets. Advances in Neural Information Processing Systems 36 (2024).
[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] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. arXiv preprint arXiv:2110.01963 (2021).
[13] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021).
[14] Theodore S Boone. 2023. The challenge of defining artificial intelligence in the EU AI Act. Journal of Data Protection & Privacy 6, 2 (2023), 180–195.
[15] Karen L Boyd and Katie Shilton. 2021. Adapting Ethical Sensitivity as a Construct to Study Technology Design Teams. Proceedings of the ACM on Human-Computer Interaction 5, GROUP (2021), 1–29.
[16] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. Qualitative research in psychology 3, 2 (2006), 77–101.
[17] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. Qualitative Research in Sport, Exercise and Health 11, 4 (2019), 589–597.
[18] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency. PMLR, 77–91.
[19] Kelly Caine. 2016. Local standards for sample size at CHI. In Proceedings of the 2016 CHI conference on human factors in computing systems. 981–992.
[20] Coleen Carrigan, Madison W Green, and Abibat Rahman-Davies. 2021. "The revolution will not be supervised": Consent and open secrets in data science. Big Data & Society 8, 2 (2021), 20539517211035673.
[21] Henriette Cramer, Jean Garcia-Gathright, Sravana Reddy, Aaron Springer, and Romain Takeo Bouyer. 2019. Translation, Tracks & Data: an Algorithmic Bias Effort in Practice. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, 1–8. https://doi.org/10.1145/3290607.3299057
[22] Yogesh Dabholkar, Haritosh Velankar, Sneha Suryanarayan, Twinkle Y Dabholkar, Akanksha A Saberwal, and Bhavika Verma. 2018. Evaluation and customization of WHO Safety Checklist for patient safety in otorhinolaryngology. Indian Journal of Otolaryngology and Head & Neck Surgery 70, 1 (2018), 149–155.
[23] Asaf Degani and Earl L Wiener. 1993. Cockpit checklists: Concepts, design, and use. Human factors 35, 2 (1993), 345–359.
[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. (2022), 473–484.
[25] Wesley Hanwen Deng, Nur Yildirim, Monica Chang, Motahhare Eslami, Kenneth Holstein, and Michael Madaio. 2023. Investigating Practices and Opportunities for Cross-functional Collaboration around AI Fairness in Industry Practice. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. 705–716.
[26] Mateusz Dolata, Stefan Feuerriegel, and Gerhard Schwabe. 2022. A sociotechnical view of algorithmic fairness. Information Systems Journal 32, 4 (2022), 754–818.
[27] Mateusz Dolata and Gerhard Schwabe. 2024. Towards the Socio-Algorithmic Construction of Fairness: The Case of Automatic Price-Surging in Ride-Hailing. International Journal of Human–Computer Interaction 40, 1 (2024), 55–65.
[28] DrivenData. 2019. Deon: An ethics checklist for data scientists. (2019). http://deon.drivendata.org/
[29] Diana Forsythe. 2001. Studying those who study us: An anthropologist in the world of artificial intelligence. Stanford University Press.
[30] Atul Gawande. 2009. The Checklist Manifesto: How to Get Things Right. Metropolitan Books, New York.
[31] Timnit Gebru. 2021. Hierarchy of Knowledge in Machine Learning and Related Fields and Its Consequences. https://www.youtube.com/watch?v=OL3DowBM9uc
[32] Brigette M Hales and Peter J Pronovost. 2006. The checklist: A tool for error management and performance improvement. Journal of Critical Care 21, 3 (2006), 231–235.
[33] Amy K Heger, Liz B Marquis, Mihaela Vorvoreanu, Hanna Wallach, and Jennifer Wortman Vaughan. 2022. Understanding machine learning practitioners' data documentation perceptions, needs, challenges, and desiderata. Proceedings of the ACM on Human-Computer Interaction 6, CSCW2 (2022), 1–29.
[34] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. 2019. Designing for Complementarity: Teacher and Student Needs for Orchestration Support in AI-Enhanced Classrooms. In International Conference on Artificial Intelligence in Education (AIED) 2019 Proceedings. Springer, 157–171.
[35] 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. ACM, 1–18. https://doi.org/10.1145/3290605.3300830
[36] Karen Holtzblatt and Sandra Jones. 1995. Conducting and analyzing a contextual interview (excerpt). In Readings in Human–Computer Interaction. Elsevier, 241–253.
[37] Matthew K Hong, Adam Fourney, Derek DeBellis, and Saleema Amershi. 2021. Planning for natural language failures with the ai playbook. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 1–11.
