Designing Responsible AI: Adaptations of UX Practice to Meet Responsible AI Challenges

Qiaosi Wang*
Georgia Institute of Technology
Atlanta, Georgia, USA
qswang@gatech.edu

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

Shaun Kane
Google Research
Boulder, Colorado, USA
shaunkane@google.com

Shivani Kapania
Google Research
Bangalore, India
kapania@google.com

Michael Terry
Google Research
Cambridge, Massachusetts, USA
michaelterry@google.com

Lauren Wilcox
Google Research
Mountain View, California, USA
lwilcox@google.com

ABSTRACT

Technology companies continue to invest in efforts to incorporate responsibility in their Artificial Intelligence (AI) advancements, while efforts to audit and regulate AI systems expand. This shift towards Responsible AI (RAI) in the tech industry necessitates new practices and adaptations to roles—undertaken by a variety of practitioners in more or less formal positions, many of whom focus on the user-centered aspects of AI. To better understand practices at the intersection of user experience (UX) and RAI, we conducted an interview study with industrial UX practitioners and RAI subject matter experts, both of whom are actively involved in addressing RAI concerns throughout the early design and development of new AI-based prototypes, demos, and products, at a large technology company. Many of the specific practices and their associated challenges have yet to be surfaced in the literature, and distilling them offers a critical view into how practitioners' roles are adapting to meet present-day RAI challenges. We present and discuss three emerging practices in which RAI is being enacted and reified in UX practitioners' everyday work. We conclude by arguing that the emerging practices, goals, and types of expertise that surfaced in our study point to an evolution in praxis, with associated challenges that suggest important areas for further research in HCI.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**;

KEYWORDS

responsible AI; industry practice; UX; interview

ACM Reference Format:
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 (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3544548.3581278

*The work was done when the author was an intern at Google Research

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581278

1 INTRODUCTION

Technology companies continue to invest in efforts to work toward responsible design and development of AI, responding to increased demand to account for and mitigate the social risks posed by AI technology. As responsible AI (RAI[1]) efforts become more established as organizational practices [84], an increasing number of individuals and groups within both the private and public sector are responding to RAI-related research and product needs.

Much of the literature focusing on RAI efforts is devoted to understanding and improving machine learning (ML) models, through improvements to data annotation practices [9, 28, 31, 71], model evaluations [32, 56, 88], data and model documentation [36, 47, 83], or building developer-facing tools to help engineers interpret models post-hoc [3, 6, 7, 10, 68, 107]. However, in contrast with a development paradigm in which the design of ML models occurs in parallel with their use in products, in many cases, ML models are increasingly used to power user-facing AI applications developed by wholly distinct product teams—in some cases, by product teams at other organizations [15, 58]. Thus, RAI efforts in industry continue after models *leave the research lab*.

Before deploying AI applications in deployment contexts where they could cause far-reaching societal consequences, many practitioners are undertaking RAI work, in more or less visible, and more or less formal, positions. Although prior RAI research in CHI and CSCW has focused on opportunities to intervene in AI development practices [e.g., 53, 69, 70, 84], it has (with few exceptions) focused on the work practices of data scientists and ML engineers in the model development process, rather than how a wider set of

[1]Efforts to identify and address the ways that algorithmic systems perpetuate or amplify societal inequities and biases have a long history [e.g., 45], with more recent research focusing on issues related to fairness, accountability, transparency, and ethics in AI, or what is often referred to as ethical or "responsible" AI [67, 84]. Although numerous companies, government agencies, and civil society organizations have developed principles and guidelines for ethical or responsible AI, broadly converging on high-level values, there is a wide variety of ways that these principles are operationalized in practice across sectors and organizations [61]. We use the term "responsible AI" by its general definition throughout this paper, although we acknowledge that its meaning and scope remain in flux, and wherever possible, we defer to our participants' definitions and operationalization. However, as we note in the Findings section, participants primarily opted to define responsible AI in terms of the practices they used to enact it, rather than offering a conceptual definition of what it means.

practitioners are involved in *applying* existing models for novel AI applications. Among this wider set of practitioners applying the models, many focus on user-facing and user-centered aspects of technology (e.g., interaction designers, user experience (UX) designers, or UX researchers). Yet, the specific practices of such user-centered practitioners involved in RAI have yet to be identified or formalized, and distilling them offers a critical view into how they are adapting to meet present-day challenges in RAI.

In this paper, we explore an emerging set of RAI practices carried out by user-centered practitioners when applying large models during early stages and refinement of AI application design. To understand these emerging user-centered RAI practices, we conducted interviews with UX practitioners, who are often involved in early-stage AI application ideation, design, prototyping, and evaluation, and with RAI subject matter experts (whom we refer to as "RAI experts" throughout the paper), who often perform evaluation and offer consultation about responsibility in AI projects and products or product features, at a U.S. site of a large technology company that has a large AI function. These two sets of practitioners were actively involved in addressing RAI concerns, formally by RAI experts and informally by UX practitioners, both early in design and refinement of new, AI-based prototypes, demos, and product features. We aim to identify and understand the emerging RAI work carried out by UX practitioners in their design processes, by investigating these practices in light of the RAI work conducted by RAI experts in their formal job capacity. We thus seek to answer two research questions:

RQ 1: How do UX practitioners currently incorporate and examine RAI considerations during early stages and refinement of AI application design given their organizational context?

RQ 2: What challenges do UX practitioners encounter in their current RAI practices during early stages and refinement of AI application design given their organizational context?

We conducted a reflexive thematic analysis of interviews with participants to make the following research contributions:

(1) An identification of three key emerging practices that UX practitioners are developing to meet evolving RAI needs: *a) building and reinforcing an RAI lens, b) responsible prototyping, and c) responsible evaluation of AI applications.*
(2) A reflection on the hidden RAI work UX practitioners are carrying out and ways to support this evolution in UX praxis moving forward.
(3) A discussion of the implications of UX practitioners' current challenges in designing with ML models and the need to reconfigure the role of the user when designing RAI.

We first situate our study with respect to prior work on RAI work practices in industry contexts, existing research on values and ethics in UX practice, and prior literature on designing and prototyping with ML models. After introducing our study and data analysis process, we identify three emerging RAI practices developed and carried out by UX practitioners in order to adapt existing UX practices to meet evolving RAI challenges. These practices—building and reinforcing an RAI lens, responsible prototyping, and responsible evaluation of AI applications—are not linear, but are embedded throughout work practices in iterative ways. Within each emerging practice, we situate the RAI work of UX practitioners' with respect to the RAI work carried out by RAI experts in their formal job capacity, and highlight strategies and techniques that UX practitioners adopted, to sensitize people to RAI concerns, to communicate RAI issues with the teams, to consider potential consequences of users' mental models of AI systems (e.g., over-reliance on AI [cf. 80]), and to surface and mitigate potential RAI issues through an emerging AI prototyping technique called prompt programming.[2] Finally, we present practices that place RAI concerns in direct conversation with traditional user evaluation approaches. We reflect on these findings and what they mean for the evolution of UX praxis in the Discussion, and conclude by highlighting opportunities for further HCI research to support the work of designing responsible AI.

2 RELATED WORK

2.1 Responsible AI Practices Among Industry Practitioners

A growing body of work in HCI examines the work practices of industry practitioners as they address RAI issues during the design and development process [53, 67, 69, 70, 84]—in contrast with prior work that has studied data annotation [9, 28, 31, 71] and model evaluation [32, 56, 88], or developed resources for data and model documentation [36, 47, 83] or post-hoc interpretation of model performance [3, 6, 7, 10, 68, 107]. Prior work has found that AI practitioners seek to uncover fairness issues prior to deployment [53], yet many RAI approaches or mitigations are reactionary, initiated as a result of public relations issues, media attention, or customer complaints [53, 69, 84].

Practitioners have emphasized intervening in training datasets [31, 32, 53], as well as algorithmic mitigations [2], as critical in working toward RAI. However, AI practitioners find it challenging to represent diverse demographic groups in datasets and in fairness assessments, given that practitioners tend to draw on their (often homogenous [106]) personal experiences and perspectives when it comes to RAI [30, 53, 69]. Madaio et al. [2022] also note that the heuristics practitioners used to determine priorities in assessing fairness issues (e.g., perceived fairness severity, perceived brand impact, etc.) could compound existing inequities of AI systems [69]. Given these challenges, practitioners call for more guidance and support (e.g., practices, tools, and other resources) in working toward the design, development, and deployment of RAI [53, 69, 84].

To fulfill practitioners' needs for RAI support, a plethora of toolkits, guidelines, and other resources have been developed [e.g. 1, 4, 7, 10, 54, 68, 72]. Recent work has explored the sociotechnical practice of how RAI toolkits are integrated into practitioners' workflows [30, 70, 73, 114]. For example, Deng et al. [2022] point out that these toolkits are often less prescriptive than required—containing guidance on what to do (e.g., engage stakeholders), but not how to

[2]Prompt programming is a practice leveraged by people who interact and communicate with pre-trained large models through natural language prompts instead of writing programming code. Prompt programming has been gaining popularity as more people interact with large language models (LLMs) like GPT-3 [41] and text-to-image generation models like DALL-E [85] and attempt to understand these models' capabilities and limitations.

carry out those recommendations [30]. In addition, many RAI toolkits are also designed to primarily support specific types of technical work from technical practitioners to address RAI issues, rendering themselves less accessible to contributors with varying types of expertise [30, 114]. Relying on toolkits may also enact a form of techno-solutionism, by promoting or incentivizing solely technical solutions to the *socio*technical work of responsible AI [70, 90, 114].

Existing literature has also highlighted the role of organizational factors in RAI practices [53, 69, 70, 84], including the lack of organizational incentives to address RAI issues (or the *dis*incentives to this work) [53, 70, 84]. Coupled with the lack of clarity over roles and responsibilities in RAI work [72, 84], addressing RAI issues during AI development often relies on individuals who are able to dedicate time to developing and promoting RAI processes—processes that may or may not be formally adopted within the rest of the organization [70, 84]. Identifying, assessing, and mitigating RAI issues in datasets, algorithms, and model behavior often require practitioners (and their organizations) to invest significant time and resources, yet AI practitioners often face time and resource constraints in carrying out their RAI practices (including incentives to ship products on fast-paced timelines) [30, 53, 69, 84, 112]. Many scholars have called for changes in organizational structures and processes, to support practitioners' RAI practices instead of hindering them [53, 69, 70, 72, 84].

