Learning about Responsible AI On-The-Job: Learning Pathways, Orientations, and Aspirations

Michael A. Madaio
Google Research
New York, NY, USA

Shivani Kapania
Carnegie Mellon University
Pittsburgh, PA, USA

Rida Qadri
Google Research
Mountain View, CA, USA

Ding Wang
Google Research
Atlanta, GA, USA

Andrew Zaldivar
Google Research
Los Angeles, CA, USA

Remi Denton
Google Research
New York, NY, USA

Lauren Wilcox
eBay
San Jose, CA, USA

ABSTRACT

Prior work has developed responsible AI (RAI) toolkits and studied how AI practitioners use such resources when practicing RAI. However, AI practitioners may not have the relevant skills or knowledge to effectively use RAI resources—particularly as pre-trained AI models have enabled more people to develop AI-based applications. In this paper, we explore current practices and aspirations for learning about RAI on-the-job, by interviewing 16 AI practitioners and 24 RAI educators across 16 organizations. We identify AI practitioners' learning pathways for RAI, including information foraging and interpersonal learning; the orientations of RAI learning resources towards computational and procedural approaches to RAI; and aspirations for RAI learning, including desires for more sociotechnical approaches to understand potential harms of AI systems—aspirations that can be in tension with organizational priorities. We contribute empirical evidence of what and how AI practitioners are learning about RAI, and we suggest opportunities for the field to better support sociotechnical approaches to learning about RAI on-the-job.

CCS CONCEPTS

• **Social and professional topics** → *Computing education*; **Codes of ethics**; • **Human-centered computing** → **Empirical studies in HCI**.

KEYWORDS

Responsible AI, sociotechnical AI, on-the-job learning, training

ACM Reference Format:
Michael A. Madaio, Shivani Kapania, Rida Qadri, Ding Wang, Andrew Zaldivar, Remi Denton, and Lauren Wilcox. 2024. Learning about Responsible AI On-The-Job: Learning Pathways, Orientations, and Aspirations. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3630106.3658988

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658988

1 INTRODUCTION

Artificial intelligence (AI)[1] systems are increasingly integrated into public life, where they have led to harms, particularly for already marginalized communities [85, 115]. To address this, researchers, policymakers, civil society organizations, and many others have developed principles [70], toolkits [29, 42, 91, 141], playbooks [100, 133], documentation tools [32, 50, 90, 100, 105, 106, 122], and other interventions to lead to more responsible AI (RAI) development practices [23, 37, 54, 93, 130]. Policymakers have begun to formalize such processes as guidelines, standards, or requirements for AI design and deployment [1, 46, 67, 68, 126]. However, recent work suggests that AI practitioners may lack the skills and knowledge needed to incorporate RAI practices into their AI design and development workflows [9, 30, 135, 142]. University courses on ethics in technology and computer science [39, 40, 48, 119] may help train future AI practitioners in potentially relevant topics for addressing RAI issues, but many working practitioners may not have had the opportunity to take university ethics courses [cf. 72, 76]. Although prior research investigated how AI practitioners are engaging in responsible AI [e.g., 6, 64, 81, 102, 132, 135] and how ethics is incorporated into universities' computer science courses [e.g., 39, 40, 48, 101], this paper instead asks how AI practitioners are learning about RAI on-the-job. In this study, we explore:

RQ1: What and how are AI practitioners currently learning about responsible AI?

RQ2: What are AI practitioners' and RAI educators' challenges and aspirations for learning about RAI?

To investigate these questions, we interviewed AI practitioners (across job roles, application types, and domains) with experience with responsible AI (n=16), and people with experience developing learning resources for AI practitioners (n=24), whether in a formal educational role or not. Participants were from 16 organizations

[1]We follow the definition of AI from the U.S. National Institute for Standards and Technology: *"an engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments"* [126]. However, it is important to note that the term AI has been critiqued as a *"floating signifier"*, which may perpetuate beliefs about technological inevitability [17, 104, 125, 128].

of varying sizes and types, including technology companies and nonprofits. In this paper, we identify AI practitioners' learning pathways for RAI, including information foraging and interpersonal learning; we highlight how the orientations of RAI learning resources tend to reinforce computational and procedural approaches to RAI; and we identify practitioners' and educators' aspirations for RAI learning that draws on sociotechnical ways of understanding potential harms and helps learners apply RAI in the workplace. We contribute empirical evidence of AI practitioners' and educators' current practices, challenges, and aspirations for learning about RAI; and we contribute implications for the design of RAI learning opportunities that emphasize the sociotechnical nature of algorithmic impacts and that open space for critical reflection in AI development.

2 RELATED WORK

2.1 Ethics in computing education

Recent calls have urged computer scientists to anticipate and proactively address the potential negative impacts of computing [e.g., 33, 45, 73, 74]. In response, computer science (CS) departments in higher education are embedding tech ethics in their curricula [56, 63, 66, 69, 96, 97, 113, 143] either through standalone courses or modules in standard CS courses [40, 48]. Fiesler et al. [40] reviewed syllabi from hundreds of tech courses with ethics content to identify the topics covered, the instructors, and the hosting departments. A related analysis revealed that only 12% of nearly 200 AI/ML courses included any mention of an ethics-related topic on the syllabus, and among those that did, these topics were covered in the last few weeks of the course *"as time allows"* [48]. Moreover, Raji et al. [101] reported a lack of support in tech ethics courses for cross-disciplinary work, where the language of the syllabi may reinforce a hierarchy of knowledge [cf. 49] that implicitly values *"hard"* or *"practical"* skills from computer science over *"soft"* skills from humanities disciplines. A complementary study explored CS educators' perspectives on tech ethics, finding that while many CS educators felt it was important to teach ethics, they found it difficult to make time to teach it alongside other topics [119].

Some courses focus on ethics in machine learning (ML) and AI, as opposed to CS more generally [e.g., 14, 62, 80, 94, 103, 107, 114, 116, 137]. Weerts and Pechenizkiy [137] discuss the challenges of encouraging engineering students to link engineering and modeling choices to real-world outcomes and impacts. To address this gap, Rea et al. [103] and Orchard and Radke [94] demonstrate how using scenarios and case studies may help ML students better understand and identify the social implications of AI systems. Others, like Lewis and Stoyanovich [80] and Shapiro et al. [114] use stages in a typical AI development lifecycle to foster reflection on ethical issues, using students' personal data [114] and transparency tools as *"objects-to-think-with"* [80]. Meanwhile, Shen et al. [116] and Hod et al. [62] foster dialogue and reflection among students using *"value cards"* [116] and case studies from law and data science to encourage multidisciplinary dialogue [62]. However, formal university courses are not the only pathway for learning about CS, data science, or AI [41, 72, 76, 109, 144]. Thus, in this paper, we investigate working AI professionals' ongoing learning about RAI on-the-job.

2.2 Educational needs for RAI in the workplace

Recent surveys reveal most data scientists and ML engineers acquire and refine their skills on-the-job [71, 72]. Many transition into AI from other roles, learning from online courses [72], from *"practitioner–instructors"* [76], or from other self-directed learning methods [26]. However, this prior work on data scientists' learning pathways has not focused on the ways that data scientists and ML engineers learn about topics related to ethics or responsible AI. Substantial prior work has empirically studied how AI practitioners (including data scientists, ML and software engineers, user experience (UX) researchers and designers, and others involved in building AI products) *engage in* the work of responsible AI in the workplace, including the organizational dynamics and incentives that shape that work [6, 64, 81, 82, 102, 132, 135]. For instance, AI practitioners are conducting assessments of the fairness of models [e.g., 81, 136], documenting datasets and models [e.g., 100], exploring how AI models might lead to harms during UX prototyping and evaluation processes [e.g., 135], and leading adversarial testing of potential model failures [38, 47, 95].

Research suggests that the work of responsible AI entails new forms of work practices that may be outside the norm for traditional machine learning and AI development [9, 13]. For instance, empirical studies of how AI teams are adopting responsible AI practices suggests that members of AI teams are informally taking on educator roles to support their peers' learning, as in AI teams using resources such as the People+AI Research guidebook to learn (and teach others on their team) about human-centered AI [142], cross-functional teams creating their own educational resources about RAI to bridge disciplinary boundaries [30], and UX professionals leading *"responsibility lifts"* at the start of a new project to foster learning about RAI [135]. Recognizing this need for additional learning about RAI topics, researchers have identified trainings as a key dimension of organizational maturity for RAI [60, 131, 134]. At the same time, some companies have developed some resources for informal learning or training about responsible AI [27, 112]; however, it is unclear to what extent these resources are meeting the needs of AI practitioners to effectively engage in the work of responsible AI.

3 METHODS

3.1 Participants

To investigate our research questions, we conducted semi-structured interviews with two groups of participants, for a total of 40 participants across 16 organizations. We recruited both groups of participants using a combination of direct emails to contacts in our professional networks, recruitment messages on email lists and social media, and snowball sampling.[2] We recruited n=16 industry practitioners working on teams designing or developing AI products or services at four technology companies of various sizes, who we refer to in this paper as *"AI practitioners."* Our inclusion criteria was that they reported some prior experience engaging with RAI in their work (e.g., contributing to evaluations of fairness of models, adversarial testing, privacy for generative models, etc), though learning about RAI on-the-job was not a recruitment criterion.

[2]In sections 5.4 and 7.2, we discuss how our positionality may have impacted our sampling approach.

