“It’s Everybody’s Role to Speak Up... But Not Everyone Will”: Understanding AI Professionals’ Perceptions of Accountability for AI Bias Mitigation

CAITLIN M. LANCASTER, KELSEA SCHULENBERG, CHRISTOPHER FLATHMANN, NATHAN J. MCNEESE, and GUO FREEMAN, Clemson University, United States

In this paper, we investigate the perceptions of AI professionals for their accountability for mitigating AI bias. Our work is motivated by calls for socially responsible AI development and governance in the face of societal harm but a lack of accountability across the entire socio-technical system. In particular, we explore a gap in the field stemming from the lack of empirical data needed to conclude how real AI professionals view bias mitigation and why individual AI professionals may be prevented from taking accountability even if they have the technical ability to do so. This gap is concerning as larger responsible AI efforts inherently rely on individuals who contribute to designing, developing, and deploying AI technologies and mitigation solutions. Through semi-structured interviews with AI professionals from diverse roles, organizations, and industries working on development projects, we identify that AI professionals are hindered from mitigating AI bias due to challenges that arise from two key areas: (1) their own technical and connotative understanding of AI bias and (2) internal and external organizational factors that inhibit these individuals. In exploring these factors, we reject previous claims that technical aptitude alone prevents accountability for AI bias. Instead, we point to interpersonal and intra-organizational issues that limit agency, empowerment, and overall participation in responsible computing efforts. Furthermore, to support practical approaches to responsible AI, we propose several high-level principled guidelines that will support the understanding, culpability, and mitigation of AI bias and its harm guided by both socio-technical systems and moral disengagement theories.

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

Additional Key Words and Phrases: Responsible AI, bias, bias mitigation, accountability

ACM Reference format:
Caitlin M. Lancaster, Kelsea Schulenberg, Christopher Flathmann, Nathan J. McNeese, and Guo Freeman. 2024. “It’s Everybody’s Role to Speak Up... But Not Everyone Will”: Understanding AI Professionals’ Perceptions of Accountability for AI Bias Mitigation. *ACM J. Responsib. Comput.* 1, 1, Article 5 (March 2024), 30 pages.
https://doi.org/10.1145/3632121

This work is supported by the Office of Naval Research under Grant No.: N000142112336.

Authors’ address: C. M. Lancaster, K. Schulenberg, C. Flathmann, N. J. McNeese, and G. Freeman, Clemson University, 821 McMillan Road, Clemson, South Carolina, USA; e-mails: cma8@clemson.edu, kelseas@clemson.edu, cflathm@clemson.edu, mcneese@clemson.edu, guof@clemson.edu.

2832-0565/2024/03-ART5
https://doi.org/10.1145/3632121

1 INTRODUCTION

In October 2022, The White House announced the creation of the *Blueprint for an AI Bill of Rights*, establishing five core protections against **artificial intelligence (AI)** bias and discrimination due to growing concerns over how AI technologies "*are increasingly used to make everyday decisions affecting people's rights, opportunities, and access,*" [66]. Indeed, a considerable amount of research in **Human-Computer Interaction (HCI)**, **Human-AI Interaction (HAI)**, and responsible AI details how AI-powered decision-making systems can perpetuate harmful human biases and damage society. Examples include criminal justice algorithms inordinately predicting higher re-offense rates for Black Americans compared to White Americans [2, 4, 20, 68]; insurance and credit risk predictions disproportionately disfavoring people of color [136]; recruitment and hiring AI systems discriminating against candidates based on their perceived gender [81, 92, 101, 104]; and beyond that demonstrate AI's evident social impact [85].

In response, governmental bodies and researchers within HCI, HAI, and responsible AI have put forth effort to identify policy-driven [28, 49, 50, 66, 71] and data/algorithmic-driven [17, 94, 129, 145] approaches to mitigating AI bias under the responsible AI umbrella. In this paper, we define responsible AI as "the chain of human and organizational control which governs responsible behavior for the AI system" [44], focusing on ethical principles and practices that center human values in AI development as part of a larger socio-technical system of responsibility [13, 33]. Responsible AI literature often tackles socially harmful AI bias mitigation at the governmental, societal, and organizational levels through governance and policy practices [28, 49, 50, 71].