2.2 Values and Ethics in UX Practice

Although recent work has focused on AI practitioners' work practices for RAI, substantial prior research has identified the ways that values are instantiated in technology design more broadly [40, 57, 79, 103, 110], as well as through ongoing practices of technology use, appropriation, maintenance, and repair [e.g., 55]. As such, various methods, tools, and theoretical frameworks [e.g., 5, 24, 42–44, 92, 115] have been developed to support designers (including UX practitioners) in surfacing relevant values throughout the design process and bringing those values to bear on design decisions (what JafariNaimi et al. [2015] refer to as the "identify/apply" logic of values in design) [57].

For instance, Value-Sensitive Design (VSD) is a framework intended to support designers in understanding the role of specific values in the design of technology, including how they might be promoted or undermined through a given system [42, 44]. In addition, numerous other methods and tools have been developed to support designers' "values work" in technology design [112], including resources to identify relevant values and surface potential harms of technology, such as Envisioning Cards [43], Judgment Call [5], Timelines [115] and more (see Chivukula et al. [2021] for a review). Although VSD and related methods are not the focus of our work in this paper, they represent a key aspect of the training and resources that may inform how UX practitioners address values as part of the work involved in addressing RAI.

2.2.1 Situated Work Practices of UX Practitioners. As part of a broader turn to practice within HCI scholarship [66], recent research has focused not only on developing methods and tools for designers to use, but has also explored the situated work practices of UX practitioners in engaging with values as part of their everyday work [e.g., 23, 25, 50, 112, 113]. For instance, Chivukula et al. [2021] have identified a set of "identity claims" that UX practitioners articulate for the various roles they take on as part of doing values work in their practice, including learner and educator (i.e., learning and teaching others on their team about ethics), translator (i.e., taking resources about values in one domain and translating them to their own work), and advocate/activist, among others. Situating individual UX practitioners within larger social and organizational contexts, Gray and Chivukula [2019] describe various factors that mediate the relationship between designers' individual ethical awareness and action, including the role of organizational practices [25, 50].

This body of scholarship has identified the social and political work that UX practitioners engage in, in addition to their technical work; including rhetorical work to convince leadership of the value of UX (and of the importance of values work in UX) [e.g., 87]—an issue of critical importance given the relatively lower status of UX and design compared with software engineering roles in large technology companies [112]. Given the role that organizational factors play in mediating UX practitioners' values work in practice, Wong [2021] has identified tactics for "soft resistance" that UX practitioners engage in to create space to address values in their work. In part, this involves making values visible and relevant to others in their organization (similar to the advocate or activist identity described by Chivukula et al. [2021]), as well as working to change organizational norms and practices from within, by tactically expanding the definition of who is considered to be the "user" [cf. 116], through leveraging (and trying to change) organizational goals and priorities (e.g., Objectives and Key Results, or Key Performance Indicators), or more broadly leveraging corporate logics to make a business case for values in UX [112].

However, as Wong [2021] point out, each of these tactics are partial and embody contradictions, in that they are trying to challenge, contest, or change organizational practices, while leveraging those same logics and discourses. Finally, as Wong [2021] identify, the values work of UX practitioners often involves "work outside of the technology design process," work that often involves substantial emotional labor [cf. 96], and which may not be valued as part of their everyday work [113]. Relatively little literature, however, has explored the values work that UX practitioners engage in as part of designing and developing AI applications, or how that work is aligned with the work practices of RAI more generally.

2.3 Designing and Prototyping with Machine Learning Models

Prior research has studied UX practitioners' current practices and challenges when designing and prototyping ML-powered AI applications [38, 117, 119, 120, 122]. Existing work has found that UX practitioners, given their training in HCI and User-Centered Design (UCD), commonly leverage traditional HCI and UCD methods and toolkits when designing with AI, yet these are often insufficient [117, 120, 122]. For example, traditional methods such as Wizard-Of-Oz, sketching, paper prototyping, and rapid prototyping can fail to accommodate the non-deterministic behaviors and opaque mechanisms of AI systems [117, 119, 122].

Prior work has highlighted the need for high-fidelity prototypes that can generate tangible, realistic behaviors, instead of toy scenarios, to elicit user feedback and better assess the potential human and societal impacts of AI applications [54, 119, 122]. Zdanowska and Taylor [2022] note that UX practitioners found that most HCI and UCD methods and tools place too much emphasis on the design and evaluation of the user interface, which is only one of many components when designing with AI [122]. Users' mental models of the AI system [120, 122] and the feasibility and user acceptance of the design [122] are also key considerations for UX practitioners.

When designing and prototyping with AI, UX practitioners also frequently collaborate with technical experts such as engineers and data scientists [82, 97, 118] to gain a deeper understanding of the capabilities and limitations of the models [38, 117, 119]. While high-level abstractions are sufficient for UX practitioners to design AI applications [118, 122], a deeper understanding of ML model functionality could help UX practitioners better envision use cases that may not yet exist [38, 119]. However, this collaboration also poses challenges in that UX practitioners and engineers don't always share a common perspective, language, or workflow [82, 97, 119], leading UX practitioners to take on extra work to bridge disciplinary boundaries through sharing user stories and raw user feedback from user testing video recordings to help engineers understand user needs [98], and sometimes, adapting to and embracing a more data-centric culture to communicate user needs through both qualitative and quantitative metrics [118].

2.3.1 Prototyping with Prompt Programming. To meet the increasing demand for new UX prototyping and design tools when designing with AI, an emerging practice involves prototyping with ML models through prompt programming [20, 27, 60, 121]. Using prompting techniques [27, 60], UX practitioners are able to send natural language prompts to ML models as inputs and interact with the models directly to test-drive their capabilities and limitations [60]. Prior work has found that prompt-based prototyping with large language models (LLMs) helps UX practitioners reduce their reliance on engineers and developers to understand model capabilities, speed up the prototyping process to test out initial ideas and "fail fast" [cf. 117], and better communicate with collaborators using prototypes as boundary objects [60].

Prompt programming is usually conducted with pre-trained, large-scale models such as LLMs [27, 60] and text-to-image generation models [124], which are prone to generate outputs that may perpetuate social stereotypes, toxicity, discrimination, and exclusionary norms [15, 33, 48, 105]. Researchers have been exploring ways to evaluate LLMs and other generative models prior to putting them into use. Common evaluation methods include manually generating general test cases, or tests targeted at specific failure modes [59, 86], as well as automatically generating test inputs using the model itself [46, 81]. However, others have noted that such "behavioral tests" and use of benchmarks to prompt models to intentionally generate harmful outputs (so they can be prevented) often come with pitfalls that render these methods invalid [14]. As such, there are increasing calls for human-centered approaches from HCI and UX practitioners to support the work of identifying and mitigating RAI issues with LLMs [e.g., 12–15]. Yet, insights into how prompt programming can facilitate or hinder UX practitioners' work on RAI has not been explored.

3 METHODS

3.1 Recruitment

To investigate our research questions, we conducted semi-structured interviews with both UX practitioners (n = 15) and subject matter experts in a designated RAI role (n = 8). We recruited participants through snowball sampling at our study site (via direct emails to contacts). Our study site was chosen due to the company's large AI function, as well as the depth of researcher access that could be achieved to participants' work practices and teams. All participants were recruited from the same company that the authors were employed at during the time of the study, a decision we discuss further in section 3.2, below. Inclusion criteria included UX practitioners who had worked on or were currently working on the design, prototyping, user research, or user evaluation of AI applications (prototypes, demos, product features) that were powered by large-scale models.[3] We specifically sought out UX practitioners who were either directly or tangentially involved with addressing RAI concerns as part of their work with these applications. For RAI experts, our inclusion criteria included experts in a formal Responsible AI role, who had experience evaluating or being consulted about responsibility in AI projects and products or product features.

Information about participants' job roles can be found in Table 1. UX participants had an average of 5.6 years experience at the company (*SD*=2.3), an average of 11.8 years working in UX (*SD*=7.3), and an average of 7.8 years working with AI (*SD*=4.4); RAI expert participants had an average of 2.3 years working at the company (*SD*=1.6), an average of 2.4 years working in RAI (*SD*=1.2), and an average of 5.9 years working with AI (*SD*=2.6). We also asked participants to optionally share their gender identity: among UX participants, seven were women, eight were men, and one was non-binary. Among RAI experts, five were women, two were men, and one was non-binary. Each participant was compensated through a donation to their charity of choice valuing $40 USD.

3.2 Data Collection

Our study ran from June through July, 2022. All of the participants worked in the United States, in hybrid or fully remote roles at the time of study; hence, all interviews were conducted virtually through an internal virtual meeting platform. Except for two 30-minute interviews, and one 90-minute interview, all interviews lasted about 60 minutes. All participants provided written consent to participate in the research study before interviews began. Scoping the present work to a specific technology company as a research site allowed us to take advantage of internal AI resources (described below in 3.2.3) to ground interview discussions. As researchers were also company employees, participants could provide more details of their AI projects and context for their RAI practices, while upholding confidentiality and IP protection. However, we

[3]We define large-scale models for the purposes of this paper as machine learning models trained on large amounts of text or image data (e.g., at the terabyte or petabyte scale), with model parameter estimates in the billions.

Table 1: Interview participant information. UX participants worked across different AI product and research areas. Listed AI areas are based on the specific projects/products mentioned in interviews. For RAI expert participants, we list their background.

Participant Group	Professional Role	AI Product/Project Area or Background
UX Practitioners (Participant ID contains "U")	UX Designer (n = 2) Interaction Designer (n = 4) UX Researcher (n = 8) UX Engineer (n = 1)	**AI Product/Project Areas:** Language translation (n = 1) Generative language models (n = 4) Conversational AI (n = 4) Text generation (n = 1) Medical imaging (n = 1) Speech recognition (n = 1) AI-enhanced audio (n = 1) Text analysis (n = 1)
Responsible AI Experts (Participant ID contains "R")	Ethics Advisor (n = 2) Responsible AI Researcher (n = 4) Ethics specialist/reviewer (n = 2)	**Background:** Machine Learning (n = 2) Ethics, Philosophy, and Law (n = 3) Science and Technology Studies (n = 2) Design Ethics (n=1)

did not limit interview discussions to participants' current work experiences; many had experience working in multiple settings and companies, and we prompted them to reflect more broadly on their experiences over the course of their careers.