These participants had various roles including software engineers, data scientists, program/product managers, UX researchers and designers, and more, and they worked on AI systems across multiple application areas, such as finance, education, and healthcare. Then, to gain a complementary perspective on learning about responsible AI, we interviewed n=24 people across 13 technology companies, universities, or nonprofits who had developed RAI trainings or learning resources for AI practitioners, whom we refer to in this paper as *"RAI educators."* We recruited people who had developed resources or led trainings for industry AI practitioners—either in their own organization or elsewhere—rather than university students. Participants held a variety of formal roles, some explicitly related to education or RAI, such as technical writer, head of curriculum, or responsible AI lead, while others were AI practitioners taking on educational responsibilities in more or less formal ways.[3] See Table 1 for a summary of participants' roles.[4]

3.2 Semi-structured interviews

We conducted semi-structured interviews with both groups of practitioners, using different protocols for each group. Interviews were an average of 60 minutes, and participants were compensated an average of $54 USD, in either gift cards or donations to a charity, based on participants' choice. For both groups, we started the interview by asking them to describe their role and what responsible AI means to them in their work. We asked AI practitioners about their *learning process* for RAI, including the context(s) for learning (e.g., university, bootcamps, online courses, on-the-job training, etc), their motivations for learning about RAI, the skills and concepts they felt were the most important, and their specific processes, modalities, or pedagogies for learning. We then asked about how they *applied* what they learned in their current work (including challenges to that application). Finally, we asked about their *aspirations* for ideal learning experiences for RAI, and which skills or concepts they wish they had learned (and how). For RAI educators, we asked them to give an overview of the learning resources or trainings they developed about RAI. We asked them to *describe a single resource* in depth, including their motivation for creating it, the intended audience, learning goals, and specific skills or concepts taught, and how they were assessed. We asked about their *design process* for RAI learning resources, such as how they decided on the specific topics, the topics they felt were most important, and which were easier or harder to teach or assess. We asked about their *aspirations* for learning resources for RAI, including the skills and concepts they feel future RAI learning resources should focus on and what may be preventing them from realizing those aspirations. See Appendix A for the interview protocols. When RAI educators were also AI practitioners, we asked about their learning process for RAI.

[3]In section 4.1.3, we talk more about the informal ways that AI practitioners educate each other about responsible AI.

[4]We collected demographic data on participants' gender and race and ethnicity to recruit a diverse set of perspectives. We include the question items and a summary table of responses in Appendix B in the spirit of transparency in research reporting. However, we did not analyze our data with respect to participants' demographics, and we acknowledge the risk that including such information may inadvertently retrench existing biases in computing [120].

3.3 Data analysis

We took a reflexive thematic analysis approach to analyze the interview data, following Braun and Clarke [21, 22]. We started by discussing epistemological trade-offs of different approaches to thematic analysis, deciding to take a reflexive approach [21], as all of the authors are or were employed as researchers at an industry research group (see Section 7.2), and we thus wanted to grapple with and reflect on how our position as industry researchers who have variously contributed to the design of responsible AI tools and resources may impact our approach to data collection and analysis. We first coded the 40 transcripts from the interviews, with all seven authors coding at least two transcripts each, meeting regularly to discuss and reflect on our codes (and the assumptions underpinning them) throughout the coding process. Following the initial coding, we met regularly as a group to inductively generate themes that captured patterns of shared meaning across the interviews. We used the digital whiteboard Mural to iteratively cluster the codes into larger themes, discussing the relationship between codes and themes as we went and resolving any disagreements in synchronous group discussions. For instance, one of the themes that we developed in this process was *"quantifiable and technical solutions are prioritized over other forms of knowledge."* After several rounds of iterating on the themes and their relationship over several weeks, we generated the final set of themes, which we report on in the following sections.

4 FINDINGS

In this section, we describe three high-level themes in our findings. First, we identify three primary pathways by which AI practitioners learn about RAI on-the-job—by applying ethics knowledge from previous experiences; by foraging for learning resources; and by learning from coworkers and impacted communities. Second, we then identify several orientations towards RAI that participants identified and critiqued, including computational orientations to RAI and a procedural orientation that focuses on teaching learners how to use RAI toolkits or how to comply with their company's RAI processes. Finally, we present RAI educators' and learners' aspirations to teach and learn about RAI in more sociotechnical and relational ways, and the organizational pressures that can come into tension with these aspirations.

4.1 Learning Pathways for Responsible AI

4.1.1 Adapting knowledge and skills from previous education and work experiences. Some participants reported applying their prior knowledge of ethics from their university training, which included design, information schools, philosophy of technology, social psychology, and medical ethics—but, with few exceptions, participants did not report learning about ethics in their CS courses, either because they came from different disciplines or their CS courses did not include ethics. Some participants described how they applied skills and knowledge from their previous work experiences—typically in other industries. For instance, participants described grappling with ethical issues in a wide range of contexts: working on data privacy in city government, data ethics in education, medical ethics, and more. Because the focus of this paper is learning on-the-job, we do not discuss details of what these participants

AI Practitioners	Professional Role	RAI Educators	Professional Role
Participant ID contains "P" (n = 16)	Software Engineer (n = 4) Program Manager (PM) (n = 3) Trust and Safety / RAI Lead (n = 3) UX Researcher/Designer (n = 2) Data Scientist (n = 2) Researcher (n = 2)	Participant ID contains "E" (n = 24)	Researcher (n = 6) Program Manager (PM) (n = 4) Curriculum Lead (n = 3) Technical Writer (n = 3) Software Engineer (n = 3) Trust and Safety / RAI Lead (n = 2) C-Suite (n = 2) Data Scientist (n = 1)

Table 1: Summary of participants' groups and roles.

learned in their university programs, but we highlight that applying such knowledge to their current roles is a challenge (which we discuss more in section 4.3).

4.1.2 On-the-job foraging for responsible AI learning resources. The majority of our participants reported learning about responsible AI on-the-job. Although some participants were required by their employer to complete RAI trainings, many practitioners described being self-motivated to learn about RAI. AI practitioners engaged in self-driven learning by adopting *information foraging* tactics [98] to find, compile, and share resources on RAI from different sources. Participants told us: *"everything I know about [R]AI, I just learned on the job...but it was all self-learning..."* (P14), and *"it was much more grassroots, you have to go out and find information"* (E31). Participants foraged for resources either within their company's internal repositories or by searching for external resources. Participants described searching within their company for training videos or other professional development courses, talks, best practice guides, or educational games (P2, P5, P24, E33). Externally, some participants found online courses or other semi-structured resources like bootcamps (P7, P15, E18). Participants also described reading books on algorithmic bias, such as *Algorithms of Oppression* and *Race after Technology* (P11, P12, E40). Participants also described looking for research papers related to fairness, interpretability, or other RAI topics (P6, P7, E10, P23, P24). Finally, participants reported using social media to search for articles or testimonials about harms or failures of algorithmic systems (P4, P9, E10, P27).

Although foraging can be a useful part of sense-making in a new domain [98], participants pointed out several drawbacks and challenges with this approach for RAI. Many participants expressed anxiety about the quality of the information and the reliability of the content they found. There was an aspiration to learn from what some referred to as *"authoritative"* sources of information, although it was not clear what such authority might look like. One PM described wanting to *"know who are folks that regularly publish digestible updates or information as things develop that I can learn from... an authorized dealer of information... to not do a course or read something that is not actually grounded in research, is not grounded in facts"* (P14). Participants' access to information was often dependent on *"influencers"* (P14) whose posts on social media were shared widely: *"part of my job was tapping into the sort of academic influencer community online... and most of it was just following folks on Twitter"* (P14). Thus, their searches were shaped by who they followed or the content that was amplified on their social media feeds. However, social media content that goes viral may not include the most critical or comprehensive topics in responsible AI. Participants also noted how the learning resources they found may depend on the search terms they use, which may require background knowledge that some learners may not have. For instance, one participant described searching for their company's RAI learning resources: *"I suspect that maybe someone newer to [the company], might not know that we use the term responsible AI. I know fairness was a term that was used before. I think human-centered ML has been used... but I'm not sure what someone who's completely new to this space might search"* (E21). This may be exacerbated by disciplinary divisions that lead people to search for—and find—resources that appeal to their disciplinary identity. We return to this in Section 5.3.

4.1.3 Interpersonal learning about responsible AI. Participants also described learning from (and in some cases educating) other people about RAI. This includes learning from coworkers in the form of informal discussions, via *"casual conversations with collaborators or friends"* (P2), *"team chats"* (P9), and *"word of mouth talking with other people"* (P7). Participants also learned about RAI from users, impacted communities, or other stakeholders. For example, a product manager described how social media exposed them to advocacy from artists and impacted communities:

> *"The first time I experienced [responsible AI] was when Lensa launched and everyone was posting their personal portraits on Instagram stories... There was a huge AI strike of a lot of these artists whose image was used to train a model without being acknowledged... and having a family member who worked in that field and has spoke out so clearly against it on social platforms was what first got me thinking of... there's people that helped build this that didn't consent to their art being used."* (P9)

Other participants (particularly in user-facing roles like UX), described learning about impacts on users via *"conversations with real people who are telling me what their issues with a certain product are"* (P34). Other participants described a similar user-focused (P4) or customer-focused (P8) approach by *"talking to the clinicians that are part of our pilot program and seeing how they will use it"* (P6). However, this may privilege issues surfaced by the largest or highest-paying groups of customers or users, rather than communities most severely impacted by a given technology [cf. 81].