Yet, such governance and policy approaches often fail to prevent AI bias for various reasons (e.g., slow policy adoption [90, 122] or vague, ineffectual, and corporatized "tech ethics" strategies [58, 67, 112]) and often neglect to consider how individuals may inadvertently contribute to these failures. These larger socio-technical systems of responsibility rely on the individuals who design, implement, and maintain AI systems (i.e., AI professionals) to interpret and enact proposed solutions to AI bias mitigation, capitalizing on responsible autonomy among the system members. For instance, using the AI Fairness 360 toolkit to identify and correct bias relies on human interventions, such as interpreting and implementing metrics, developing solutions, and maintaining the system [148]. Given this human reliance, the responsible AI field, while continuing to explore policies and data/algorithmically-driven mitigation, must make explicit efforts to understand the factors that can prevent individuals from feeling capable of implementing these solutions.

More specifically, there is a critical gap in understanding how actual AI professionals might feel prevented from feeling accountable (i.e., responsible) for mitigating AI bias *even if they have the technical ability to do so.* This lack of understanding results from two primary research gaps. First, previous theoretical work in responsible AI has focused on potential individual imperatives toward and gaps in responsibility for ethical AI outcomes [60, 121, 132, 135], but empirical, human-centered research with AI professionals must ground these insights in real socio-technical environments, building on key guidance from socio-technical systems theory (e.g., responsible autonomy). Second, the limited previous empirical work primarily focuses on technical knowledge as a hurdle for individuals [108] without considering social factors such as motivation, agency, and personal experience, despite their importance for supporting collective action in larger socio-technical systems [95, 123]. For instance, defining bias is socially and mathematically complex and often determined by an individual's context, experience, ethics, and morality, shaping how they might approach a shared problem in AI development [71, 81, 90, 94]. What is not well-understood, however, is how this defining process might act as a mediator to individual-level actions for real AI professionals, introducing cognitive dissonance and allowing certain processes, such as moral disengagement) to fill in gaps, further underscoring the critical need for empirical engagement.

Building upon calls to explore how humans developing these technologies perceive their projects' impact on society [34, 89] and the outlined gaps, our work seeks to understand AI professionals' perceptions of accountability for AI bias mitigation as both individuals and contributors to their organizations, and how this connects to their larger responsibilities in the socio-technical system in which they operate. Through semi-structured interviews (N=20) with AI professionals with a wide range of experience (i.e., entry-level through senior executive roles), diverse employers, and varying positions across the AI development lifecycle (e.g., data collection and tagging, algorithmic training, front-end design, and client-facing model deployment), we explore the following research questions:

> **RQ1:** How do AI professionals understand AI bias and bias mitigation practices?
> **RQ1a:** Given their understanding of bias and mitigation practices, what factors affect their perceptions of accountability for mitigating AI bias?
> **RQ2**: How do their organizations influence AI professionals' perceptions of accountability for mitigating AI bias?

In answering these research questions, our work contributes to responsible computing and responsible AI literature in several ways. First, we demonstrate the various intrapersonal, interpersonal, and intra-organizational dynamics that shape and inhibit AI professionals' sense of accountability for AI bias mitigation. Second, our work underscores the need for socially responsible AI professionals and researchers to explicitly consider and seek out AI professionals working in various roles across the AI developmental lifecycle to gain a more holistic picture of bias mitigation. Third, our work demonstrates how organizations often fail to support their employees to contribute to and collaborate on responsible AI bias mitigation efforts, instead prioritizing client and budgetary concerns over ethical dilemmas. Finally, in support of more practical approaches to responsible AI, our work presents several high-level principled guidelines to support greater collective understanding, culpability, and mitigation of AI bias and its harms.

2 BACKGROUND

To better understand the value in investigating AI bias conceptualization and mitigation responsibility at the individual level (**RQ1**), we first outline in 2.1 how pre-existing works have broadly tackled both AI bias conceptualization (2.1.1) and mitigation (2.1.2) to provide preliminary insights into the challenges AI professionals might be facing in these areas (2.1.3). In doing so, we underscore the critical need to go beyond challenge identification to specifically interrogate the role that an individual AI professional's understanding of these concepts might play in their perceptions of bias mitigation responsibility (**RQ1a**). Furthermore, 2.2 grounds our work in the foundational theories of **Socio-Technical Systems (STS)** and moral disengagement (2.2.1) in the context of responsible AI, with particular emphasis on how such works characterize the roles that governance & policy (2.2.2) and organizational dynamics (2.2.3) might play within these systems to create potential challenges for AI professionals (2.2.4). In doing so, we identify organizational dynamics in existing fields as a critically understudied factor affecting perceptions of responsibility (**RQ2**).