Given the sensitive nature of RAI topics, we also sought to foster a high level of transparency and trust in our communications with each practitioner, before, during, and after the qualitative interview. For example, we held pre-interview calls and shared research briefs with our participants to answer questions about the research, and shared our data interpretations and findings back with each participant by email, noting their specific quotes in the paper and asking for any feedback before submission.

To inform our research questions, we first observed a conversational AI design sprint.[4] We did not include our observations as research data, but instead followed up with four sprint participants to enroll them in our interview study. We also studied any artifacts that were provided in the interview, including sprint artifacts discussed. We discuss each interview protocol below, and they are each provided in Supplementary Materials. We also describe a prompt programming tool called PromptMaker [60] below, which we used as a probe during interviews with both UX practitioners and RAI experts.

3.2.1 UX Practitioner Protocol. In the interviews with UX practitioners, we asked them to broadly describe the types of UX work they do related to AI-based prototypes, demos, and/or product features, and then to dive into more details by having them walk us through one or two specific projects they had worked on. We specifically focused on how RAI issues surfaced in their work, mitigation strategies or precautions they used to address RAI issues, and how RAI issues might have influenced the project direction. At the end of the interview, we also asked UX practitioners about their thoughts on how RAI processes for early-stage AI application design could be improved.

3.2.2 RAI Expert Protocol. In our interviews with RAI experts, we asked participants in designated RAI roles about their experiences reviewing, analyzing, evaluating, and/or consulting on AI technologies as part of organizational RAI practice. We then asked the experts to walk us through an AI project that they had been involved with, which drew on their expertise, the RAI issues that were present, and how they worked with the project/product teams to address those RAI concerns. Towards the end of the interview, we also asked the experts to envision possible improvements to current RAI practices and possible processes or tools to support RAI in early-stage AI application design.

3.2.3 PromptMaker As a Probe. During interviews with all participants, we introduced and described an LLM prototyping tool called PromptMaker, as described in Jiang et al. [2022]. PromptMaker provides a web-based interface to LLMs, enabling users to interactively write and test LLM prompts.[5] PromptMaker also enables practitioners to remotely execute a prompt. For example, a basic prompt to translate English into French could be created (see example in footnote), then embedded within a prototype to test out a translation feature. Collectively, PromptMaker's capabilities enable practitioners to rapidly prototype and test new AI features in hours or days, without requiring significant machine learning experience.

Many of the interview participants were familiar with PromptMaker and seven UX participants indicated in the interviews that they use PromptMaker in their daily job during early-stage AI application design. In our interviews, we used PromptMaker as a probe to understand how emerging AI design and prototyping tools were shaping UX practitioners' RAI practices and to understand the potential opportunities and challenges new AI design and prototyping tools present for RAI.

[4] In this context, a sprint is a multi-day, time-bound, focused series of collaborative activities in which ideation about a problem or problem space is followed by discussions of design ideas relating to potential solutions, often represented in some form of design representation or prototype, which are then further narrowed down and evaluated with stakeholders or prospective technology users.

[5] At the most basic level, a prompt is simply text-based input to an LLM. For example, the following is a very simple prompt: "The opposite of hot is". LLMs generate text likely to appear after the input text. Thus, given this latter example, an LLM is likely to produce text that starts with the word "cold". This prompt is an example of a *zero-shot* prompt, as it provides no sets of examples in the prompt. In contrast, a *few-shot* prompt includes examples to help steer the model toward a particular type of output, such as in this prompt: "English: Hello. French: Bonjour. English: Goodbye. French:". See [20] for more on prompting.

3.3 Data Analysis

All interviews were video-recorded and later transcribed verbatim for data analysis purposes. To analyze the interview data, we drew on Braun and Clarke's [2019–2021] reflexive thematic analysis approach [17–19]. Reflexive Thematic Analysis is a post-positivist approach that emphasizes researchers' role in knowledge production, including the philosophical stance and theoretical assumptions that they make in informing their data analysis approach [18]. This differs from other qualitative data analysis approaches such as constructing codebooks and establishing inter-rater reliability metrics, which might not offer the flexibility needed for researchers to actively participate in the analytic process in a systematic and rigorous way [18]. A reflexive thematic analysis approach also encourages researchers to collaborate and discuss interpretations throughout the process to facilitate the generation of themes.

Five authors participated in the interview data analysis process and continuously and collaboratively discussed the codes and themes throughout. We followed the analysis process outlined in Braun and Clarke [2006]. First, all five researchers familiarized ourselves with the data by reading through the transcripts and taking notes. We then began generating initial codes and divided the transcripts among the five researchers—each interview transcript was read and used in initial code generation by at least two researchers. Each researcher reviewed nine to 11 transcripts at this phase and each generated hundreds of open codes. After initial codes were generated, we frequently met to search, review, discuss, and define themes based on initial codes. In the early stages of our codes-to-themes process, we generated four domain categories (e.g., early-stage RAI challenges, RAI conceptualizations, RAI strategies and practices, aspirational RAI and improvements) and 53 preliminary themes through continuous discussions and iterations. All data was analyzed using shared spreadsheets and text documents. In parallel, two authors reviewed artifacts provided by participants, such as design sprint materials, framework documents, and user study findings, to understand practices associated with early-stage AI application design. After discussing observations together, they shared them with the entire study team.

After further review and discussion of all observations and themes, we distilled three emergent RAI practices: building and reinforcing an RAI lens, responsible prototyping, and responsible evaluation of AI applications, each of which comprises two high-level themes, which we present in the Findings section.

3.4 Author Positionality

Our author team is comprised of researchers with both academic and industry research backgrounds, with varied professional experiences that shape our perspectives. All researchers were employees of the company that served as the research site during the research period. One author has experience in an industry UX role at a large technology company. Two authors have experience developing applications that incorporate large-scale ML models.

Two authors were born in and are currently, or have previously, lived in APAC countries, and four identify as white Americans. All authors completed the bulk of their research training, and work in, predominantly Western institutions. Five identify as having experience with marginalization in computing, either as a member of a marginalized group themselves and/or through many years of conducting HCI research with marginalized groups.

The authors' background and experiences influence our positionality: as HCI researchers trained and working in predominantly Western organizations, we acknowledge that complementary scholarship related to our research questions is needed, to extend and further the understandings presented in this paper. Our positionality has also influenced the subjectivity inherent in framing our research questions, a snowball sampling approach that makes use of our professional networks, the study protocol design, and our data interpretation and analysis.

4 FINDINGS

Through our data analysis, we identified key emerging practices that UX practitioners carried out to meet evolving RAI needs. These practices are not linear, but are embedded throughout their work in iterative ways. In this section, we present three emerging RAI practices: building and reinforcing an RAI lens, responsible prototyping, and responsible evaluation of AI applications. For each practice, we first introduce how RAI experts (labeled with 'R' throughout) and UX practitioners (labeled with 'U' throughout) enact and operationalize RAI.[6] We then highlight UX practitioners' RAI practice in light of the RAI work conducted by RAI experts in their formal job capacity, and present how their practices, goals, challenges, and aspirations for the role of UX in RAI take shape.

4.1 Building and Reinforcing an RAI Lens

We begin by highlighting an overarching practice that influences our subsequent findings: the importance of seeing responsibility as a "lens". An RAI lens positions RAI as a reflexive, ongoing, and holistic perspective that influences practices and decisions throughout design and development, as they arise in context. It is thus not bound to a specific artifact, protocol, or type of analysis of data or model outputs—though it can incorporate them. Instead, it represents an ongoing, shared mindset that acknowledges and seeks to account for the social position of those who shape technology—and the company itself—and the intersecting relationships between specific design and development choices and their societal implications.

As such, both RAI experts and UX practitioners went to great lengths to cultivate and reinforce an RAI lens, not only in their work, but in their teams and the broader culture. This practice was carried out explicitly by RAI experts who inhabited established and prominent RAI positions with focused RAI expertise (e.g., ethics, philosophy, law) in the company, and implicitly by UX practitioners who considered RAI as an emerging yet crucial piece of the UX of AI projects/products.

Below, we characterize the work both groups of practitioners carried out to pursue this goal of an RAI lens: from self-education and sensitization to RAI issues, to actively communicating and working with teams to surface and mitigate RAI issues.

4.1.1 Sensitizing to RAI Concerns. Both UX practitioners and RAI experts were often called upon by product teams to address RAI issues: UX practitioners given their expertise in accounting for the

[6] Although we asked RAI experts and UX practitioners how they defined responsible AI, they primarily responded with descriptions of how they enacted and operationalized RAI in practice in their company, rather than providing a conceptual definition.

impacts of technology on users; RAI experts given their formal job positions and expertise in RAI and ethics. While RAI is not part of UX practitioners' formal training and role, by accumulating experience and knowledge of technologies' impact on users, RAI had already been integrated into many UX practitioners' day-to-day practices and became what UX practitioners described as a *lens*: *"With my lens, I make [RAI issues] surface, so I'm not sure that [RAI issues] would just naturally bubble up to the surface. But I think part of my research ethos is to look for those gaps and those red flags."*(U7)

For RAI experts, a large part of their job included sensitizing [cf. 16] teams across the company to RAI concerns by offering RAI resources and support such as RAI reviews, office hours, consultations, and RAI workshops. When asked about their processes of RAI review and consultations, many RAI experts mentioned the usage of RAI frameworks and guidelines, many of which they (or their colleagues) had developed. RAI experts not only used these RAI frameworks and guidelines to sensitize teams to RAI considerations, but also to help themselves be more aware of potential RAI issues in their work. For instance, R6, R8, and their colleagues are creating an AI harms framework, outlining domain-specific RAI considerations, and best practices to help engage practitioners with RAI issues. R8 told us:

> *"I've found [the AI harms framework] to be pretty helpful for me because even if I do this all the time, I can still forget about one of these possible negative implications that could happen. But teams also like it because they tend to be new to these ideas, so it's helpful for them to just be like, here is the landscape of things that could happen in a pretty digestible format."*

These RAI frameworks, guidelines, and best practices established by RAI experts were commonly referred to by UX practitioners during their interviews. Given that RAI was not part of their formal training or job requirement, to sensitize themselves to RAI considerations, UX practitioners often explicitly and actively sought out these internal RAI resources as well as external literature to better understand the issues and integrate them into their work.