[38] OECD Legal Instruments. 2019. Recommendation of the Council on Artificial Intelligence. Organization for Economic Cooperation and Development (2019).
[39] Steven J Jackson. 2017. Speed, time, infrastructure. The sociology of speed: Digital, organizational, and social temporalities 169 (2017).
[40] Seyyed Ahmad Javadi, Chris Norval, Richard Cloete, and Jatinder Singh. 2021. Monitoring AI Services for Misuse. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. 597–607.
[41] Christopher M Kelty. 2018. The Participatory Development Toolkit. https://limn.it/articles/the-participatory-development-toolkit/
[42] Seungjun Kim, Dan Tasse, and Anind K Dey. 2017. Making machine-learning applications for time-series sensor data graphical and interactive. ACM Transactions on Interactive Intelligent Systems (TiiS) 7, 2 (2017), 1–30.
[43] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences 117, 14 (2020), 7684–7689.
[44] PM Krafft, Meg Young, Michael Katell, Jennifer E Lee, Shankar Narayan, Micah Epstein, Dharma Dailey, Bernease Herman, Aaron Tam, Vivian Guetler, et al. 2021. An action-oriented AI policy toolkit for technology audits by community advocates and activists. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. 772–781.
[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] Leonardo Leite, Carla Rocha, Fabio Kon, Dejan Milojicic, and Paulo Meirelles. 2019. A survey of DevOps concepts and challenges. ACM Computing Surveys (CSUR) 52, 6 (2019), 1–35.
[47] Kornel Lewicki, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. 2023. Out of Context: Investigating the Bias and Fairness Concerns of "Artificial Intelligence as a Service". In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 1–17.
[48] 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.
[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] Olivera Marjanovic, Dubravka Cecez-Kecmanovic, and Richard Vidgen. 2021. Algorithmic pollution: Making the invisible visible. Journal of Information Technology 36, 4 (2021), 391–408.
[51] Olivera Marjanovic, Dubravka Cecez-Kecmanovic, and Richard Vidgen. 2022. Theorising algorithmic justice. European Journal of Information Systems 31, 3 (2022), 269–287.
[52] Shannon Mattern. 2021. Unboxing the Toolkit. https://tool-shed.org/unboxing-the-toolkit/
[53] Jacob Metcalf, Emanuel Moss, and danah Boyd. 2019. Owning ethics: Corporate logics, Silicon Valley, and the institutionalization of ethics. Social Research 86, 2 (2019), 449–476.
[54] Eric Monteiro, Neil Pollock, Ole Hanseth, and Robin Williams. 2013. From artefacts to infrastructures. Computer supported cooperative work (CSCW) 22, 4 (2013), 575–607.
[55] Jessica Morley, Luciano Floridi, Libby Kinsey, and Anat Elhalal. 2020. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Science and engineering ethics 26, 4 (2020), 2141–2168.
[56] Ziad Obermeyer, Rebecca Nissan, Michael Stern, Stephanie Eaneff, Emily Joy Bembeneck, and Sendhil Mullainathan. 2021. Algorithmic bias playbook. Center for Applied AI at Chicago Booth (2021).
[57] 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.
[58] Chrysanthi Papoutsi and Ian Brown. 2015. Privacy as articulation work in HIV health services. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing. 339–348.
[59] Samir Passi. 2021. Making Data Work: The Human and Organizational Lifeworlds of Data Science Practices. Ph. D. Dissertation. Cornell University.
[60] Samir Passi and Solon Barocas. 2019. Problem formulation and fairness. In Proceedings of the Conference on Fairness, Accountability, and Transparency. 39–48.

[61] David Piorkowski, Soya Park, April Yi Wang, Dakuo Wang, Michael Muller, and Felix Portnoy. 2021. How ai developers overcome communication challenges in a multidisciplinary team: A case study. Proceedings of the ACM on Human-Computer Interaction 5, CSCW1 (2021), 1–25.
[62] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data cards: Purposeful and transparent dataset documentation for responsible ai. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 1776–1826.
[63] Inioluwa Deborah Raji, Morgan Klaus Scheuerman, and Razvan Amironesei. 2021. You can't sit with us: Exclusionary pedagogy in ai ethics education. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 515–525.
[64] 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.
[65] 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.
[66] Rashida Richardson. 2021. Defining and demystifying automated decision systems. Md. L. Rev. 81 (2021), 785.