2.1 Understanding AI Bias, Bias Mitigation, & Their Challenges

Recently, the rapid adoption of AI-powered decision-making across domains has been linked to discriminatory outcomes [69, 118], such as racist judgments in criminal justice systems [2, 4, 68] and sexist practices in hiring [81, 92]. Due to these issues, researchers [17, 94, 129, 145], governmental bodies [28, 49, 50, 66, 71] and technology organizations [48, 97, 109] are seeking to tackle issues of bias in and stemming from AI systems. Unfortunately, the continued proliferation of AI bias demonstrates how often these mitigation efforts fail, partly because individuals' roles are not considered in the outcome. As such, we explore (1) how AI bias is broadly defined, (2) what strategies for mitigating AI bias exist, and (3) how each presents challenges for AI professionals to understand these efforts and their challenges better.

2.1.1 *Defining AI Bias.*

AI bias is defined in various, sometimes mutually exclusive ways throughout the literature and practice. For example, from a computational perspective, AI bias is an "expected or average value that differs from the true value that it aims to estimate" [12]. From a socio-technical perspective, however, AI bias is inherent favoritism toward an individual, group, or concept caused by intentional or unintentional discrimination from an AI-powered system [147]. Based on the need for standardization, researchers have created taxonomies to define and differentiate AI bias types [2, 3, 91], primarily focusing on data, algorithmic, and societal forms of bias [94, 105]. Srinivasan and Chander synthesized these efforts into four parts: (1) data-creation bias (bias in data collection, categorization, and deployment) [91, 94, 105], (2) problem formulation bias (bias stemming from how a problem is defined for an AI) [94], (3) algorithm/data analysis bias (bias derived from the algorithm, model, and its output) [17, 63], and (4) evaluation bias (bias from the humans evaluating and validating the models) [129].

2.1.2 *AI Bias Mitigation Strategies.*

Grounded in the preceding definitions, AI and machine learning researchers and practitioners have promoted three central mitigation strategies: data-, algorithmically-, and human-driven mitigation. Data-driven (i.e., preprocessing) approaches to AI bias mitigation aim to balance the datasets used to build and train an AI system to prevent improper skewing [105] using techniques such as changing input attributes and labels [29, 74, 144] or output labels [75]; reviewing data sets for abnormal distributions using toolkits or standard statistical practices [119], and pre-collecting bias impact statements to prescreen historical data [14]. Algorithmic bias mitigation strategies target both in-processing and post-processing stages. For in-processing, these approaches improve the more commonly used and researched supervised learning models [27] by either explicitly incorporating algorithmic identification of discriminatory behavior [105] (e.g., algorithmic prejudice removers in IBM's AI Fairness 360 [18, 148]) or balancing target labels [79] (e.g., iterative training with latent fair classes [82]). Post-processing approaches focus on improving AI model predictions after creation [105], including ensuring protected and unprotected groups are held proportional [75], adding bias-aware classifiers to existing black-box ones [1], and introducing **explainable AI (XAI)** to interpret the system decisions [114]. Finally, human-driven bias mitigation approaches (e.g., human-in-the-loop) use human evaluators to review AI development and deployment to assess their potential or actual biased outcomes [96] and collaborate with AI to correct identified biases [47]. Practitioners often chose this approach because humans tend to understand social contexts better, which can be necessary to recognize bias [42, 98, 142] and mitigate it [84].

2.1.3 *Challenges in AI Bias Conceptualization & Mitigation for AI Professionals.*

Despite abundant strategies accommodating various definitions of bias (e.g., algorithmic mitigation strategies for algorithmic bias), AI bias remains challenging to mitigate for several reasons, including systemic issues of resource constraints, lack of diversity in development and data, the black-box design of algorithms, and beyond [64, 100, 128]. Furthermore, individual evaluators' biases (e.g., evaluation, automation biases) often limit human-driven bias mitigation strategies [110, 129, 140]. For the AI professionals who design and deploy these systems, prior work highlights two crucial challenges at the individual level: (1) conceptual issues leading to individual interpretation and application of bias, and (2) lack of awareness about and knowledge of mitigation strategies.