However, a few UX practitioners also mentioned building frameworks from scratch to meet the specific needs of an application domain. Similar to RAI experts' practices, many of the UX practitioners in our interviews talked about their efforts to compile their RAI knowledge and experience into actionable RAI guidelines and best practices. Some UX practitioners even integrated RAI into their standard practices and design pipeline: *"One of our team's RAI standard practices is just how we collect data, that we're putting in a process on collecting data fairly [...] [F]olks have built out a much more rigid pipeline for how we collect that stuff [...] it's just part of the standardized practice now."* (U14).

To further reinforce their RAI perspective during the design process, UX practitioners implemented *responsibility lifts*, a series of activities at the beginning of the design process to reinforce RAI as a lens to inspect and filter design ideas. U4 talked about a design sprint on ideating about AI products powered by LLMs: *"[RAI] was called out in the brief of this sprint as a whole to think about responsibility for any ideas that we come up with. And what are the responsibility implications of those?"* Similarly, U14 who organized the conversational AI design sprint and invited internal guest speakers to talk about RAI at the beginning of the sprint, explained his rationale during the interview:

> *"Even just doing stuff like that [having an RAI lightning talk at the beginning of the sprint], [...] having a space put in to any workshop moving forward, or any team doing this kind of work with AI and everything, you make the space for someone to give a presentation like that, where it can at least put those ideas in the designer, and the prototyper, and the engineers' brains, so it is at least top of mind [...] you make space for responsible design and thinking when you're in the thick of it."*

4.1.2 Organizational Challenges of Communicating about RAI with Teams. As RAI is a rather new and still-emerging discipline, those with expertise are limited, and practitioners with demonstrated expertise are often called upon to help shepherd projects through the examination of RAI considerations and mitigation strategies. Many RAI experts and UX practitioners in our study were members of centralized teams and were often positioned to apply RAI expertise horizontally. As such, both RAI experts and UX practitioners described "dropping in" and then out of specific teams to conduct RAI work in particular phases.

Due to this form of centralized organizational structure (rather than, for instance, embedded RAI experts and UX practitioner with permanent roles on a single AI team), many RAI experts and UX practitioners thus invested significant time in sensitizing members of the teams they were "dropping in[to]" to RAI considerations. Oftentimes, RAI experts and UX practitioners joined AI projects that were still relatively early in development, but which had already begun, making it more difficult to reverse decisions that had already been made or shape the design direction in fundamental ways, particularly as they worked to understand and navigate the power dynamics among the AI team with whom they were working.

Both RAI experts and UX practitioners talked about how the norms and implicit values associated with the larger organizational culture incentivized "moving fast" and emphasizing positive outcomes of AI systems [cf. 53, 70, 84]. This remained a challenge even for RAI experts who were in formal positions as RAI reviewers or ethics consultants that had more power, sometimes with blocking power, over project/product directions and releases. In their interviews, many RAI experts said they mostly worked with teams that voluntarily reached out to them for RAI consultations and reviews. These teams tend to be more open to suggestions and critiques that RAI experts brought up during the process. Many RAI experts emphasized teams' openness in discussing RAI and initiatives in making changes as paramount when working with the teams on RAI issues:

> *"[RAI consultations] is very much conversation-based. And we steer the conversations. But the important thing is that the interest has to come from [the product/project teams]. It has to be that they want to change their beliefs and behaviors. We can try and impose it, of course. It's always better if this kind of culture change comes internally rather than externally." (R3)*

However, this was not always the case for UX practitioners who had less power to sway the project directions over potential RAI

concerns. UX practitioners also worked closely with the teams on the design and development of the products on a day-to-day basis and faced the same time constraints alongside the product teams. In the face of these pressures, UX practitioners described using a combination of communication techniques and product team activities to reconcile these incentives for speed with the introduction of new, more intentional RAI processes that encourage reflection on a range of potential outcomes and harms of AI systems.

Here, UX practitioners emphasized the importance of communicating potential RAI issues strategically and presenting themselves in a supportive role, rather than being seen as a "blocker." U8 told us, *"Making it as blameless as possible is one of the best things we've learned. I'm not trying to tell anyone they're bad, I'm not trying to freak anyone out. I more want to highlight, here's something that could go wrong. Here are the ways that people could be affected, here's the ways the business could be affected. You can choose to act on that or not, but I would strongly advise you to do so."*

In the face of organizational pressures to ship products rapidly, UX practitioners took on additional hidden work to sensitize their team and organizational leadership to the potential harms of AI systems and the importance of mitigating those harms prior to deployment—crucial labor that was not always recognized as being core to their work by their organization.

When not directly negotiating with team members or organizational leadership about RAI concerns, UX practitioners would sometimes create activities or documents meant to enable team members to respond to RAI issues. One strategy included showing potential impacts through user testing that surfaced concerns, or by creating artifacts that illustrate potential harms. For example, U12 had difficulty getting their team to respond to potential concerns and ran a team activity to create fake newspaper headlines [cf. 115] to illustrate how things could go wrong: *"I had put together a deck of like fake headlines of how this could go wrong. [...] I think at the time I think it had a big effect. People backed off the idea, we didn't have to go any further with it."* This activity and others like it were part of the additional work that UX practitioners took on to sensitize others in the AI teams they worked with (and the company more broadly) to the range of potential harms from AI systems.

4.2 Responsible Prototyping: Ideating and Building with Machine Learning Models

During the interviews, both sets of practitioners described how they understood RAI in terms of anticipating and surfacing harms that AI technology could bring to the end-users, society, and the public. They both brought up similar sets of harms in their interviews that they tried to anticipate and surface, which included, but were not limited to: safety, misinformation, inequity and/or culture and identity erasure, stereotyping, over-reliance on AI, anthropomorphism of AI, toxicity and/or offensiveness, and privacy.

However, RAI experts and UX practitioners approached this operational definition of RAI differently based on their respective methods and perspectives. In their interviews, RAI experts told us that given their expertise and the nature of their jobs, they were often invited by the teams to explicitly and intentionally look for RAI issues and concerns, often through question-asking and adversarial testing. As R8 described, *"I think the vast majority of people are not thinking adversarial-y. I'm here to think about the worst of the worst. That's what I was hired to do."* When asked about how they helped the teams to surface harms, R6 said, *"It's honestly just asking questions. This job is mostly just knowing what questions to ask. It's just critical thinking, and it's helping other teams think critically about their projects. We're always asking, what is the worst-case scenario?"*

In contrast, while UX practitioners were not trained or required to surface and anticipate potential AI harms in their day-to-day work, we found that UX practitioners were committed to surface and anticipate harms from their unique human-focused perspective and with their unique skills, methods, and tools during their design and prototyping process.

In this section, we describe how UX practitioners anticipated potential harms by envisioning how user interface design decisions could influence the users' mental model of AI, and how they attempted to surface harms by leveraging both traditional and new approaches to design and prototyping.

4.2.1 Considering the Consequences of Users' Mental Models of AI. During the interviews, UX practitioners told us that due to the stochastic nature of ML models they typically work with (in particular, generative language models, which often include some stochasticity to aid in producing variety in the model outputs), they were often concerned about how users would perceive AI applications driven by these models. In part, UX practitioners were concerned that users might overestimate the capabilities of the language model or treat the model as if it were human-like (i.e., anthropomorphizing it [80, 104]). UX practitioners felt responsible for appropriately communicating model capabilities to users, both through how they designed the user studies as well as the design of the application interface. For example, U1 discussed the potential misinterpretations of AI technologies such as LLMs: *"I think one of the other things about this technology [LLM] that is like, really a misnomer, is that, a lot of people see it as this, like general-purpose chatbot first and then, 'oh, it can do all this other stuff.'"*

To mitigate the potential consequences of people's mental models of AI technology, UX practitioners leveraged different design strategies and techniques. During the interviews, UX practitioners mentioned techniques such as putting constraints on user input and model output, to limit model behavior (e.g., preventing it from generating answers to off-topic user questions) in AI applications; designing LLM-driven AI applications that are not chatbots to expand people's understanding of LLMs; or changing the appearance of the UI of a chatbot application when they discovered users' expectations of an LLM-based chatbot didn't match the actual language interaction: *'So we had to reconfigure how we show these prototypes visually and also how we communicate about what these prototypes are. And that [higher-fidelity UI] was vital to change up. So the team quickly responded and changed the way that it looked, and they changed it from looking like a product to looking more like, terminal, like a very simple, old-school Lo-Fi terminal chatbot."* (U7).

Other UX participants brought up the need to build features and functions to allow users to dig deeper into the AI technology's capability, to improve their understanding of the model. For example, U14 said *"Providing the insight for the user on what features and capabilities are possible helps expand the user's mental model of how*

they can interact with the system. That all develops trust [built on] knowing what's possible and what they can use the system for."

During her interview, U10, a UX researcher, also reflected on the UX practitioners' role and obligation in design to help users with low technology literacy understand the AI system and avoid overestimating the model's capability or erroneously anthropomorphizing the LLM:

> *"[W]e might have many users who don't even know what AI or ML is. And so there were definitely components that I think [raised] a very big, open question for me, in interviews where I asked people how they thought it worked, and they'd be like, 'There's a person that's looking at these, and they get back to you really quickly.' And so, there's this question about what obligation, if any, do we have to correct that misconception, and what potential downstream harm could having that misconception lead to, and what might onboarding to a system like this look like for people who have lower tech literacy? [...] How do you help people understand that this is an AI system and what that means?"*

4.2.2 Examining ML Models through "Test-Driving". A large part of UX designers' job during early-stage AI application design, besides representing the user perspective as part of standard UX practice, was to understand the potential harm and capability of the ML model underlying the AI application. To do this, UX participants described how they supplemented traditional UX design and prototyping methods with novel emerging methods to surface potential RAI issues that traditional methods may not have been able to uncover.

Several participants mentioned using traditional UX design methods such as Wizard-of-Oz studies or toy examples to quickly test out their AI design ideas. However, they also pointed out that these traditional design methods are flawed, as their ability to surface real-world RAI issues is limited. The scope, time frame, and researcher supervision constraints of typical user studies don't allow users to bring *in situ*, authentic personal data, needs, and use cases to bear on model interaction (U3). In part, the organizational factors that we described in section 4.1.2 impact UX practitioners' ability to conduct more longitudinal studies of the harms of AI systems in a more ecologically valid context. To work around those constraints, UX practitioners used toy examples they believed to be representative of usage scenarios.