[67] Lionel P Robert, Casey Pierce, Liz Marquis, Sangmi Kim, and Rasha Alahmad. 2020. Designing fair AI for managing employees in organizations: a review, critique, and design agenda. Human–Computer Interaction 35, 5-6 (2020), 545–575.
[68] Yvonne Rogers. 1997. Reconfiguring the social scientist: Shifting from telling designers what to do to getting more involved. Social science, technical systems, and cooperative work: Beyond the great divide (1997), 57–77.
[69] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-Imagining Algorithmic Fairness in India and Beyond. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (Virtual Event, Canada) (FAccT '21). Association for Computing Machinery, New York, NY, USA, 315–328. https://doi.org/10.1145/3442188.3445896
[70] Daniel Schiff, Bogdana Rakova, Aladdin Ayesh, Anat Fanti, and Michael Lennon. 2021. Explaining the principles to practices gap in AI. IEEE Technology and Society Magazine 40, 2 (2021), 81–94.
[71] Kjeld Schmidt. 1997. Of maps and scripts—the status of formal constructs in cooperative work. In Proceedings of the international ACM SIGGROUP conference on Supporting group work: the integration challenge. 138–147.
[72] Kjeld Schmidt. 2000. The critical role of workplace studies in CSCW. Cambridge University Press Cambridge.
[73] Andrew D Selbst. 2021. An Institutional View of Algorithmic Impact. Harvard Journal of Law & Technology 35, 1 (2021).
[74] 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.
[75] Ben Shneiderman. 2021. Responsible AI: Bridging from ethics to practice. Commun. ACM 64, 8 (2021), 32–35.
[76] Susan Leigh Star. 1991. The Sociology of the Invisible: The Primacy of Work in the Writings of Anselm Strauss. Transaction Publishers.
[77] Anselm Strauss. 1988. The articulation of project work: An organizational process. Sociological Quarterly 29, 2 (1988), 163–178.
[78] 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.
[79] Lucy Suchman. 2002. Located accountabilities in technology production. Scandinavian Journal of Information Systems 14, 2 (2002), 7.
[80] Lucy Suchman. 2020. Agencies in technology design: Feminist reconfigurations. In Machine Ethics and Robot Ethics. Routledge, 361–375.
[81] Lucy Suchman. 2023. The uncontroversial 'thingness' of AI. Big Data & Society 10, 2 (2023), 20539517231206794.
[82] Elham Tabassi. 2023. AI Risk Management Framework: AI RMF (1.0). Technical Report. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1
[83] Helena Tendedez, Maria Angela Felicita Cristina Ferrario, and Jonathan Nicholas David Whittle. 2018. Software development and CSCW: standardization and flexibility in large-scale agile development. Proceedings of the ACM on Human-Computer Interaction-CSCW 2, CSCW (2018).
[84] Emily Tucker. 2022. Artifice and intelligence. Center on Privacy & Technology at Georgetown Law Blog (2022).
[85] Rama Adithya Varanasi and Nitesh Goyal. 2023. "It is currently hodgepodge": Examining AI/ML Practitioners' Challenges during Co-production of Responsible AI Values. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 1–17.

[86] Qiaosi Wang, Michael Madaio, Shaun Kane, Shivani Kapania, Michael Terry, and Lauren Wilcox. 2023. Designing Responsible AI: Adaptations of UX Practice to Meet Responsible AI Challenges. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 1–16.
[87] Zijie J Wang, Chinmay Kulkarni, Lauren Wilcox, Michael Terry, and Michael Madaio. 2024. Farsight: Fostering Responsible AI Awareness During AI Application Prototyping. arXiv preprint arXiv:2402.15350 (2024).
[88] Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. 2023. Fairlearn: Assessing and Improving Fairness of AI Systems. Journal of Machine Learning Research 24, 257 (2023), 1–8.
[89] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359 (2021).
[90] David Gray Widder and Dawn Nafus. 2023. Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. Big Data & Society 10, 1 (2023), 20539517231177620.
[91] David Gray Widder, Sarah West, and Meredith Whittaker. 2023. Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI. Concentrated Power, and the Political Economy of Open AI (August 17, 2023) (2023).
[92] Richmond Y Wong, Michael A Madaio, and Nick Merrill. 2023. Seeing like a toolkit: How toolkits envision the work of AI ethics. Proceedings of the ACM on Human-Computer Interaction 7, CSCW1 (2023), 1–27.
[93] JoAnne Yates and Wanda J Orlikowski. 1992. Genres of organizational communication: A structurational approach to studying communication and media. Academy of management review 17, 2 (1992), 299–326.

Received January 2023; revised October 2023; accepted January 2024