For (1), AI bias' numerous denotative and connotative definitions often result in situations where individuals working across the AI development lifecycle must make interpretations of these terms based on their personal experiences [56, 99, 130]. For example, some researchers argue that bias is often inappropriately conflated with the term fairness [81] when it should be considered a facet of fairness instead [77, 90]. Fairness is a complex concept, with over 21 mathematical definitions and many more social definitions that are contextually, ethically, and morally situated [71, 90], which are difficult or impossible to address simultaneously [57]. This definitional ambiguity creates confusion for AI professionals that bleeds into professional practice across domains [102] by forcing them to rely on their own interpretations stemming from personal experience to do their job effectively [56, 99, 130]. When these personal experiences do not align with larger socio-technical understandings of bias and how it manifests in society [34, 105], there is an inadvertent failure to notice the presence of bias. Research seeking to understand how this definitional issue plays out tends to examine the problem from a societal, or system-wide, level [78] and compares understanding of bias across communities to find common ground [102]. While these high-level perspectives provide a foundational understanding of this problem, our work seeks to further this body of knowledge by engaging individual AI professionals on **how their definitions of bias might affect not only recognition of a problem (RQ1) but also perceptions of individual accountability for fixing the problem (RQ1a).**

For (2), limited previous work identifies that practitioners' lack of knowledge on bias mitigation strategies and implementation prevents AI professionals from acting on AI bias [64]. Indeed, one of the few prior works focused on how individuals attribute responsibility for AI's ethical outcomes within a socio-technical system found that practitioners cite a lack of awareness and technical knowledge as barriers to addressing these problems [108]. Sparse corporate reports echo these challenges, finding that barriers to individual bias mitigation efforts include too much focus on mitigating technical bias, obliviousness about the effect of team design decisions, limited team agency, and lack of accountability [128]. Despite the foundational understandings provided by these sources, little is known about the challenges to individual accountability for bias mitigation when AI professionals do not perceive AI bias knowledge and mitigation strategies as a primary concern. Thus, research must explore the mitigation hurdles beyond technical prowess and what creates a disconnect between knowledge and action. More critically, through our empirical investigations of how AI professionals conceptualize bias **(RQ1)** and how said conceptualizations affect perceptions of accountability **(RQ1a)**, our work fills this gap by being one of the first empirical works to our knowledge to **uncover various factors that affect perceptions of accountability *even when AI professionals have the knowledge needed to enact bias mitigation.***

2.2 Socio-technical Systems for Responsible AI & System Challenges

Recognizing the need for value-centered ethical principles when deploying AI technologies [33, 55], the responsible AI field broadly focuses on investigating and designing for AI systems' **fairness, accountability, transparency, and explainability (FATE)** [54, 139]. Recent calls by entities such as the **National Institute of Standards and Technology (NIST)** concentrate on one facet of the problem responsible AI aims to address, bias [46], and emphasize a socio-technical approach to mitigating AI bias by addressing the values and behaviors within, the humans working on, and organizational dynamics in the "commission, design, development, and ultimate deployment" of AI systems [127]. In doing so, researchers have begun to interrogate the socio-technical systems reliant on ethical prerogatives for the outcomes of AI [76], mainly at the governance, policy, [28, 49, 50, 66, 71] and organizational [48, 97, 109] levels.

2.2.1 Theories Foundational to Understanding Responsible AI Challenges. The previous works are primarily grounded in the conceptual roots of **Socio-Technical Systems (STS)** theory, which explores the intricate relationship between humans, their needs, and the technologies they create to fulfill those needs [15, 117, 133], as well as the systems that facilitate this interaction, i.e., the *hierarchy of action systems* [117]. More specifically, the *hierarchy of action systems* considers how individuals, organizations, and society interact and influence the outcomes of the STS, framing individuals as equal contributors to the system despite STS research's focus on small group behaviors [117]. Furthermore, a key principle within STS is *responsible autonomy*, which calls for internal supervision and accountability among group members to hold each other responsible for their role within the system [137]. Various factors, such as trust between group members, can impact the efficacy of responsible autonomy efforts in any STS [43]. In the context of AI development, scholars like Green [58] emphasize the need for a socio-technical perspective to address ethical considerations, which requires reframing responsible autonomy for human stakeholders and autonomous technology toward multi-level accountability [45].