However, as described by U10, the use of toy examples during user studies makes it difficult to surface RAI issues based on real-world usage:

> *"[O]ne of the hardest parts of prototyping a tool like this, is that it's a highly personal experience when you have a health condition, and your level of anxiety might be incredibly high or your level of investment in finding relevant information may be way higher than you can get in an actual study [using toy examples]. And so, that's been a big challenge for us, is reading the tea leaves a little bit..."*

Many of the UX participants in our study were able to get access to emerging UX prototyping tools that allowed UX practitioners to interact with the ML models directly. For example, the PromptMaker tool we provided in the interview was already used by several UX participants. During prototyping activities that involved model prompting, UX practitioners performed what many referred to as "test-driving," which refers to the activity of continuously probing model outputs using different natural language prompts (that serve as input to the model) to understand model capabilities and limitations, in order to determine fit between a design idea and model capabilities. UX practitioners considered good prompt design to be an "art form," in that it was difficult to identify reasons why prompt construction led to certain model outputs, making repeatable best practices and rules hard to distill. U5 recounted:

> *"I think it's [PromptMaker] a great tool to 'scratch pad' with, to move really quickly to just try something new. But generally, prompting is really weird. It's a kind of weird art form. Little weird things like just having one space key at the end of your prompt can change the complete output of what comes out. That's really subtle and hard to have somebody to understand what that means. So it's a useful tool, I love it, it was the most accessible thing I think that we've created so far. But it's not just self-service yet, or intuitive enough on its own."*

This lack of consistency and intuitiveness of model prompting made it challenging to come up with a consistent practice to address the need to test-drive model interaction during early-stage AI application design. During his interview, U3 also raised the issue that prompting could further introduce bias, requiring intentional efforts to avoid *"codifying our own belief systems. I think what happens is the minute you get a new user that starts to ask questions to your system, then we need different sets of belief systems. And so if you're not thoughtful to [that] fact, your prompt might represent your beliefs in a stronger way..."* To mitigate this RAI issue during responsible prototyping, U3 said he liked to crowdsource few-shot[7] examples, and intentionally eliminated prompts that looked too similar: *"Because what I'm really looking for is, I'm looking for few shots that represent a diversity of types of input that might come in. And sometimes I even model, slightly adversarial examples."* However, most UX practitioners currently lack a standard practice to address this, beyond trying to diversify few-shot examples or deferring to user testing, which we discuss below.

4.3 Responsible Evaluation of AI Applications: Involving Users to Assess Responsible AI

In our interviews, both sets of practitioners explained how they saw inclusion of a diversity of experiences and perspectives as a core dimension of RAI. Almost all RAI experts in our study pointed out the importance of user research and evaluation in ensuring accountability and responsibility, as well as validating and reflecting on the social benefits of the AI products during the early-stage design process.

However, several RAI experts called for longer-term engagement with users throughout the AI product lifecycle given that traditional user evaluation and testing might not be sufficient in designing

[7]"Few shots" here refers to the small number of examples that UX practitioners feed into PromptMaker as inputs to the LLM.

RAI. For instance, R5 critiqued the short-term nature of user testing and questioned the assumptions in traditional user testing and evaluation processes:

> *"I think at base level, not making research just be like a short-term, one-off thing, but having it be something that you are doing constantly at all different parts of this design process. Making sure that [...] you are not just going in with pre-thought of categories where you are like, 'Ok, this is how we define failure. Has this failure happened? No, it hasn't. Ok, great.' but kind of like collectively defining some of those terms with the users. I think just having that process be integrated throughout all of these different steps would probably be a little bit more helpful."*

Other RAI experts in our study more explicitly called out their aspiration of taking more participatory approaches to involve users as well as external stakeholders in the design and evaluation of RAI. For example, R2 said, *'I don't think responsible AI can be achieved without some form of robust participation in a nutshell. For me, responsible AI is the extent to which it can be made participatory. Responsible AI is participatory AI."*

Through our interviews with both RAI experts and UX practitioners, we found that while involving users and taking participatory approaches to evaluate RAI remained largely an aspiration from RAI experts, UX practitioners, with their disciplinary perspective towards involving users and other stakeholders in design and evaluation, were already carrying out this aspiration through involving prospective users of a future AI application in the design process to surface and assess RAI issues.

In this section, we outline UX practitioners' processes and challenges for involving users in the process of evaluating potential RAI issues: from preparing the model prior to involving users, to coping with unexpected model output during user evaluations.

4.3.1 Preventing Harmful Model Output Prior to User Involvement. Due to the inherent stochasticity of generative language models, UX practitioners frequently witnessed model outputs that they considered to be toxic[8] when testing ideas through direct model interaction. Many UX practitioners, while hoping to get users involved as early as possible in RAI research to surface and identify potential harms, were also concerned about exposing users to these outputs. U8 told us that he was primarily *"concerned [with] what can I do before this gets to someone who's not on the team..."*

To address the uncertainty associated with potentially harmful model outputs, before inviting users in to interact with these models, teams generated a wide range of constraint-based input and output suppression techniques that often relied on classifiers and filters. U15 described approaches to handling input and output that was "inequitable" in terms of offending particular groups: *"One [technique] is [to] detect if the user[s] themselves are initiating something that is inequitable, if the model is giving out something that is inequitable, and [another] thing that we did was reduce the scope of what the model can do in terms of its output. The [other] is reduce the scope of what a person can do in terms of its input into the model."* Sometimes, UX practitioners curated resources used to classify and mitigate toxicity: *"We put in place a bunch of filters to make sure that it's on topic, that it's not offensive, that it doesn't use a set of words or phrases that we've specifically banned."* (U1)

However, it was not always possible to prevent harmful model output entirely, in which case practitioners had to come up with other user study safeguards. U14 described these challenges in the context of their experiences in an AI application design sprint, during which they met milestones related to ideating, designing, prototyping, and conducting user testing with the prototypes all within a matter of days:

> *"I think we had two and a half days until we were actually testing in front of [external, prospective] users, so it was so fast and rapid, which is one of the pros about it [...] but if something did go wrong, and if [the model] did come up with bad suggestions, or if we did have an answer that went off the rails, we couldn't just shut it down ... [all we could do] was just make the screen go blank, and hopefully they didn't see it in that second."*

Although developing guardrails and constraints is a common strategy to mitigate and prevent RAI issues with large-scale models (albeit one that may reproduce existing structural inequities in AI [37, 89]), some practitioners pointed out the importance of exploring other approaches, so as to not over-limit the types of inputs and topics that users can discuss. As R1 told us, *"I think we need to figure out how to teach the model in a controlled-generation way, to respond more appropriately so you're not always taking the sledgehammer and just suppressing results."*

However, as R8 told us, UX practitioners and product teams also struggle with developing more flexible mitigation approaches for complex algorithmic assemblages, for which they believed a constraint-based mitigation strategy was the best feasible option for mitigating potentially toxic output: *"[for some applications] it's actually many models at once. [...] For [product] purposes, a blocklist at the code level is really the most feasible thing because [a product application team] can't go back and retrain the model."*

4.3.2 User Evaluation of the AI Application. Through our interviews, we found that UX practitioners ascribed specific purposes and meaning to user evaluation in order to meet RAI challenges in early-stage AI application design. They saw user testing as an avenue for surfacing and identifying levels of comfort with large-scale model interactions and identifying possible RAI concerns through user evaluation of AI application prototypes. Here, the line between user feedback about a potential AI application, and adversarial testing[9] of the model underlying it, has the potential to blur. As U7 told us, *"We have to change our perspective on how we user test these things [AI prototypes] and think about the greater good because if we just focus on 'oh the button needs to change' and 'people didn't like the font size,' we're in trouble."*

[8]We note that there is no single, agreed-upon definition for 'toxic' language model output, though proposed definitions often include "profanities, identity attacks, slights, insults, threats, sexually explicit content, demeaning language, language that incites violence" [104] or language that targets specific people or groups with hostile, malicious or marginalizing intent [cf. 63, 104]. Practitioners in our study also referred to toxicity in these ways. Debates in the RAI community more broadly problematize definitions of toxicity that position it as a property of the language artifacts themselves, detached from the social event or contexts surrounding specific uses of language.

[9]Adversarial testing here refers to intentionally creating inputs that could lead to harmful model output, to better understand triggers for such output and the potential forms such output might take [e.g., 46, 81].

To surface and mitigate RAI issues during user evaluations, UX practitioners often sought to recruit a broad range of users to improve diversity in user testing. This often includes, but is not limited to, recruiting for users across different demographics, race, gender, tech literacy, age, etc. However, UX practitioners also needed to balance resource and time constraints while trying to diversify user testing efforts. This often leads to the need to prioritize certain RAI issues for user testing. For example, U15 described how he gave suggestions on prioritizing the RAI issues that the team knew the least about, given limited time and resources: *"There are 21 things that you are doing. My suggestion would be to look at these five initially and redesign because my understanding of the model is that these five things have not been tested previously. [...] Let's focus on things that we know the least about."*

Many UX practitioners talked about the importance of preparing users for potentially toxic or otherwise harmful outputs that the AI application might generate during user evaluation sessions. Many also described their anxiety around viewing unpredictable model outputs with the users during these sessions. To mitigate this, UX practitioners discussed the importance of setting expectations and communicating with users at the beginning of user testing sessions: *"You are showing them to an end user at the same time you're seeing them yourself. And so there's a certain amount of anxiety there. So, in the research protocol, it's really important to be able to set people's expectations: 'This is early technology, if you see some things that are harmful, I want you to be able to talk to me about it. I'll have some narrative to help you understand what you're seeing.'"*(U3)

UX practitioners also pointed out the difficulties users could face in providing honest feedback and surfacing RAI issues that were meaningful to them during user study sessions. User studies, especially those in which the researcher and participant have not built a relationship, have always grappled with the potential for social desirability bias, participant conformity to researchers' expectations, and preferences to avoid taboo topics or embarrassing (or personally invasive) social interactions. These concerns are made more salient when evaluating high-fidelity AI applications, given that large models could generate unpredictable and harmful outputs. To help mitigate this challenge, UX practitioners also sought to create a safe and encouraging environment for users to feel comfortable discussing RAI issues. U13 described one strategy they used to accomplish this by matching identity characteristics like practitioners' race and ethnicity with those of users, to help them feel comfortable talking about potential RAI concerns:

> *"I think one of the things [for] responsible AI consideration is when you're doing user research, does the makeup of your team and the people who are interviewing those folks, are users comfortable talking to them about fairness issues? We always try to match the race of the moderator to the race or the ethnicity of the participant for psychological safety. Because a lot of times, users don't really feel comfortable talking about inequity with a person who may not have experiences with systemic inequity. How are you making sure that you're conducting not just research with users responsibly and ethically, but also you are pairing them with interviewers that could lead to honest feedback."*

Next, we reflect on our findings and the ways in which, taken together, they suggest an evolution of UX praxis. We conclude by reflecting on opportunities for HCI research to move our field closer to designing responsible AI.