Participants also described taking on roles as informal educators for their peers—or even their managers or organizational leaders—while they were learning themselves. Thus, many of the participants who signed up for the study based on their experience creating or delivering RAI learning experiences (i.e., RAI educators) were themselves AI practitioners. As one participant described: *"I was also educating based on what we learned... and everybody just started*

turning to me basically. And so I was happy to consult as best as I can" (P29). Other participants created reading groups or discussion groups, shared newsletters, or created Slack groups to discuss RAI topics with their teams (P4, P6, E11, P14, P25, P37), while one participant described how they *"start[ed] a community of practice"* (P27) with others interested in RAI within their company. Meanwhile, others found opportunities to educate *"senior leadership"* (P2), where they *"have to engage with the founders and educate them on like, 'Hey, this is why this [model output] is actually racist"'* (P37).

Although interpersonal learning was a common learning pathway, participants brought up tensions with this pathway that may impact what and how practitioners are learning about RAI, including differences in values among co-workers that shape what resources or guidance they give or receive from their peers, demands on practitioners' time and other organizational (dis)incentives that may impede whether and how they help others learn [cf. 30, 82, 102, 135, 139], such as power dynamics involved when people are learning from (or trying to teach) their managers or other organizational leaders—particularly if *"you have a team lead who just shuts everyone down"* (E32). As one participant told us: *"at the end of the day, responsible AI is really values-laden and people will have different values and different ideas of what's normatively ideal"* (E18). This potential difference in values and normative ideals may prohibit AI practitioners from engaging in this type of informal, interpersonal learning. As they went on to describe: *"working in a technical role in industry, you learn so much from the people... who've been there for decades... and so from a software engineering perspective, that's how you learn how to become a better software engineer"* (E18). However, exclusively learning from mentors who align with one's values [cf. 16, 49, 101] may inadvertently reproduce the current status quo in AI—we return to this in Section 5.

4.2 Orientations Towards Responsible AI in Learning Resources

4.2.1 Computational orientation to RAI. Because of the differences in values in an interdisciplinary field like RAI [cf. 16, 34, 101], we focus here on the framing, or orientation, of RAI learning resources, which may implicitly communicate to practitioners (particularly those who are foraging for resources) what is important to learn. We find that many RAI learning resources focus on computational implementations of RAI concepts, or what one participant referred to as *"a pure technical approach, [despite] also knowing that [they'll] have to eventually go back and look at how this thing actually affected end users or measure the impacts"* (E19). Some of the most commonly taught RAI concepts and skills involved computational evaluations of a model's performance for different demographic groups using quantifiable fairness metrics and assessment processes (sometimes referred to as dis-aggregated evaluations [e.g., 12, 81]). Similarly, RAI educators are teaching adversarial testing[5] [e.g., 44, 95] to evaluate potential harms of generative models (P2, E10, E16). Learning goals for adversarial testing were often oriented around how to conduct adversarial testing rather than interrogating who is involved in such testing, and for what purpose [cf. 44].

[5]This involves recruiting people to play the role of adversaries, prompting models to produce harmful outputs that system designers can ideally mitigate.

RAI educators described why they took a computational orientation to the topics and pedagogical approaches: to appeal to engineers' disciplinary backgrounds, via teaching with computational notebooks (e.g., Jupyter Notebooks) or framing fairness *"as an optimization problem"* (E20). However, this computational orientation was prevalent even when the target audience for RAI learning was *"non-technical people"* (E30), with numerous participants describing how a pre-requisite to learning about RAI was first knowing how AI systems work, saying: *"the foundational piece of being involved in responsible AI is you have to understand at a very basic level how the technology works"* (E31).

Despite this perceived need to appeal to engineers by translating social or philosophical concepts into computational or quantifiable forms, many participants revealed misgivings that a computational orientation to RAI might unintentionally reinforce technosolutionism, or the belief that technical interventions are able to solve fundamental societal challenges [92]. One RAI educator described an exercise they used with learners, to re-implement a research paper on gender de-biasing in word embeddings, but they shared misgivings that learners might come away thinking that *"'we can solve gender bias by just doing some linear algebra,' and is that the takeaway that we want people to have? But the counterfactual is, if we didn't have this kind of [training], then people probably wouldn't even care about it at all"* (E18). As several participants identified, choices about who is testing systems and for what types of potential harms are never value-neutral [e.g., 31, 44]. Shying away from teaching a socio-political analysis, however, meant that even measuring differences in group-level metrics was difficult to teach, as such evaluations may rely on practitioners asking sociopolitical questions like *"what does the hierarchy of social favorability... the hierarchy of privilege look like in this community?"* (E16).

This primarily computational orientation to RAI impacted pedagogical decisions, such as the choice of learning goals, instructional formats, and ways of demonstrating mastery over the material. We heard about *"the pedagogical challenges of creating content for engineers in areas that are outside the areas that they've been trained"'* (E20). Many RAI educators described reservations about adopting assessments from CS courses: *"for the more technical things, we can just borrow from how other technical projects are assessed...like calculate the group conditional true positive rate or whatever and then you can check if they implemented that correctly. So stuff that's technical is much easier to assess...if you're talking about more qualitative [methods]...it's not a math test where you test whether or not someone got the right answer"* (E18). Others felt pressure to develop quantifiable assessments, reflecting on how *"there's more that we could be doing in the evaluation space, but it's just really hard to figure out how you can quantify and assess that"* (E20). In the absence of readily available approaches to evaluating less computational or quantifiable skills, RAI educators drew on *"anecdotal"* course evaluations (E20), or *"testimonials"* of learners' takeaways from the course (E22).

This desire for a tighter integration of ethics content into ML-focused trainings [cf. 39, 48] was accompanied by language that revealed anxieties about the values that were prioritized in the course design: *"[ideally] incorporating [ethical content] more naturally into everything, not kind of shoving it into people's faces, right? And having it overtake the actual technical concepts that students are learning"*

(E11). Other RAI educators reflected on the difficulty of these pedagogical decisions: *"how do you think about whether something is normatively good or not, right? That's just really hard to do within the bounds of the CS discipline"* (E18). However, left unsaid here is the more difficult normative question of what *should* be within the bounds of the CS discipline, which we return to in the Discussion.

4.2.2 Procedural orientation to RAI: corporate processes and toolkits. In addition to a computational orientation, participants described how RAI learning experiences are oriented around procedures—teaching practitioners how to use RAI toolkits or how to comply with their companies' RAI policies. This included how to use RAI toolkits [see 9, 29, 79, 141, for a review], including fairness toolkits such as Fairlearn [136], AI Fairness 360 [4], as well as transparency tools like Datasheets [50], Data Cards [100], and Model Cards [90]. The focus on toolkits was important for some RAI educators: *"not [just] to teach people what is fairness? What is transparency?... [but] how to actually practice it, here's how to actually implement it"* (E22). Analogizing RAI to agile development, one educator noted: *"it's similar to when you're doing a daily stand-up, the objective is a process objective, it's a 'did we do the process?' 'Yes.' As opposed to an outcome objective... like did the process produce like X number of actionable items or Y number of mitigations"* (E32). Similarly, many RAI educators oriented their trainings around their companies' internal RAI principles and RAI review processes. For instance, one RAI educator *"socialized the [company]'s responsible AI principles and created all sorts of games and challenges that incentivize employees to complete them to try and build awareness. One thing that [my colleague] found was that if you had asked employees if they had heard of the RAI principles, they would say, yes, I know there's AI principles. If you ask them to name just one of them or explain it, they could not do it"* (E20). For them, building awareness of their companies' RAI principles among employees—including naming or explaining what those principles were—was a critical precursor to enacting broader organizational change [cf. 60, 134]. Another RAI educator justified this theory of change: *"practices are the most tangible thing that exist inside of a business in terms of how people experience a business, its values and its decision making"* (E32).

However, a procedural orientation to teaching about RAI toolkits and corporate AI policies may limit AI practitioners to learning only those aspects of RAI that the creators of RAI tools and policies deem relevant (or the types of algorithmic harms that such toolkits and policies are able to address—potentially acting as a *"technology of de-politicization"* [cf. 53, 61, 141]), and not, for instance, focusing on *"making systems more accountable to the public in the form of transparency or having public input into how the systems operate"* (E28). Similarly, participants reflected on the tensions inherent in developing educational resources to teach corporate principles and practices, perhaps at the expense of a focus on values or priorities that may not be aligned with corporate business imperatives [cf. 81, 141], referring to companies' trainings on RAI principles as *"this very weird sanitized version of ethics"* (E18). however, that participant expressed ambivalence, voicing a theory of change that corporations developing AI products *did* need to get their AI developers to align with some set of values or practices, because *'"companies are the ones that are [developing AI] that we need them to [develop responsibly]. I kind of feel extremely ambivalent about that"* (E18). This ambivalence was echoed by others who acknowledged that corporations' RAI processes may be *"PR, but it also makes sense"* (E30).

4.3 Aspirations for Responsible AI Learning Resources

4.3.1 Understanding harms and impacts. RAI educators and AI practitioners shared aspirations for sociotechnical approaches to RAI learning that could integrate RAI topics *"across disciplinary divides"* (E28) and enable practitioners to identify potential social impacts of algorithmic systems early in the design process. Participants wanted to *"shift away from the technical components and more on the social, cultural components"* of RAI (E13), or *"not just the technical angles, but the sociological and anthropological angles... [which is] outside the scope of what [AI practitioners] typically do"* (E20). Some described how they used case studies of AI harms across domains to help learners *"be able to foresee potential harm of an AI case... to understand the ramification and the impact of deploying an AI model within a larger system"* (E30). Others tried to foster the skill of identifying potential harms, either via consulting with product teams, using approaches like value-sensitive design (E28), or trying to make time early in an ideation phase to *"think more widely about all the potential harms as well as the opportunities"* (E32).