However, individuals within AI development may face obstacles in fulfilling their accountability, some of which we have already outlined in 2.1 (e.g., AI bias definitions). Compared to informational inadequacy, researchers do not fully understand how factors related to an individual's social cognitive processes impact responsible autonomy within the STS of AI development. Similarly to STS theory's on *responsible autonomy* [117, 137], **social cognitive theory (SCT)** assert that individuals possess agency in their choices (i.e., an *agentic* perspective) while also being influenced by the collective efficacy of those operating within the same system [9]. More specifically, SCT processes such as Bandura's *moral disengagement* can act to reduce one's perceptions of their own agency. Moral disengagement refers to the cognitive mechanisms that work to separate an individual's moral values (e.g., the value to reduce bias in AI) from their actions (e.g., implementing mitigation strategies) [8] to avoid experiencing cognitive dissonance, or mental discomfort stemming from a mismatch between one's beliefs and actions [51]. Such mechanisms include moral justification (i.e., the reconstructing of a moral belief to make one's negative intents appear permissible or obligatory) and the displacement or diffusion of responsibility (i.e., the belief that others will or should take action rather than oneself) [6, 8, 10]. These mechanisms allow individuals to cognitively distance themselves from their own agency to address problems, thereby hindering ownership and accountability [124]. In the context of AI development, then, integrating moral disengagement with STS theory offers a nuanced approach to understanding responsible AI and ethical development [9], as the lens of moral disengagement can help us understand why individual AI professionals might be inhibited in their ability to mitigate bias beyond technical issues.

2.2.2 Governance & Policy. Responsible AI literature primarily explores governance efforts from a societal and governmental regulation perspective [28, 49, 50, 71]. In conjunction with the bias mitigation strategies outlined in 2.1.2, policy-driven mitigation approaches are increasingly seeking to support bias mitigation efforts and accountability for AI-related harms, including those from the European Commission [38], Singapore [107], and, most recently, the United States [66]. These policies introduce legal frameworks [49, 66], legislative debiasing measures, such as Art. 9(2)(g) of the European Union's General Data Protection Regulation [105], and regulatory bodies for enforcement [50]. While responsible AI is often situated at this high level, research suggests that AI governance may fail to address the societal perception of fairness from these systems [53, 90]. Additionally, given the complex designs and applications of these systems, regulatory approaches may develop too late or slowly [122], become increasingly fragmented and ineffectual [36], and be co-opted to create surveillance and privacy issues [116]. Similarly, while international governing bodies have considerable authority to set norms related to responsible AI, questions remain about their ability to overlap with other global AI and general technology governance approaches, fracturing recommendations and regulations [125]. Given these concerns, more research is needed at the granular levels supporting these larger efforts (i.e., individual AI professionals) to understand the gaps that make widespread responsible AI governance difficult.

2.2.3 Organizational Dynamics. Responsible AI governance has pointed out the advantages of organizational governance practices, highlighting the strategic advantage [109] and risk mitigation potential for the brand and overall cost of amending biased AI [48]. Several private organizations, including Accenture, PwC, and Google, publicly endorse responsible AI tools and governance, citing this responsibility as a core value [97]. These organizations, however, face competitive forces that often lead to underfunding these responsible AI practices despite their value [5]. As such, researchers postulate that increasing bias education across roles and levels in an organization can help to alleviate these pressures and support mitigation practices [30]. Furthermore, the multidisciplinary and collaborative nature of AI development teams [113] makes this cross-role and multi-level approach more imperative [23]. Indeed, Rakova et al. suggest that organizational support for responsible AI practices might influence the effectiveness of responsible AI approaches [115]. However, this work focuses on those who have received explicit training in responsible AI to understand how these concepts can be applied and the organizational implications stemming from these individuals to understand how to exercise academic ideas in practice. Of the few empirical works that have sought to understand the experiences of actual AI professionals within their natural environments, Holstein et al. found AI professionals cited a lack of organizational support and auditing practices as primary challenges to mitigation [64].