5 DISCUSSION

In our findings, we highlight three types of emerging RAI practices carried out by UX practitioners to meet RAI challenges: building and reinforcing an RAI lens, responsible prototyping, and responsible evaluation of AI applications. Specifically, we identified strategies that UX practitioners employed to self-educate and communicate with others about potential RAI issues; challenges during prototyping, such as communicating model capabilities (and limitations) to users; and current approaches to responsible evaluation of AI applications, including preparing both the models and the users for potentially harmful model outputs.

In this section, we first discuss implications of the hidden work of RAI conducted by UX practitioners, highlighting the labor associated with these practices. We then discuss research opportunities and ways to support UX practitioners in their RAI practice, particularly for applying large-scale language models to develop applications. Finally, we identify opportunities to reconfigure the role of the user in designing RAI.

5.1 Supporting the Hidden Work of RAI in UX Practice

In light of increasing calls for UX practitioners to be involved in the work of identifying and addressing issues of RAI in technology design [e.g., 13–15, 53, 69, 70, 84], we highlight the hidden work that UX practitioners are conducting to meet RAI challenges. We call out this hidden work by situating UX practitioners' emerging RAI practices in the context of the work conducted by RAI experts in their formal RAI roles. We find that UX practitioners and RAI experts share similar conceptualizations and aspirations for the human impacts of AI—they both saw RAI as, in large part, anticipating, surfacing, and mitigating a variety of potential harms for users and other stakeholders who may be impacted by a given AI system. Through our study, we see RAI experts carrying out the work of developing the RAI agenda, guidelines, toolkits, evaluation, and foundational research at an organizational level throughout the entire AI design and development process, while UX practitioners operate in specific application areas, contexts, and domains to actively promote, implement, and adapt RAI to fit their day-to-day work in early-stage AI application design.

Drawing upon their unique human-centered values and design techniques, UX practitioners are adapting their UX practices, negotiating their respective roles in RAI work, and learning and educating others about human-centered values and RAI concerns. However, carrying out these RAI efforts often requires UX practitioners to devote additional time and efforts in their day-to-day work, or what Star and Strauss [1999] have described as "invisible work," or akin to what Strauss [1988] has referred to as "articulation work," or the work required to make other work happen.

We find that UX practitioners take on a translation role [cf. 23] to adapt high-level RAI guidelines into specific practices applicable for their teams, given that RAI issues are often domain- and technology-specific [e.g. 30]. Our findings thus encourage more HCI research efforts to understand how to design RAI guidelines and frameworks that can be easily tailored to fit into existing UX workflows and methods. A good starting point would be to understand how UX practitioners adapt and apply existing RAI guidelines and frameworks to fit their workflows and, accordingly, provide design implications on RAI tools, practices, and frameworks for UX practitioners.

Oftentimes, UX practitioners also need to take on additional work to learn more about RAI given that RAI and design ethics are often not included in many HCI and UX curricula [108], as well as educating others on their team about UX methods, values, and impacts of AI systems on people through strategic communication. Our findings suggest that these emerging roles as learners and educators for the UX aspects of RAI may need to be explicitly articulated as part of UX practice in order to be valued by UX practitioners' organizations [cf. 23]. Organizations could also provide UX practitioners with resources and guidance to help them raise RAI concerns, as prior work has proposed for AI practitioners more broadly [70, 84, 114]. UX-specific resources might thus be designed to support UX practitioners in taking on more of such an advocate or activist role [cf. 23, 112] in working towards more responsible AI design and development processes.

This labor that we outlined above not only requires additional efforts and time from UX practitioners, but also takes the form of emotional (or affective) labor, which Wong [2021] has identified as being a crucial part of designers' ethics work, despite not being valued as part of the typical technology design process. The hidden work of RAI conducted by UX practitioners is often not recognized or valued by their managers or organizations, and therefore constantly at risk of being deprioritized in the face of competing priorities and limited resources [113], or leading others to view UX practitioners as a "blocker" of the design and development process.

In our study, we also find that in order to surface potential RAI issues in application design, UX practitioners intentionally and repeatedly generate and witness harmful and toxic outputs through model prompting, as well as prepare participants in their studies to view such toxic content. Much like Gray and Suri [2019] identified for content moderation, the burden of viewing harmful content often falls to those who are least likely to be protected from or compensated appropriately for handling the toxic externalities of large-scale sociotechnical systems.

5.2 Challenges and Opportunities in Responsibly Designing with Large Language Models

In addition to implications for the work of UX practitioners in RAI more broadly, our findings suggest specific implications for the design of new tools and methods for UX work in responsible design and evaluation of AI applications powered by LLMs. Existing ethics-focused design methods [24] or frameworks such as value-sensitive design [42] largely do not focus on AI or language models, while existing AI ethics toolkits are largely designed to support "technical" work of RAI, rather than UX practices [114]. Indeed, despite its prevalence in the academic literature, none of our participants mentioned using value-sensitive design as an approach to designing RAI. In our study, we find that traditional HCI methods such as Wizard-of-Oz studies are not only insufficient to support UX practitioners in their ideation and design processes for AI applications powered by LLMs [38, 118, 119, 122], but they are also insufficient in helping UX practitioners identify and evaluate potential RAI issues through real use cases in early-stage AI application design.

Our study sheds light on the opportunities and challenges for RAI from the emerging design and prototyping practice of prompt programming with LLMs. In their interviews, while many UX practitioners talked about the many advantages that the prompt programming tool PromptMaker brought to them (e.g., test-driving model capabilities directly, reduced reliance on other experts like engineers, and data scientists [97, 118]), practitioners also pointed out that it is not yet a mature tool for RAI design.

One substantial risk of using prompt programming to uncover RAI issues during the design and prototyping of high-fidelity AI prototypes, as pointed out by the UX practitioners in the interviews, is the danger of embedding their belief systems into users' experience of the AI system via prompting. Similar concerns were also raised regarding the use of prompting or generating "behavioral tests" [14, 86] to prompt models to generate harmful outputs, which may be limited by the positionality of the practitioner or researcher creating those tests [14, 123].

While some UX practitioners in our study attempted to mitigate this issue by crowdsourcing prompts [e.g., 75, 77, 78], this mitigation strategy may be hindered by a lack of representative groups[10] of participants as well as a disconnect between the abstract model the tests are created for and the downstream application of that model within a particular sociocultural context and use case. As a result, new tools and methods are needed that can ground the often speculative work of anticipating potential harms of language models in their specific contexts of use, rather than abstracting that social context away [cf. 117].

Substantial prior work has identified the potential harms of technologies built on LLMs [e.g., 8, 15, 105], while at the same time acknowledging the challenge of evaluating and mitigating such harms [14, 100, 123] and calling for HCI and UX researchers and practitioners to contribute a human-centered perspective to identifying and addressing RAI issues in LLMs [12–15, 123]. Our findings suggest that constraint-based input and output suppression techniques such as blocklists are a commonly used strategy to prevent users from encountering toxic output *during user testing*, in addition to preventing models from generating such output after deployment as well [37, 89].

As prior work has identified [37, 89], blocklists are crude instruments that may lead to harms of erasure [33, 37] if and when they fail to account for the social context of language use [cf. 11, 14, 123]. However, although practitioners acknowledged in their interviews that such suppression techniques are not ideal, it is often the most

[10]The question of what, precisely, "representative" might mean may be complicated by the lack of a specific deployment context or use case for LLMs prior to their use in an application, as Chasalow and Levy [2021] has discussed for the concept of representativeness in machine learning more broadly.

feasible option they have, given that product teams often deal with cascading RAI issues from upstream models they may not have access to or control over. As such, more research is needed to understand how—in development paradigms where pre-trained models are fine-tuned or used in downstream applications [15, 58]—RAI issues may be propagated from the beginning of the model development to the design process of AI applications. In addition, more research is needed to understand how constraint-based approaches such as classifiers and blocklists are developed and used, and by whom, to interrogate the assumptions underlying their design.

5.3 Reconfiguring the Role of the "User" in RAI

Recognizing the limited perspectives of researchers and practitioners involved in AI design, there are increasing calls for more user-centered or participatory approaches to responsible AI [e.g., 4, 29, 35, 52, 64, 91, 102, 111]. In our study, we found that UX practitioners involve members of the public, potentially impacted user groups, and domain experts in the design and evaluation of RAI applications; however, this new practice of involving users and other impacted stakeholders in RAI work poses new challenges and research questions for the HCI community, including when and how people should be involved in RAI design and evaluation, as well as how to protect people from any potential harms of participating in those processes.

Our findings suggest opportunities to rethink how UX practitioners conceptualize and draw on people's[11] mental models of AI applications during the responsible design and evaluation of AI systems. For instance, prior work on folk theories of algorithms [e.g., 34, 39, 62, 93] suggests that the *accuracy* of folk theories of how algorithms work may be less critical for UX practitioners than what those folk theories (or mental models) of algorithms reveal about people's orientations towards algorithmic systems. In our findings, UX practitioners' concerns about people anthropomorphizing LLMs may suggest opportunities (e.g., more seamful design [21]) to reveal the capabilities and limitations of applications based on LLMs.

In addition, our findings suggest the need to reconsider how UX practitioners configure the role that users and other potentially impacted stakeholders play in identifying and mitigating RAI issues. For instance, we see UX practitioners using user testing sessions to conduct adversarial testing of potential model harms; prior literature has suggested crowdsourcing [77] or using "bias bounties"[12] [49] or "crowd audits" to identify potentially harmful model outputs [35, 91]. However, these approaches are still nascent, and UX education and praxis has not yet developed robust methods, frameworks, and practices for UX practitioners to either lead such efforts themselves or incorporate their results into UX design and evaluation. Moreover, as we find in our work, the nature of identifying potential RAI harms may lead to unintended consequences for the participants in such studies who may either have to generate offensive, toxic output themselves, or be exposed to offensive language as a result of prompts that the UX practitioners or other participants create. Future research should thus explore ways to protect participants from these toxic externalities of RAI work.