Participants also wanted learning resources to help incorporate perspectives from external stakeholders, such as members of communities impacted by AI systems, into RAI design and evaluation approaches, but they felt this was not typically covered by RAI learning resources. Participants reflected on how helping AI practitioners learn about community engagement or participatory approaches [cf. 15, 28, 59] could involve learning theories, methods, and skills for how to, for instance, establish relationships with community groups and *"being involved with the community, listening to these communities and, and just sitting at the same table basically"* (E13)—but this was not a part of typical RAI learning resources. To the extent that current RAI learning resources *do* discuss engaging with stakeholders, participants shared that it is often via adversarial testing or more traditional user research paradigms where the goal is to identify *"issues with a certain product"* (P34), which orients the matters of concern around product improvements, rather than systemic harms or impacts.

4.3.2 Building capacity to engage in RAI in the workplace. Practitioners described feeling unprepared to apply what they learned about RAI to their day-to-day work, due in large part to the value-laden and highly contextual nature of RAI. Some desired more support to be able to have potentially difficult, value-laden conversations with coworkers. For instance: *"I don't know if I feel prepared to go into that conversation [with other AI practitioners] and breaking down some misconceptions that could be harmful. How do you start a conversation about responsible AI?"* (P14). Educators reflected on the ability to identify one's values and how they manifest in design choices for algorithmic systems:

> *"You should know about bias and that's important, but there's a difference between that and knowing, 'I am a software engineer at this company and I know how to articulate how I feel about this thing I'm being asked to build,' or 'I know how to engage with my coworkers in*

> *a way that makes them feel safe and respected'... [or] what should we choose as a target variable, and does it have value-related implications? People might have different beliefs about that."* (E18)

These tensions in values were common, and many educators wanted to help learners develop the skill of recognizing that one's values may be different from others on a product team—and how to negotiate (and ideally resolve) the tensions in those values [cf. 78, 84, 89]. Many participants described how they wanted to learn how to raise issues or concerns about potential harms to their manager or other leadership, but these conversations about values may be difficult due to power dynamics within tech companies [cf. 82, 102, 139, 140]. RAI educators struggled to teach learners how to share concerns with their manager (E36), while learners felt there was *"a business case to be made [for RAI]. It's not just about doing the right thing, which is super important..."*, but they struggled to know *"...how can I justify this to stakeholders?"* (P37).

Participants also described the gap they felt between learning about RAI, and being able to apply this knowledge in their development practices, leading to desires for prescriptive guidance that would reduce the *"burden for [practitioners] to interpret it"* (E21). Yet, both practitioners and educators aspired to build capacity to apply learned RAI concepts and skills to new use cases, domains, contexts, applications. To close this gap, practitioners wanted resources that were situated in real-world examples of harms or tailored to different use cases or domains, via case studies or scenarios of RAI issues. Others wanted customized trainings for different geographic or cultural contexts [cf. 108] to help support AI practitioners who are "*looking [at RAI issues] in this specific country, here's a process for fairness or how to test your models*" (E22).

4.3.3 Organizational tensions in pedagogical aspirations. Finally, RAI educators' aspirations for what they saw as pedagogically beneficial approaches to RAI were often in tension with the incentives or requirements from their organizations. RAI educators described the pedagogical benefits of live instruction, especially in synchronous, small-group learning settings. This included the accountability of showing up for a course led by an instructor, as well as having the ability to ask questions or get help from the instructor for particular topics (P2, E3, E36). Some pointed out the value of being able to *"go off script and bring some of their own personal experiences to the class"* (E20). Participants noted that sociotechnical topics in particular were easier to learn in collaborative, conversational learning settings instead of, e.g., remote, asynchronous learning such as watching a training video or working through a Jupyter notebook:

> *"the sociotechnical concepts are easier to do in person, right? Because you can talk to people about why discriminating by gender is bad... It's a little bit harder to do that in a one-way reading text-based delivery versus the sort of Socratic conversation that can help bring people to the table better."* (E19)

However, numerous RAI educators felt that organizational constraints (e.g., a lack of budget, time, personnel) impacted their decisions about the design of learning resources, making it difficult for them to achieve their aspirations [for similar organizational impacts on RAI work practices, see 6, 64, 81, 82, 86, 102, 132]. RAI educators described how shifting organizational priorities made it difficult to allocate time to create trainings, including teams that created tutorials and trainings being *"re-orged"* into other teams, while others were laid off, or felt that creating educational resources was "*not my job anymore*" (E16) [cf. 6]. Educators noted a tension between their pedagogical aspirations for instructor-led, synchronous learning opportunities, and their organization's pressure to develop RAI trainings for large numbers of employees (i.e., *"scalability"* [cf. 58, 127])—*"most people would prefer some sort of instructor-led experience.... but self-study scales, that's the main asset of it"* (E20).

Similarly, despite some educators' aspirations to create a progression of learning resources from basic to more advanced topics in RAI (e.g., weighing trade-offs between different dimensions of responsible AI, such as fairness evaluations and privacy [7]), RAI educators felt organizational pressure to create *"lightweight"* trainings that could be quickly completed. For instance, *"the first [requirement] is that it was really important to create something that was lightweight, which meant that it did not require a lot of prep and it didn't take a lot of time and it was very easy to understand how to do. So anyone could just take like 45 minutes to an hour and just do consequence scanning"* (E32). This was a common theme across RAI educators, who told us how they *"optimized for speed"* (E21), and how this shaped the types of resources they created: *"maybe it's a short video like a five minute tech talk or some way of synthesizing this rich stuff into a quick way that [they] can absorb it and move on to [their] job?"* (E21). Some RAI educators discussed how they balanced this tension by providing multiple formats for learning resources, in varying lengths and complexity (E35, E40).

Such organizational pressures may have similarly shaped *AI practitioners'* aspirations for learning resources. AI practitioners told us how they wanted learning resources to give them practical, actionable guidance that they could use immediately, a pragmatic desire that was at odds with the desire of many participants (both practitioners and educators) to develop reflexive mindsets and value-driven ways of conceptualizing and designing AI systems, as discussed in Section 4.3.2. RAI educators described how practitioners taking their RAI trainings wanted prescriptive guidance to meet their companies' RAI requirements: *"I'm seeing people say, 'I know we have an [AI review process] that you have to make sure that you do. So just tell me what the thing is so that I can do it... if you can make it clearer, then you're removing some of that, that burden for me to like interpret it"* (E21). However, RAI educators noted that despite this desire from practitioners, there *was* no automated process or single solution to anticipating and avoiding harms of algorithmic systems. Participants told us how there was no *"roadmap"* (E20) or *"quick guide"* (E26) for RAI, or in some cases *"there isn't really a right answer"* (P9) at all. RAI educators described how they wanted to foster critical thinking, to shape new ways of thinking towards a *"cultural shift"* in AI development (E17), to enable *"people to make better decisions in the long term"* (E18). They raised concerns that providing guidance that was overly prescriptive (E13) would encourage a mindset of *"a tick box [approach]"* (E19) [cf. 9, 141]. As others told us, *"what [learners] need is not what they want. Because there's no one answer. It depends on who's your audience, what is your product, what is the risk tolerance of your executives?"* (E16).

5 DISCUSSION

5.1 Implications of learning environments for responsible AI

The environments or sites in which learning about RAI occurs shape the learning process in crucial ways. In our study, we found RAI trainings are often oriented around companies' AI policies or organizational processes for RAI review, which can be understood as one element of organizational cultural change or RAI organizational *"maturity models"* [60, 102, 134]. While such trainings may provide learners with opportunities to directly apply what they learn in their work, corporate sites for learning may also have mixed incentives: trainings may be required by their employers [e.g., 10, 11, 24, 134], but AI practitioners may face challenges applying what they learn, due to the organizational pressures for speed and scale that impede RAI work practices [6, 64, 81, 82, 102]. In our study we see similar incentives impacting opportunities for learning about RAI, via constraints on aspirations for RAI learning resources. On the one hand, learners and educators described aspirations to foster reflexive mindsets [cf. 20] and build capacity to apply learned RAI concepts and skills to new use cases, domains, contexts, applications. Yet a fast-paced development environment contributed to a desire for prescriptive guidance that removes the burden of high-stakes decision-making from practitioners [cf. 139]. The pedagogical approaches RAI educators adopt are also shaped by organizational pressures. For example, participants described desires for curricula of increasing depth and instructor-led collaborative learning. Yet, organizational pressures to develop scalable learning opportunities drove educators towards developing self-study resources and training.

We also found that AI practitioners are learning about RAI in unstructured, self-directed ways outside of their companies—from books, documentaries, blogs, social media, or larger communities of AI practitioners. Much like prior work has found for self-directed learning of web developers [36] and data scientists [72], design choices in learning environments implicitly communicate the objects of concern for the AI community and how such topics should be approached. Prior work on self-directed learning for ML has identified learners' challenges finding the right resources for their learning goals [e.g., 26]—however, given the risk of epistemological bifurcation in learning about responsible AI (i.e., into social and technical goals), these challenges become even more salient. As RAI researchers, we can support informal learning opportunities by developing learning resources that are able to reach a much wider audience of people developing AI models or AI-infused applications in the open-source community. However, our findings raise questions about precisely how such informal learning resources might be designed—if integrated into existing toolkits [e.g., 136], how might they avoid the de-politicization or technosolutionism of that genre [cf. 141]? If integrated into online leaderboards or communities such as Kaggle competitions or HuggingFace [cf. 3], how might such learning resources resist the technical orientation of such approaches, rather than inadvertently reinforcing them?