Furthermore, despite calls for broader participation, participatory design (PD), or community-based design of AI, the current modes for engaging people in responsible AI evaluation may not deliver on the empowering goals of participatory approaches [29]. For instance, relying on users (and other stakeholders) to identify potential harms may inadvertently relegate them to a more *consultative* or extractive mode of engagement, rather than empowering them to have more generative, creative input into RAI design, as suggested by traditions such as participatory design [e.g., 29, 74], aspirations-based design [65, 101], and community-collaborative design approaches [26]. Moreover, the design paradigm of training large-scale AI models and applying them in downstream AI-powered applications poses serious questions for HCI research about how we might develop modes of participation in RAI design and evaluation that empower participants to have meaningful control over the design of AI applications.

6 LIMITATIONS

While our work provides valuable insights into and implications of emerging RAI practices carried out by UX practitioners, our study has limitations. First, all study participants were recruited from one large technology company and their RAI perspectives and practices could be shaped or limited by the organization's processes and culture around RAI; hence, more research is needed to identify the relevance and applicability of our findings and implications in other industry contexts—including at smaller technology companies. That almost all participants in our study had prior experience working at different technology companies allowed us to draw from their prior work experience during the interviews. Second, our participants mostly discussed their experiences related to computer vision and language-based ML models, as participants were primarily working within these areas. Interviewing participants with expertise in other types of ML models, such as sound-based models, could identify different challenges, strategies, or tensions. Third, our study focused on one specific style of designing and deploying AI systems, i.e., a model-first trajectory [76], in which ML models were developed before practitioners designed and built AI products around the model. We acknowledge that there are other forms of AI application design and development, such as product-first, or integrated approaches, and that our findings might be more or less transferable to these processes. Future studies should replicate our work across organizations of different sizes, AI development approaches, and maturity levels [cf. 109], to expand on the practices, challenges, and needs associated with RAI in the UX practices that we identified through our study.

7 CONCLUSION

This paper reports on interviews with fifteen UX practitioners and eight RAI subject matter experts, to understand and situate the emergent RAI practices of UX practitioners in a large technology company. Through the interviews, we identify three emerging RAI practices conducted by UX practitioners in AI application design: building and reinforcing an RAI lens, responsible prototyping, and

[11] Although standard UX practice may be to refer to "users," given the wider-reaching effects of AI systems on people beyond just their users (e.g., data- or decision-subjects, or society more broadly), it is critical to consider stakeholders beyond just end-users [cf. 99].

[12] https://blog.twitter.com/engineering/en_us/topics/insights/2021/algorithmic-bias-bounty-challenge

responsible evaluation of AI applications. We distill challenges and strategies UX practitioners employed to self-educate and communicate RAI issues with the team, to surface and identify potential harms when designing and prototyping with ML models, and to involve users in RAI application design and evaluation processes. Based on our findings, we discuss and highlight the hidden work of RAI carried out by UX practitioners. We then outline research opportunities and questions for the HCI community, to increase practitioner support for managing RAI challenges, and to move towards best practices for participatory involvement of impacted stakeholders in RAI-related processes.

ACKNOWLEDGMENTS

We thank our study participants. We also thank anonymous reviewers for their valuable feedback on the paper.

REFERENCES

[1] 2019. People+AI Guidebook. https://pair.withgoogle.com/guidebook/.
[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning*. PMLR, 60–69.
[3] Yongsu Ahn and Yu-Ru Lin. 2019. Fairsight: Visual analytics for fairness in decision making. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 1086–1095.
[4] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.
[5] Stephanie Ballard, Karen M Chappell, and Kristen Kennedy. 2019. Judgment call the game: Using value sensitive design and design fiction to surface ethical concerns related to technology. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. 421–433.
[6] Niels Bantilan. 2018. Themis-ml: A fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation. *Journal of Technology in Human Services* 36, 1 (2018), 15–30.
[7] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.
[8] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Inc, 610–623.
[9] Elena Beretta, Antonio Vetrò, Bruno Lepri, and Juan Carlos De Martin. 2021. Detecting discriminatory risk through data annotation based on bayesian inferences. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 794–804.
[10] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020).
[11] Su Lin Blodgett. 2021. Sociolinguistically driven approaches for just natural language processing. (2021).
[12] Su Lin Blodgett, Hal Daumé III, Michael Madaio, Ani Nenkova, Brendan O'Connor, Hanna Wallach, and Qian Yang. 2022. Proceedings of the Second Workshop on Bridging Human–Computer Interaction and Natural Language Processing. In *Proceedings of the Second Workshop on Bridging Human–Computer Interaction and Natural Language Processing*.
[13] Su Lin Blodgett, Q Vera Liao, Alexandra Olteanu, Rada Mihalcea, Michael Muller, Morgan Klaus Scheuerman, Chenhao Tan, and Qian Yang. 2022. Responsible Language Technologies: Foreseeing and Mitigating Harms. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–3.
[14] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 1004–1015.
[15] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).
[16] 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.
[17] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qual. Res. Psychol.* 3, 2 (Jan. 2006), 77–101.
[18] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (Aug. 2019), 589–597.
[19] Virginia Braun and Victoria Clarke. 2021. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qual. Res. Psychol.* 18, 3 (July 2021), 328–352.
[20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
[21] Matthew Chalmers and Ian MacColl. 2003. Seamful and seamless design in ubiquitous computing. In *Workshop at the crossroads: The interaction of HCI and systems issues in UbiComp*, Vol. 8.
[22] Kyla Chasalow and Karen Levy. 2021. Representativeness in statistics, politics, and machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 77–89.
[23] Shruthi Sai Chivukula, Aiza Hasib, Ziqing Li, Jingle Chen, and Colin M Gray. 2021. Identity Claims that Underlie Ethical Awareness and Action. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.
[24] Shruthi Sai Chivukula, Ziqing Li, Anne C Pivonka, Jingning Chen, and Colin M Gray. 2021. Surveying the landscape of ethics-focused design methods. *arXiv preprint arXiv:2102.08909* (2021).
[25] Shruthi Sai Chivukula, Chris Rhys Watkins, Rhea Manocha, Jingle Chen, and Colin M Gray. 2020. Dimensions of UX Practice that Shape Ethical Awareness. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
[26] Ned Cooper, Tiffanie Horne, Gillian R Hayes, Courtney Heldreth, Michal Lahav, Jess Holbrook, and Lauren Wilcox. 2022. A Systematic Review and Thematic Analysis of Community-Collaborative Approaches to Computing Research. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 73, 18 pages. https://doi.org/10.1145/3491102.3517716
[27] Hai Dang, Lukas Mecke, Florian Lehmann, Sven Goller, and Daniel Buschek. 2022. How to Prompt? Opportunities and Challenges of Zero-and Few-Shot Learning for Human-AI Interaction in Creative Applications of Generative Models. *arXiv preprint arXiv:2209.01390* (2022).
[28] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics* 10 (2022), 92–110.
[29] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2021. Stakeholder Participation in AI: Beyond" Add Diverse Stakeholders and Stir". *arXiv preprint arXiv:2111.01122* (2021).
[30] 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. *arXiv preprint arXiv:2205.06922* (2022).
[31] Emily 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] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. 2020. Bringing the people back in: Contesting benchmark machine learning datasets. *arXiv preprint arXiv:2007.07399* (2020).
[33] Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff M Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. *arXiv preprint arXiv:2108.12084* (2021).
[34] Michael A DeVito, Jeffrey T Hancock, Megan French, Jeremy Birnholtz, Judd Antin, Karrie Karahalios, Stephanie Tong, and Irina Shklovski. 2018. The algorithm and the user: How can hci use lay understandings of algorithmic systems?. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–6.
[35] Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. 2022. Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *CHI Conference on Human Factors in Computing Systems*. 1–19.
[36] Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily 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

[37] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758* (2021).

[38] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX design innovation: Challenges for working with machine learning as a design material. In *Conference on Human Factors in Computing Systems - Proceedings*, Vol. 2017-May. Association for Computing Machinery, 278–288.

[39] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I" like" it, then I hide it: Folk Theories of Social Feeds. In *Proceedings of the 2016 cHI conference on human factors in computing systems*. 2371–2382.

[40] Mary Flanagan, Daniel C Howe, and Helen Nissenbaum. 2008. Embodying values in technology: Theory and practice. *Information technology and moral philosophy* 322 (2008), 24.

[41] Luciano Floridi and Andrew Strait. 2020. Ethical Foresight Analysis: What it is and Why it is Needed? *Minds Mach.* 30, 1 (March 2020), 77–97.

[42] Batya Friedman. 1996. Value-sensitive design. *interactions* 3, 6 (1996), 16–23.

[43] Batya Friedman and David Hendry. 2012. The envisioning cards: a toolkit for catalyzing humanistic and technical imaginations. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1145–1148.

[44] Batya Friedman, Peter Kahn, and Alan Borning. 2002. Value sensitive design: Theory and methods. *University of Washington technical report* 2 (2002), 12.

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

[46] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. [n. d.]. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. ([n. d.]).

[47] 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.

[48] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462* (2020).

[49] Ira Globus-Harris, Michael Kearns, and Aaron Roth. 2022. An Algorithmic Framework for Bias Bounties. (2022).

[50] Colin M Gray and Shruthi Sai Chivukula. 2019. Ethical mediation in UX practice. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[51] Mary L Gray and Siddharth Suri. 2019. *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books.

[52] Anna Lauren Hoffmann. 2020. Terms of Inclusion: Data, Discourse, Violence. *New Media & Society* (2020).

[53] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miroslav Dudík, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery.

[54] Matthew K Hong, Adam Fourney, Derek DeBellis, and Saleema Amershi. 2021. Planning for natural language failures with the ai playbook. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–11.

[55] Lara Houston, Steven J Jackson, Daniela K Rosner, Syed Ishtiaque Ahmed, Meg Young, and Laewoo Kang. 2016. Values in repair. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 1403–1414.

[56] Ben Hutchinson, Negar Rostamzadeh, Christina Greer, Katherine Heller, and Vinodkumar Prabhakaran. 2022. Evaluation Gaps in Machine Learning Practice. *arXiv preprint arXiv:2205.05256* (2022).

[57] Nassim JafariNaimi, Lisa Nathan, and Ian Hargraves. 2015. Values as hypotheses: design, inquiry, and the service of values. *Design issues* 31, 4 (2015), 91–104.

[58] Seyyed Ahmad Javadi, Chris Norval, Richard Cloete, and Jatinder Singh. 2021. Monitoring AI Services for Misuse. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 597–607.

[59] Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328* (2017).

[60] Ellen Jiang, Kristen Olson, Edwin Toh, Alejandra Molina, Aaron Donsbach, Michael Terry, and Carrie J Cai. 2022. PromptMaker: Prompt-based Prototyping with Large Language Models. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. ACM, New York, NY, USA, 1–8.

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

[62] Nadia Karizat, Dan Delmonaco, Motahhare Eslami, and Nazanin Andalibi. 2021. Algorithmic folk theories and identity: How TikTok users co-produce Knowledge of identity and engage in algorithmic resistance. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–44.

[63] Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C. Fraser. 2021. Confronting Abusive Language Online: A Survey from the Ethical and Human Rights Perspective. *J. Artif. Int. Res.* 71 (sep 2021), 431–478. https://doi.org/10.1613/jair.1.12590

[64] Bogdan Kulynych, David Madras, Smitha Milli, Inioluwa Deborah Raji, Angela Zhou, and Richard Zemel. 2020. Participatory Approaches to Machine Learning. International Conference on Machine Learning Workshop.

[65] Neha Kumar, Marisol Wong-Villacres, Naveena Karusala, Aditya Vishwanath, Arkadeep Kumar, and Azra Ismail. 2019. Aspirations-based design. In *Proceedings of the tenth international conference on information and communication technologies and development*. 1–11.

[66] Kari Kuutti and Liam J Bannon. 2014. The turn to practice in HCI: towards a research agenda. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 3543–3552.

[67] Min Kyung Lee, Nina Grgić-Hlača, Michael Carl Tschantz, Reuben Binns, Adrian Weller, Michelle Carney, and Kori Inkpen. 2020. Human-centered approaches to fair and responsible AI. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.

[68] Michelle Seng Ah Lee and Jatinder Singh. 2021. The landscape and gaps in open source fairness toolkits. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery.

[69] 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.

[70] 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 *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery.

[71] 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.

[72] Jessica Morley, Luciano Floridi, Libby Kinsey, and Anat Elhalal. 2020. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and engineering ethics* 26, 4 (2020), 2141–2168.

[73] Jessica Morley, Libby Kinsey, Anat Elhalal, Francesca Garcia, Marta Ziosi, and Luciano Floridi. 2021. Operationalising AI ethics: barriers, enablers and next steps. *AI and Society* (2021).

[74] Michael J Muller and Sarah Kuhn. 1993. Participatory design. *Commun. ACM* 36, 6 (1993), 24–28.

[75] Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456* (2020).

[76] Nadia Nahar, Shurui Zhou, Grace Lewis, and Christian Kästner. 2022. Collaboration Challenges in Building ML-Enabled Systems: Communication, Documentation, Engineering, and Process. *Organization* 1, 2 (2022), 3.

[77] Nikita Nangia, Saku Sugawara, Harsh Trivedi, Alex Warstadt, Clara Vania, and Samuel R Bowman. 2021. What ingredients make for an effective crowdsourcing protocol for difficult NLU data collection tasks? *arXiv preprint arXiv:2106.00794* (2021).

[78] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133* (2020).

[79] Helen Nissenbaum. 2001. How computer systems embody values. *Computer* 34, 3 (2001), 120–119.

[80] Samir Passi and Mihaela Vorvoreanu. [n. d.]. Overreliance on AI: Literature review. ([n. d.]).

[81] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286* (2022).

[82] David Piorkowski, Soya Park, April Yi Wang, Dakuo Wang, Michael Muller, and Felix Portnoy. 2021. How ai developers overcome communication challenges in a multidisciplinary team: A case study. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–25.

[83] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1776–1826. https://doi.org/10.1145/3531146.3533231

[84] 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.

[85] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*. PMLR, 8821–8831.

[86] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. *arXiv preprint arXiv:2005.04118* (2020).

[87] Emma Rose and Josh Tenenberg. 2016. Arguing about design: A taxonomy of rhetorical strategies deployed by user experience practitioners. In *Proceedings of the 34th ACM International Conference on the Design of Communication*. 1–10.

[88] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 2019. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–33.

[89] Ari Schlesinger, Kenton P O'Hara, and Alex S Taylor. 2018. Let's talk about race: Identity, chatbots, and AI. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–14.

[90] 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.

[91] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–29.

[92] Katie Shilton. 2013. Values levers: Building ethics into design. *Science, Technology, & Human Values* 38, 3 (2013), 374–397.

[93] Ignacio Siles, Andrés Segura-Castillo, Ricardo Solís, and Mónica Sancho. 2020. Folk theories of algorithmic recommendations on Spotify: Enacting data assemblages in the global South. *Big Data & Society* 7, 1 (2020), 2053951720923377.

[94] Susan Leigh Star and Anselm Strauss. 1999. Layers of Silence, Arenas of Voice: The Ecology of Visible and Invisible Work. , 9–30 pages.

[95] Anselm Strauss. 1988. The articulation of project work: An organizational process. *Sociological Quarterly* 29, 2 (1988), 163–178.

[96] Norman Makoto Su, Amanda Lazar, and Lilly Irani. 2021. Critical Affects: tech work emotions amidst the techlash. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–27.

[97] Hariharan Subramonyam, Jane Im, Colleen Seifert, and Eytan Adar. 2022. Human-AI Guidelines in Practice: Leaky Abstractions as an Enabler in Collaborative Software Teams. *arXiv preprint arXiv:2207.01749* (2022).

[98] Hariharan Subramonyam, Colleen Seifert, and Eytan Adar. 2021. ProtoAI: Model-Informed Prototyping for AI-Powered Interfaces. In *International Conference on Intelligent User Interfaces, Proceedings IUI*. Association for Computing Machinery, 48–58.

[99] Lucy Suchman. 2002. Located accountabilities in technology production. *Scandinavian Journal of Information Systems* 14, 2 (2002), 7.

[100] Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, et al. 2022. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Challenges*.

[101] Kentaro Toyama. 2017. Design, needs, and aspirations in international development. In *International Conference on Social Implications of Computers in Developing Countries*. Springer, 24–32.

[102] Kush Varshney, Tina Park, Inioluwa Deborah Raji, Gaurush Hiranandani, Narasimhan Harikrishna, Oluwasanmi Koyejo, Brianna Richardson, and Min Kyung Lee. 2021. Participatory Specification of Trustworthy Machine Learning. https://www.abstractsonline.com/pp8/#!/10390/session/446

[103] Peter-Paul Verbeek. 2011. *Moralizing technology: Understanding and designing the morality of things*. University of Chicago press.

[104] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359* (2021).

[105] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 214–229.

[106] Sarah Myers West, Meredith Whittaker, and Kate Crawford. 2019. Discriminating systems. *AI Now* (2019).

[107] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2019. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 56–65.

[108] Lauren Wilcox, Betsy DiSalvo, Dick Henneman, and Qiaosi Wang. 2019. Design in the HCI classroom: Setting a research agenda. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. 871–883.

[109] Amy A. Winecoff and Elizabeth Anne Watkins. 2022. Artificial Concepts of Artificial Intelligence. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM. https://doi.org/10.1145/3514094.3534138

[110] Langdon Winner. 2017. Do artifacts have politics? In *Computer Ethics*. Routledge, 177–192.

[111] Christine T. Wolf, Haiyi Zhu, Julia Bullard, Min Kyung Lee, and Jed R. Brubaker. 2018. The Changing Contours of "Participation" in Data-Driven, Algorithmic Ecosystems: Challenges, Tactics, and an Agenda. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Jersey City, NJ, USA) *(CSCW '18)*. Association for Computing Machinery, New York, NY, USA, 377–384. https://doi.org/10.1145/3272973.3273005

[112] 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.

[113] Richmond Y Wong. 2021. Using Design Fiction Memos to Analyze UX Professionals' Values Work Practices: A Case Study Bridging Ethnographic and Design Futuring Methods. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.

[114] Richmond Y Wong, Michael A Madaio, and Nick Merrill. 2022. Seeing Like a Toolkit: How Toolkits Envision the Work of AI Ethics. *arXiv preprint arXiv:2202.08792* (2022).

[115] Richmond Y Wong and Tonya Nguyen. 2021. Timelines: A world-building activity for values advocacy. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[116] Steve Woolgar. 1990. Configuring the user: the case of usability trials. *The Sociological Review* 38, 1_suppl (1990), 58–99.

[117] Qian Yang, Justin Cranshaw, Saleema Amershi, Shamsi T Iqbal, and Jaime Teevan. 2019. Sketching NLP: A case study of exploring the right things to design with language intelligence. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery.

[118] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. 2018. Investigating how experienced UX designers effectively work with machine learning. In *DIS 2018 - Proceedings of the 2018 Designing Interactive Systems Conference*. Association for Computing Machinery, Inc, 585–596.

[119] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery.

[120] Nur Yildirim, Alex Kass, Teresa Tung, Connor Upton, Donnacha Costello, Robert Giusti, Sinem Lacin, Sara Lovic, James M O'Neill, Rudi O'Reilly Meehan, et al. 2022. How Experienced Designers of Enterprise Applications Engage AI as a Design Material. In *CHI Conference on Human Factors in Computing Systems*. 1–13.

[121] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story Writing With Large Language Models. In *27th International Conference on Intelligent User Interfaces*. 841–852.

[122] Sabah Zdanowska and Alex S Taylor. 2022. A study of UX practitioners roles in designing real-world, enterprise ML systems. In *CHI Conference on Human Factors in Computing Systems*. 1–15.

[123] Kaitlyn Zhou, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman, and Alexandra Olteanu. 2022. Deconstructing NLG Evaluation: Evaluation Practices, Assumptions, and Their Implications. *arXiv preprint arXiv:2205.06828* (2022).

[124] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2021. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134* (2021).