Finally, interpersonal learning is one key pathway by which AI practitioners learn about RAI on-the-job. Some RAI educators in our study had formal roles related to education (e.g., technical writers), though many participants who developed learning resources did so as side projects or in informal educator roles. This echoes prior work that found practitioner–educator roles are common amongst data scientists [76] and AI practitioners [30, 135, 142]. While the predominance of practitioner–educator roles may speak to under-resourcing and under-investment in RAI education, it also suggests an opportunity to leverage RAI expertise that may be distributed across a company. Our findings suggest opportunities for organizations to support interpersonal RAI learning—e.g., via mentoring programs for RAI, providing support for informal or formal conversations with peers about RAI, or more broadly fostering a community of practice for RAI. However, prior research on corporate *"safety cultures"* [cf. 117] identifies risks of relying on approaches to safety (here, RAI) that put the onus for cultural change on workers, given the organizational pressures that may disincentivize workers from raising concerns about harms that may pose threats to their companies' business models [6, 82, 117, 132, 139, 141].

5.2 Designing sociotechnical learning opportunities for responsible AI

Throughout the interviews, we heard RAI educators grappling with underlying disciplinary tensions via pedagogical decisions about learning objectives, instructional approaches, and methods of assessing learners' understanding. Despite an acknowledgment from many participants that interdisciplinary, sociotechnical approaches to responsible AI were important [cf. 34, 55, 101], the learning design choices that RAI educators in our study described may reinforce a bifurcation between approaches that teach so-called *"technical"* AI concepts and approaches from the social sciences or humanities that grapple with social, historical, and political forces that may shape or be shaped by algorithmic systems.

We saw this disciplinary bifurcation reproduced when RAI learning materials taught sociotechnical concepts (e.g., fairness [34][6]) in primarily computational ways, to appeal to learners with ML expertise. This includes teaching concepts such as fairness in ML in ways that were removed from any social context [cf. 111], treating fairness as a metric for algorithmic optimization rather than, e.g., understanding the historical specificity of marginalization in the context(s) in which AI systems are deployed [108], in which their data was collected or annotated [87, 88], or in which AI development teams were located [cf. 123, 138]. RAI educators' reliance on computational approaches may be a response to the existing disciplinary norms of AI development more generally—and yet, such appeals may smuggle in the positivist, technical values of machine learning [e.g., 16, 55, 101], rather than the aspirations our participants had for more integrated, sociotechnical approaches to RAI. We also found that in cases where resources were designed with explicit learning objectives, those objectives often emphasized technical goals or made a distinction between social and technical goals. Given the self-directed foraging our study found, the lack of integrated sociotechnical learning objectives may lead learners to discover or complete only those learning resources that align with their disciplinary identity, further reinforcing a disciplinary division. Instead, we suggest designing RAI resources around explicitly *sociotechnical* learning objectives and adopting pedagogical approaches that involve case studies, scenarios, problem-based

[6] https://casmi.northwestern.edu/news/articles/2023/measuring-safety-in-artificial-intelligence-positionality-matters.html

learning, or other ways of understanding how harms are situated within particular historical contexts [34, 63, 69, 73, 75, 80, 113, 115].

When developing assessments, or ways for learners to demonstrate mastery of RAI skills and concepts, RAI educators in our study reported tensions between what they felt were *"scalable"* methods for learners to demonstrate mastery, like multiple choice questions or code notebooks (where learners could compare their code to a hidden code block with answers), with modes of assessment perhaps better suited to sociotechnical concepts, such as reflective writing, presentations, group discussions, but which were seen as less scalable. However, our participants were uncertain how to adopt such approaches to demonstrating mastery of RAI concepts, in part due to either their own or their learners' disciplinary training in computer science, or due to pressure from their employers to develop trainings that could be deployed to large numbers of learners across their company. One participant voiced their concern about the challenge of assessing mastery of sociotechnical concepts as *"there's no right answer."* Although this reflects the normative questions at the heart of contested concepts such as fairness or RAI [55], this is a challenge that other fields (e.g., the humanities) have long since grappled with when designing assessments. Our findings suggest that more work is needed to develop modes of assessment that are appropriate to an interdisciplinary, sociotechnical approach to RAI [e.g., 5, 75, 113].

5.3 Resisting hierarchies of knowledge in learning about responsible AI

Our findings echo others' calls to resist and destabilize existing hierarchies of knowledge in AI [cf. 49, 55, 101], to lead to a more *sociotechnical* approach to learning about responsible AI. As numerous educational philosophers have argued, choices about what and how to teach are inherently choices about values [43, 51, 52, 65]—indeed, curricular decisions and educational standards have long been sites of public contestation and negotiation [77, 110]. In AI, in addition to tech ethics courses in higher education [e.g., 39, 40, 48, 119], professional training and on-the-job learning are part of what Bourdieu and others refer to as socialization into an occupational community [8, 18, 19]. In other words, choices about designing learning opportunities for RAI (or AI more generally) communicate to members of the AI field what is important for them to know and be able to do as an AI practitioner. As we discuss in Section 5.2, many approaches to learning about RAI implicitly suggest to AI practitioners of *all* roles and disciplinary backgrounds that fairness and other RAI goals are a technical problem that can be subsumed under model optimization goals like accuracy [cf. 16] or usability, rather than socio-political forces that shape every step of the development, deployment, and use of AI systems [cf. 55]. Thus, technosolutionist orientations to learning may implicitly suggest to practitioners that critical, reflexive work is somebody else's problem [e.g., 6, 20, 55, 61, 84, 132, 141]—as part of the separation of concerns or dis-located accountability that cultures of abstraction and modularity in computer science may reinforce [83, 138]. To resolve this, prior work has suggested integrative, interdisciplinary approaches to AI, including Agre's call for *"critical technical practice"* [2], Turkle and Papert's call for epistemological pluralism [129], among many others [e.g., 49, 55, 73, 84, 101].

It is thus worth asking why, decades after such provocations, the AI field continues to reproduce dominant technical values in the occupational socialization of future AI researchers. We thus ask how the FAccT and RAI community might resist this when developing new ways for practitioners to learn about social impacts of algorithmic systems. How might we as a field shift the professional norms and identity of AI development—or what anthropologist Karin Knorr-Cetina refers to as the *"epistemic culture"* [25] of scientific practice in AI? As one approach, we may look to theories of learning as a subversive activity [43, 65, 99] for inspiration for pedagogical approaches that push learners out of their comfort zones or actively draw on disciplinary or epistemic discomfort as a generative way to support learners' growth [cf. 124].

Radical education philosophers have long argued that formal systems of schooling are likely to reproduce dominant, hegemonic views, rather than leading to more liberatory social change [e.g., 43, 52, 65]. In AI, we see signs of what historians have referred to in other fields as *"normative centering"* [57]—wherein formal RAI training may center technosolutionist, computational approaches. To counter this normative centering of technosolutionism, we call for pedagogical provocations that destabilize dominant, hegemonic values in RAI (or AI more broadly). This may entail drawing on Freire's problem-posing approach to learning that is situated in learners' contexts [cf. 43]—rather than centralized, abstracted, homogenized approaches to teaching a single set of ratified concepts, in ways that may reproduce technosolutionist ideologies about (responsible) AI. Along these lines, Malik and Malik [84] draw on Freire's work on critical consciousness to call for technologists to support one anothers' critical technical awakening via learning communities [cf. 2]. However, radical, liberatory educational philosophies and pedagogies may not fit neatly within corporate sites for learning on-the-job, which may prioritize corporate values and desiderata such as scalability. We do not provide easy answers here, but instead call for a proliferation of radical approaches to fostering critical technical awakenings [84] and the development of sociotechnical RAI knowledge and skills as a core part of the AI discipline.

5.4 Limitations

Our participants were primarily tech workers—future work should explore learning with people outside the technology industry, including nonprofits and civil society, policymakers, and the public, including communities impacted or harmed by AI [cf. 35]. Our inclusion criteria for AI practitioners were those who had some prior experiences with RAI, but this may have led to self-selection of participants already interested in the topic, or who saw their work as RAI. In addition, 35 of 40 of our participants were located in the US, and as such, our findings may be skewed towards the perspectives of US-based AI practitioners; future research should explore cross-cultural perspectives on learning about RAI. Finally, future work might analyze the content of RAI learning resources, which we did not include here in part due to access restrictions.

6 CONCLUSION

As technology companies increasingly integrate AI systems into more facets of public life and AI practitioners attempt to identify, evaluate, and mitigate potential harms of AI systems, it is critical

to understand what and how AI practitioners are learning about responsible AI on-the-job. Via interviews with 16 AI practitioners and 24 RAI educators from 16 organizations, we identify AI practitioners' learning pathways for RAI; the primarily computational and procedural orientations of many RAI learning resources; and practitioners' and educators' aspirations for sociotechnical approaches to RAI learning, impacted by organizational pressures. We close with implications of our findings for learning environments for RAI, implications for the design of sociotechnical approaches to learning about RAI on-the-job, and broader questions for the field about how to foster critical reflection among AI practitioners and resist hierarchies of knowledge in RAI.

7 RESEARCH ETHICS AND SOCIAL IMPACT

7.1 Ethical Considerations

Before recruiting participants, we reviewed participant gratuity amounts, data management plans, and study design documents, including consent forms, with experts in the RAI community with many years of human participant research experience, to review social and ethical implications of our choices. We also aligned these choices with standards of human subjects research within our institution. We followed strict protocols to ensure the confidentiality, anonymity, and privacy of our participants. We collected personally-identifiable information only for the purpose of providing gratuities, and kept that information separate from the study data. After transcribing the interviews, we replaced participants' names with unique identifiers and removed any other potentially de-anonymizing information (e.g., the name of their company). Participation in our study was voluntary. At the start of each session, we walked each participant through an informed consent process and form, sharing the study's purpose and intended use of the data, and suggesting they not share confidential information. We told participants they were free to ask us to move on to a different question if they weren't comfortable responding, and they were able to withdraw from the study at any time with no penalty for them (i.e., they would still receive the gratuity), and we would delete their study data if so, although none of the participants asked to withdraw.

7.2 Positionality

All authors are, or were at the time of conducting the research, employed as researchers at a large technology company. Many of the authors have contributed to the design, development, and implementation of RAI tools, frameworks, playbooks, or other RAI resources. Several authors have partnered with or advised product teams on RAI, such as conducting fairness evaluations, data collection best practices, and more. Multiple authors have created RAI learning resources or conducted RAI training sessions with AI practitioners. We have backgrounds in machine learning, human–computer interaction, critical computing, and tend towards post-positivist or interpretivist traditions in qualitative research [120]. These backgrounds and experiences have shaped our research, through choices about the research questions to explore (e.g., focusing on learning for working professionals, rather than learning in academic settings) and our data analysis (e.g., choices about specific codes, themes, or how to interpret them in light of our epistemological orientations and experiences with RAI in industry). Additionally, our positionality—e.g., all authors were employed at a technology company—may have shaped our approach to recruitment (and thus the set of participants). Of the 40 participants, only two were employed at non-profits and two were employed at a university (but spoke about prior experiences in industry).

7.3 Adverse/Unintended Impacts

Although we intend this paper to open a critical conversation in FAccT and RAI about the implications of learning design choices for practitioners learning about RAI, we acknowledge that this work may have unintended impacts. First, we focus in this paper on on-the-job learning, but we do not mean to suggest that formal learning pathways such as ethics courses in higher education should be de-prioritized or under-invested in. On the contrary, we see informal learning about RAI as a complementary approach to intervening in the occupational socialization of (responsible) AI practitioners. Indeed, we believe there is much that RAI and FAccT researchers can learn from prior research on integrating ethics and critical computing into formal computer science education [e.g., 40, 48, 73, 101, 119, and see Section 2.1 for more detail]. Second, although we identify participants' concerns for disciplinary divisions of RAI work into social and technical, and we argue in the Discussion for integrated, interdisciplinary approaches to learning about RAI, we acknowledge the risk that this paper may unintentionally further reinforce disciplinary divisions within the AI field.

ACKNOWLEDGMENTS

We want to thank our participants for sharing their experiences and the anonymous reviewers for their helpful feedback on the draft. We also want to thank Daniel J. Barrett, Alicia Chang, Samantha Finkelstein, Andrew Smart, Tom Stepleton, Bogdana Rakova, Zijie Jay Wang, Elizabeth Anne Watkins, Hilde Weerts, David Widder, and Richmond Wong for helpful feedback that greatly improved this work.

REFERENCES

[1] EU AI Act. 2021. *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts.* EU AI Act. Retrieved December 2023 from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

[2] Philip E Agre. 2014. Toward a critical technical practice: Lessons learned in trying to reform AI. In *Social science, technical systems, and cooperative work.* Psychology Press, 131–157.

[3] Shazeda Ahmed, Klaudia Jaźwińska, Archana Ahlawat, Amy Winecoff, and Mona Wang. 2024. Field-building and the epistemic culture of AI safety. *First Monday* (2024).

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

[5] Andrea Aler Tubella, Marçal Mora-Cantallops, and Juan Carlos Nieves. 2024. How to teach responsible AI in Higher Education: challenges and opportunities. *Ethics and Information Technology* 26, 1 (2024), 3.

[6] 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.

[7] 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.

[8] Netta Avnoon. 2021. Data scientists' identity work: Omnivorous symbolic boundaries in skills acquisition. *Work, Employment and Society* 35, 2 (2021), 332–349.

[9] Agathe Balayn, Mireia Yurrita, Jie Yang, and Ujwal Gadiraju. 2023. "Fairness Toolkits, A Checkbox Culture?" On the Factors that Fragment Developer Practices in Handling Algorithmic Harms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 482–495.

[10] Kenneth A Bamberger and Deirdre K Mulligan. 2011. New governance, chief privacy officers, and the corporate management of information privacy in the United States: An initial inquiry. *Law & Policy* 33, 4 (2011), 477–508.

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

[12] 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).

[13] Marguerite Barry, Aphra Kerr, and Oliver Smith. 2020. Ethics on the Ground: From Principles to Practice. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 688. https://doi.org/10.1145/3351095.3375684

[14] Benjamin S Baumer, Randi L Garcia, Albert Y Kim, Katherine M Kinnaird, and Miles Q Ott. 2022. Integrating data science ethics into an undergraduate major: A case study. *Journal of Statistics and Data Science Education* 30, 1 (2022), 15–28.

[15] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the people? opportunities and challenges for participatory AI. *Equity and Access in Algorithms, Mechanisms, and Optimization* (2022), 1–8.

[16] 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.

[17] 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.

[18] Pierre Bourdieu. 1990. The intellectual field: a world apart. *In other words: Essays towards a reflexive sociology* (1990), 140–49.

[19] Pierre Bourdieu. 1990. *The logic of practice*. Stanford university press.

[20] 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.

[21] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.

[22] Virginia Braun and Victoria Clarke. 2021. Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches. *Counselling and psychotherapy research* 21, 1 (2021), 37–47.

[23] Kirsten E Bray, Christina Harrington, Andrea G Parker, N'Deye Diakhate, and Jennifer Roberts. 2022. Radical futures: Supporting community-led design engagements through an afrofuturist speculative design toolkit. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–13.

[24] Ann Cavoukian, Scott Taylor, and Martin E Abrams. 2010. Privacy by Design: essential for organizational accountability and strong business practices. *Identity in the Information Society* 3 (2010), 405–413.

[25] Karin Knorr Cetina. 1999. *Epistemic cultures: How the sciences make knowledge*. Harvard University Press.

[26] Rimika Chaudhury, Philip J Guo, and Parmit K Chilana. 2022. " There's no way to keep up!": Diverse Motivations and Challenges Faced by Informal Learners of ML. In *2022 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 1–11.

[27] Paul B de Laat. 2021. Companies committed to responsible AI: From principles towards implementation and regulation? *Philosophy & technology* 34 (2021), 1135–1193.

[28] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–23.

[29] 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.

[30] 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.

[31] Remi Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. 2021. Whose ground truth? accounting for individual and collective identities underlying dataset annotation. *arXiv preprint arXiv:2112.04554* (2021).

[32] Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Remi Denton. 2022. CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2342–2351. https://doi.org/10.1145/3531146.3534647

[33] Kimberly Do, Rock Yuren Pang, Jiachen Jiang, and Katharina Reinecke. 2023. "That's important, but...": How Computer Science Researchers Anticipate Unintended Consequences of Their Research Innovations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[34] Mateusz Dolata, Stefan Feuerriegel, and Gerhard Schwabe. 2022. A sociotechnical view of algorithmic fairness. *Information Systems Journal* 32, 4 (2022), 754–818.

[35] Daniel Domínguez Figaredo and Julia Stoyanovich. 2023. Responsible AI literacy: A stakeholder-first approach. *Big Data & Society* 10, 2 (2023), 20539517231219958.

[36] Brian Dorn and Mark Guzdial. 2010. Learning on the job: characterizing the programming knowledge and learning strategies of web designers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 703–712.

[37] Salma Elsayed-Ali, Sara E Berger, Vagner Figueredo De Santana, and Juana Catalina Becerra Sandoval. 2023. Responsible & Inclusive Cards: An Online Card Tool to Promote Critical Reflection in Technology Industry Work Practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM. https://doi.org/10.1145/3544548.3580771

[38] Hayden Field. 2022. How microsoft and google use ai red teams to "stress test" their systems. *Emerging Tech Brew* (2022).

[39] Casey Fiesler, Mikhaila Friske, Natalie Garrett, Felix Muzny, Jessie J Smith, and Jason Zietz. 2021. Integrating ethics into introductory programming classes. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*. 1027–1033.

[40] Casey Fiesler, Natalie Garrett, and Nathan Beard. 2020. What do we teach when we teach tech ethics? A syllabi analysis. In *Proceedings of the 51st ACM technical symposium on computer science education*. 289–295.

[41] Casey Fiesler, Shannon Morrison, R Benjamin Shapiro, and Amy S Bruckman. 2017. Growing their own: Legitimate peripheral participation for computational learning in an online fandom community. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 1375–1386.

[42] Markkula Center for Applied Ethics. 2024. *Ethics in Technology Practice Framework and Toolkit*. Retrieved Jan 2024 from https://www.scu.edu/ethics-in-technology-practice/

[43] Paulo Freire. 1996. Pedagogy of the oppressed (revised). *New York: Continuum* 356 (1996), 357–358.

[44] Sorelle Friedler, Ranjit Singh, Borhane Blili-Hamelin, Jacob Metcalf, and Brian J Chen. [n. d.]. AI Red-Teaming Is Not a One-Stop Solution to AI Harms. ([n. d.]). https://datasociety.net/library/ai-red-teaming-is-not-a-one-stop-solution-to-ai-harms-recommendations-for-using-red-teaming-for-ai-accountability/

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

[46] G7. 2023. *Hiroshima Process International Code of Conduct for Advanced AI Systems*. G7 leaders. Retrieved Novemeber 2023 from https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-code-conduct-advanced-ai-systems

[47] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858* (2022).

[48] Natalie Garrett, Nathan Beard, and Casey Fiesler. 2020. More than "If Time Allows" the role of ethics in AI education. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 272–278.

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

[50] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[51] Henry A Giroux. 2007. *Border crossings: Cultural workers and the politics of education*. Routledge.

[52] Henry A Giroux. 2020. *On critical pedagogy*. Bloomsbury Publishing.

[53] Charles Goodwin. 2015. Professional vision. In *Aufmerksamkeit: Geschichte-Theorie-Empirie*. Springer, 387–425.

[54] Google. 2024. *Responsible AI Practices*. Retrieved Jan 2024 from https://ai.google/responsibility/responsible-ai-practices/

[55] 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.

[56] Barbara J Grosz, David Gray Grant, Kate Vredenburgh, Jeff Behrends, Lily Hu, Alison Simmons, and Jim Waldo. 2019. Embedded EthiCS: integrating ethics across CS education. *Commun. ACM* 62, 8 (2019), 54–61.

[57] Berndt Hamm and John M Frymire. 1999. Normative Centering in the Fifteenth and Sixteenth Centuries: Observations on Religiosity, Theology, and Iconology. *Journal of Early Modern History* 3, 3 (1999), 307–354.

[58] Alex Hanna and Tina M Park. 2020. Against scale: Provocations and resistances to scale thinking. *arXiv preprint arXiv:2010.08850* (2020).

[59] Christina Harrington, Sheena Erete, and Anne Marie Piper. 2019. Deconstructing community-based collaborative design: Towards more equitable participatory design engagements. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.

[60] Amy Heger, Samir Passi, and Mihaela Vorvoreanu. 2022. All the tools, none of the motivation: Organizational culture and barriers to responsible AI work. (2022).

[61] Zoë Hitzig. 2020. The normative gap: mechanism design and ideal theories of justice. *Economics & Philosophy* 36, 3 (2020), 407–434.

[62] Shlomi Hod, Karni Chagal-Feferkorn, Niva Elkin-Koren, and Avigdor Gal. 2022. Data science meets law. *Commun. ACM* 65, 2 (2022), 35–39.

[63] Anna Lauren Hoffmann and Katherine Alejandra Cross. 2021. Teaching Data Ethics: Foundations and Possibilities from Engineering and Computer Science Ethics Education. (2021).

[64] 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.

[65] Bell Hooks. 2014. *Teaching to transgress*. Routledge.

[66] Diane Horton, David Liu, Sheila A McIlraith, and Nina Wang. 2023. Is More Better When Embedding Ethics in CS Courses?. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*. 652–658.

[67] White House. 2023. *Blueprint for an AI Bill of Rights* . White House. Retrieved May 2023 from https://www.whitehouse.gov/ostp/ai-bill-of-rights/

[68] White House. 2023. *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. White House. Retrieved Novemeber 2023 from https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

[69] Iris Howley, Darakhshan Mir, and Evan Peck. 2022. Integrating AI ethics across the computing curriculum. In *The Ethics of Artificial Intelligence in Education*. Routledge, 255–270.

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

[71] Kaggle. 2021. State of Data Science and Machine Learning. (2021). https://www.kaggle.com/kaggle-survey-2021

[72] Miryung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel. 2017. Data scientists in software teams: State of the art and challenges. *IEEE Transactions on Software Engineering* 44, 11 (2017), 1024–1038.

[73] Amy J. Ko, Anne Beitlers, Brett Wortzman, Matt Davidson, Alannah Oleson, Mara Kirdani-Ryan, Stefania Druga, and Jayne Everson. 2023. *Critically Conscious Computing: Methods for Secondary Education*. https://criticallyconsciouscomputing.org'

[74] Amy J Ko, Alannah Oleson, Neil Ryan, Yim Register, Benjamin Xie, Mina Tari, Matthew Davidson, Stefania Druga, and Dastyni Loksa. 2020. It is time for more critical CS education. *Commun. ACM* 63, 11 (2020), 31–33.

[75] Ari Krakowski, Eric Greenwald, Timothy Hurt, Brandie Nonnecke, and Matthew Cannady. 2022. Authentic Integration of Ethics and AI through Sociotechnical, Problem-Based Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 12774–12782.

[76] Sean Kross and Philip J Guo. 2019. Practitioners teaching data science in industry and academia: Expectations, workflows, and challenges. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–14.

[77] David F Labaree. 1997. Public goods, private goods: The American struggle over educational goals. *American educational research journal* 34, 1 (1997), 39–81.

[78] Christopher A Le Dantec, Erika Shehan Poole, and Susan P Wyche. 2009. Values as lived experience: evolving value sensitive design in support of value discovery. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1141–1150.

[79] 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.

[80] Armanda Lewis and Julia Stoyanovich. 2021. Teaching responsible data science: Charting new pedagogical territory. *International Journal of Artificial Intelligence in Education* (2021), 1–25.

[81] 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.

[82] 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.

[83] James W Malazita and Korryn Resetar. 2019. Infrastructures of abstraction: how computer science education produces anti-political subjects. *Digital Creativity* 30, 4 (2019), 300–312.

[84] Maya Malik and Momin M Malik. 2021. Critical technical awakenings. *Journal of Social Computing* 2, 4 (2021), 365–384.

[85] Sean McGregor. 2020. Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. *arXiv 2011.08512* (2020). http://arxiv.org/abs/2011.08512

[86] 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.

[87] Milagros Miceli, Julian Posada, and Tianling Yang. 2022. Studying up machine learning data: Why talk about bias when we mean power? *Proceedings of the ACM on Human-Computer Interaction* 6, GROUP (2022), 1–14.

[88] Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between subjectivity and imposition: Power dynamics in data annotation for computer vision. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–25.

[89] Jessica K Miller, Batya Friedman, Gavin Jancke, and Brian Gill. 2007. Value tensions in design: the value sensitive design, development, and appropriation of a corporation's groupware system. In *Proceedings of the 2007 ACM International Conference on Supporting Group Work*. 281–290.

[90] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[91] 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. *Sci. Eng. Ethics* 26, 4 (Aug. 2020), 2141–2168.

[92] Evgeny Morozov. 2013. *To save everything, click here: The folly of technological solutionism*. PublicAffairs.

[93] Omidyar Network. 2018. *Ethical OS Toolkit*. Retrieved Jan 2024 from https://rri-tools.eu/-/ethical-os-toolkit#:~:text=The%20Ethical%20Operating%20System%20can,minimize%20technical%20and%20reputational%20risks.

[94] Alexi Orchard and David Radke. 2023. An Analysis of Engineering Students' Responses to an AI Ethics Scenario. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 13 (2023), 5834–15842.

[95] Alicia Parrish, Hannah Rose Kirk, Jessica Quaye, Charvi Rastogi, Max Bartolo, Oana Inel, Juan Ciro, Rafael Mosquera, Addison Howard, Will Cukierski, et al. 2023. Adversarial Nibbler: A Data-Centric Challenge for Improving the Safety of Text-to-Image Models. *arXiv preprint arXiv:2305.14384* (2023).

[96] Justin Petelka, Megan Finn, Franziska Roesner, and Katie Shilton. 2022. Principles matter: integrating an ethics intervention into a computer security course. In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education-Volume 1*. 474–480.

[97] Ajit G. Pillai, A. Baki Kocaballi, Tuck Wah Leong, Rafael A. Calvo, Nassim Parvin, Katie Shilton, Jenny Waycott, Casey Fiesler, John C. Havens, and Naseem Ahmadpour. 2021. Co-designing resources for ethics education in HCI. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–5.

[98] Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological review* 106, 4 (1999), 643.

[99] Neil Postman. 1971. *Teaching as a subversive activity: A no-holds-barred assault on outdated teaching methods-with dramatic and practical proposals on how education can be made relevant to today's world*. Delta.

[100] 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.

[101] 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.

[102] 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.

[103] Stephen Rea, Qin Zhu, Dean Nieusma, Kylee Shiekh, and Tom Williams. 2021. Cultivating Ethical Engineers in the Age of AI and Robotics: An Educational Cultures Perspective. In *IEEE International Symposium on Technology and Society*.

[104] Rashida Richardson. 2021. Defining and demystifying automated decision systems. *Md. L. Rev.* 81 (2021), 785.

[105] Negar Rostamzadeh, Diana Mincu, Subhrajit Roy, Andrew Smart, Lauren Wilcox, Mahima Pushkarna, Jessica Schrouff, Razvan Amironesei, Nyalleng Moorosi, and Katherine Heller. 2022. Healthsheet: Development of a Transparency Artifact for Health Datasets. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1943–1961. https://doi.org/10.1145/3531146.3533239

[106] Negar Rostamzadeh, Diana Mincu, Subhrajit Roy, Andrew Smart, Lauren Wilcox, Mahima Pushkarna, Jessica Schrouff, Razvan Amironesei, Nyalleng Moorosi, and Katherine Heller. 2022. Healthsheet: development of a transparency artifact for health datasets. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1943–1961.
[107] Jeffrey Saltz, Michael Skirpan, Casey Fiesler, Micha Gorelick, Tom Yeh, Robert Heckman, Neil Dewar, and Nathan Beard. 2019. Integrating ethics within machine learning courses. *ACM Transactions on Computing Education (TOCE)* 19, 4 (2019), 1–26.
[108] 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*. 315–328.
[109] Lara L Schenck and Betsy DiSalvo. 2023. From Data Work to Data Science: Getting Past the Gatekeepers. In *Proceedings of the 2023 ACM Conference on International Computing Education Research-Volume 2*. 1–2.
[110] Campbell F Scribner. 2016. *The fight for local control: Schools, suburbs, and American democracy*. Cornell University Press.
[111] 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.
[112] Akseli Seppälä, Teemu Birkstedt, and Matti Mäntymäki. 2021. From ethical AI principles to governed AI. In *Proceedings of the 42nd International Conference on Information Systems (ICIS2021)*.
[113] Ben Rydal Shapiro, Emma Lovegall, Amanda Meng, Jason Borenstein, and Ellen Zegura. 2021. Using role-play to scale the integration of ethics across the Computer Science curriculum. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*. 1034–1040.
[114] Ben Rydal Shapiro, Amanda Meng, Cody O'Donnell, Charlotte Lou, Edwin Zhao, Bianca Dankwa, and Andrew Hostetler. 2020. Re-Shape: A method to teach data ethics for data science education. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.
[115] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, et al. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 723–741.
[116] 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.
[117] Susan S Silbey. 2009. Taming Prometheus: Talk about safety and culture. *Annual Review of Sociology* 35 (2009), 341–369.
[118] Mario Luis Small and Jessica McCrory Calarco. 2022. *Qualitative literacy: A guide to evaluating ethnographic and interview research*. Univ of California Press.
[119] Jessie J Smith, Blakeley H Payne, Shamika Klassen, Dylan Thomas Doyle, and Casey Fiesler. 2023. Incorporating Ethics in Computing Courses: Barriers, Support, and Perspectives from Educators. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*. 367–373.
[120] Robert Soden, Austin Toombs, and Michaelanne Thomas. 2024. Evaluating Interpretive Research in HCI. *Interactions* 31, 1 (2024), 38–42.
[121] Katta Spiel, Oliver L Haimson, and Danielle Lottridge. 2019. How to do better with gender on surveys: a guide for HCI researchers. *Interactions* 26, 4 (2019), 62–65.
[122] Ramya Srinivasan, Remi Denton, Jordan Famularo, Negar Rostamzadeh, Fernando Diaz, and Beth Coleman. 2021. Artsheets for art datasets. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 2)*.
[123] Lucy Suchman. 2002. Located accountabilities in technology production. *Scandinavian journal of information systems* 14, 2 (2002), 7.
[124] Lucy Suchman. 2021. Border thinking about anthropologies/designs. *Designs and Anthropologies* (2021), 17–33.
[125] Lucy Suchman. 2023. The uncontroversial 'thingness' of AI. *Big Data & Society* 10, 2 (2023), 20539517231206794.
[126] Elham Tabassi. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). (2023). https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=936225
[127] Anna Lowenhaupt Tsing. 2012. On nonscalability: The living world is not amenable to precision-nested scales. *Common knowledge* 18, 3 (2012), 505–524.
[128] Emily Tucker. 2022. Artifice and intelligence. *Center on Privacy & Technology at Georgetown Law Blog* (2022).
[129] Sherry Turkle and Seymour Papert. 1990. Epistemological pluralism: Styles and voices within the computer culture. *Signs: Journal of women in culture and society* 16, 1 (1990), 128–157.
[130] Princeton University. 2019. *Dialogues on AI and Ethics Case Studies*. Retrieved Jan 2024 from https://aiethics.princeton.edu/case-studies/
[131] Ville Vakkuri, Marianna Jantunen, Erika Halme, Kai-Kristian Kemell, Anh Nguyen-Duc, Tommi Mikkonen, and Pekka Abrahamsson. 2021. Time for AI (ethics) maturity model is now. *arXiv preprint arXiv:2101.12701* (2021).
[132] 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.
[133] Mihaela Vorvoreanu. 2023. Create Effective and Responsible AI User Experiences with The Human-AI Experience (HAX) Toolkit. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23, Article 534)*. Association for Computing Machinery, New York, NY, USA, 1–2.
[134] Mihaela Vorvoreanu, Amy Heger, Samir Passi, Shipi Dhanorkar, Zoe Kahn, and Ruotong Wang. 2023. *Responsible AI Maturity Model*. Technical Report MSR-TR-2023-26. Microsoft. https://www.microsoft.com/en-us/research/publication/responsible-ai-maturity-model/
[135] 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.
[136] 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.
[137] Hilde Jacoba Petronella Weerts and Mykola Pechenizkiy. 2022. Teaching responsible machine learning to engineers. In *Proceedings of the Second Teaching Machine Learning and Artificial Intelligence Workshop*. PMLR, 40–45.
[138] 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.
[139] David Gray Widder, Derrick Zhen, Laura Dabbish, and James Herbsleb. 2023. It's about power: What ethical concerns do software engineers have, and what do they (feel they can) do about them?. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 467–479.
[140] 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.
[141] 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.
[142] Nur Yildirim, Mahima Pushkarna, Nitesh Goyal, Martin Wattenberg, and Fernanda Viégas. 2023. Investigating How Practitioners Use Human-AI Guidelines: A Case Study on the People+ AI Guidebook. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–13.
[143] Cass Zegura, Ben Rydal Shapiro, Robert MacDonald, Jason Borenstein, and Ellen Zegura. 2023. "Moment to Moment": A Situated View of Teaching Ethics from the Perspective of Computing Ethics Teaching Assistants. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.
[144] Jia Zhu, Stephanie J Lunn, and Monique Ross. 2023. Characterizing Women's Alternative Pathways to a Computing Career Using Content Analysis. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*. 158–164.

A INTERVIEW PROTOCOLS

See this section for a high-level version of the interview protocols used for both groups of participants. Note, however, that for semi-structured interviews, the actual questions asked may differ in various ways from the protocol [e.g., 118], such as follow-up questions to probe deeper on specific topics that became salient later in the interview process.

A.1 Interview protocol for AI practitioners with RAI experience

A.1.1 Overview.

- Can you tell me about your role and your team?
- What does RAI mean to you in the work that you do?
- Can you walk me through a specific project where you addressed responsible AI considerations? Which aspects of responsible AI did you focus on, and why?

A.1.2 Reflections on their RAI learning process.

- How did you learn about RAI? Talk me through your learning process.
- Why did you learn about responsible AI?
- What were more or less effective ways of learning about RAI that you've experienced?
- What were the easiest and hardest RAI skills or concepts to learn?

A.1.3 Applying RAI learning in practice.

- Is what you learned from your RAI learning process relevant in your current work? How?
- Can you give an example of how you applied what you learned about RAI on your project?
- What were some challenges you faced when applying what you learned about RAI to your current work?

A.1.4 Aspirations for RAI learning.

- What would you want out of an ideal RAI learning experience?
- What RAI skills or knowledge do you wish you had learned or been taught?
- What kinds of resources or trainings for learning RAI would you want to have?

A.2 Interview protocol for RAI educators

A.2.1 Overview.

- Can you tell me about your role and your team?
- What does RAI mean to you in the work that you do?

A.2.2 Reflection on RAI training or learning resources.

- Can you walk me through a recent RAI training or learning resource you've developed? What was the focus of it?
- Why did you develop that RAI training or learning resource?
- Who is the intended audience?
- What RAI topics did you include in your training or learning resource?
- How did you decide what topics or goals to develop content for?
- Are there easier or harder topics to develop educational content for? What are they?
- What were the specific learning goals or objectives you had for that content, and how did you determine them?
- How did you assess students' understanding or competence about the topic?

A.2.3 Reflections on their RAI learning process.

- How did you learn about RAI? Talk me through your learning process.
- Why did you learn about responsible AI?
- What were more or less effective ways of learning about RAI that you've experienced?
- What were the easiest and hardest RAI skills or concepts to learn?

A.2.4 Aspirations for RAI learning design.

- If you could go back, and create that training or resource again, what would you do differently?
- What would you want out of an ideal RAI learning resource?
- What skills or knowledge do you think RAI trainings or resources should focus on?
- Are there certain topics or skills you wish you could teach or develop content for, but aren't sure how?
- Are there certain formats or pedagogical approaches you think RAI learning resources should adopt?

B DEMOGRAPHIC QUESTIONS AND RESPONSES

This questionnaire was based on work from Spiel et al. [121] on designing better survey questions about gender.

B.1 What is your gender?

- Woman
- Man
- Non-binary
- Prefer to self-describe
- Prefer not to say
- If you would prefer to self-describe your gender, please do so here:

B.2 With which racial or ethnic groups do you identify?

Mark all boxes that apply.

- White
- Hispanic, Latino, or Spanish origin
- Black or African American
- Asian
- American Indian or Alaska Native
- Middle Eastern or North African
- Native Hawaiian or other Pacific Islander
- Prefer not to answer
- Other:

Participant Group	Professional Role	Gender	Race/ethnicity
AI Practitioners (participant ID contains "P") (n = 16)	Software Engineer (n = 4) Program Manager (PM) (n = 3) Trust and Safety / RAI Lead (n = 3) UX Researcher/Designer (n = 2) Data Scientist (n = 2) Researcher (n = 2)	Man (n = 9) Woman (n = 6) Non-binary (n = 0) Prefer to self-describe (n = 0) Prefer not to say (n = 1)	Asian (n = 8) Black or African-American (n = 1) Hispanic, Latino, or Spanish origin (n = 2) White (n = 3) Prefer not to say (n = 2)
Responsible AI Educators (participant ID contains "E") (n = 24)	Researcher (n = 6) Program Manager (PM) (n = 4) Curriculum Lead (n = 3) Technical Writer (n = 3) Software Engineer (n = 3) Trust and Safety / RAI Lead (n = 2) C-Suite (n = 2) Data Scientist (n = 1)	Man (n = 6) Woman (n = 17) Non-binary (n = 0) Prefer to self-describe (n = 0) Prefer not to say (n = 1)	Asian (n = 5) Black or African-American (n = 2) Hispanic, Latino, or Spanish origin (n = 0) Middle Eastern or North African (n = 2) White (n = 14) Prefer not to say (n = 2)

Table 2: Summary of participants' groups, roles, and demographics. Participants could select multiple options for race/ethnicity.