2.2.4 Challenges of Governance & Organization for Individual AI Professionals. Finally, despite individuals' importance to responsible AI infrastructure [11, 44, 70, 97, 121, 135], there is limited empirical responsible AI work centering on the individual and the individual's contributions to broader efforts. Research on the individuals' roles is often based on theoretical frameworks for broad responsibility gaps and assertions of ethical imperatives toward governance [121] rather than empirical examinations of actual AI professionals' perceptions of their responsibilities and ethical imperatives. Even the existing research narrowly focuses on technical knowledge rather than exploring other factors affecting AI professionals at the individual level [108]. This gap is concerning, as there is limited understanding of other factors that may act as a barrier to individual ownership of wider responsible AI aims [37], prompting calls for more empirical research on moral responsibility and individual culpability [86].

Researchers theorize one potential hurdle is an individual's place within an organization: in the absence of the establishment of task accountability respective to functional roles [24, 31], patterns of complacency may form among AI project teams that can prevent the adoption of ownership for bias mitigation [83]. The highly collaborative nature of AI development may also create the problem of "many hands," where overlapping responsibilities between roles obscure individual culpability [39, 131]. Additionally, critical evaluations of the broader ethical AI field have cited approaches that organizations have used to bypass regulation, scrutiny, and overall responsibility that place pressure on individual workers, such as establishing ineffectual ethics boards, adopting "rubber stamp" evaluations, and funding self-serving research initiatives [25, 58, 67, 112]. These explorations demonstrate both the notable pressures exerted on individuals from organizations and tech workers' capabilities to exert power and influence when they take an activist stance for responsible AI governance [16, 22, 58, 67, 103]. While these well-reasoned and foundational ideas explore organizational structures that influence accountability, they fail to explore these issues from the perspectives of these AI professionals in practice. As such, our work is one of the first to our knowledge to explore the critical question of **how organizations might influence AI professionals' perceptions of accountability for mitigating AI bias based on professionals' actual experiences in the field (RQ2)**. In doing so, our work seeks to provide the empirical foundation required to understand how best to curate a culture of ethical responsibility in which every employee is encouraged and empowered to participate actively in the collaborative process of bias mitigation.

3 METHODS

3.1 Context and Participants

In this study, we interviewed participants (N=20) who work on AI development project teams. In our recruitment approach, we purposefully sought to interview participants with diverse perceptions and opinions on AI development. As such, we recruited from different roles, experience levels, and organization types to understand how these actors perceive bias and their culpability for bias throughout the lifecycle of AI development. We recruited these individuals through snowball procedures from government or tech industry organizations. The participants have experience developing AI systems, often involving some human resource or capital decisions. By exploring perspectives across these various elements, we sought to find commonalities about perceptions of AI accountability in bias mitigation that can be applied across organization type, job type, or experience.

Our participant demographics are summarized in Table 1. Most of our participants were white (N=16, 80%) and male (N=18, 90%), a recruitment result that reflects the demographics of the AI industry in which only 22% of AI development workers are women, just 2.4% are Black or African American, and 3.2% are Hispanic [126, 146]. Participants ranged from 21 to 69 years old, with an average age of 41.2. All participants reside in the United States except for one participant, but all participants work for U.S.-based organizations. As such, their perceptions will be grounded in U.S. experiences, limiting the global applicability of these claims. In terms of organization type, participants are either government contractors (N=9), government employees (N=1), private industry/corporation workers (N=8), private industry/consulting workers (N=2), or former AI development team members working in academia (N=2). A few individuals (N=2) worked multiple jobs related to AI and, thus, were counted multiple times. Within these organizations, position titles included "Development Intern," "Project Manager," "Senior Software Developer," and "Chief Technology Officer," with a complete list shown in Table 1. As such, our sample offers a rich array of experiences across the development life cycle and the hierarchies of these organizations.

3.2 Semi-Structured Interviews

Two researchers who were cross-trained for the interviews and conducted the twenty semi-structured in-depth interviews via the Zoom video conferencing software. Before the interviews, participants were told the interviews would be about their work in AI and bias mitigation. They were provided with informed consent documentation via email, and all participants agreed to participate in the interviews. We did not collect participant names; any identifiable information was removed from the transcripts and deleted. All interviews were video and audio recorded with participant consent. Participants were not compensated and were involved voluntarily.

After commencing interviews, we asked participants basic demographic questions and questions about their experiences, roles, and responsibilities within their organization. Then, following these contextualization questions, we aimed to understand their perceptions of bias, asking them to situate their understanding in their own life experiences, perspectives, and language. Questions in this section of the interview focused on their experiences with bias (e.g., "*Please describe for me where you see bias happening in the AI system you are currently helping to develop.*"). Other questions explicitly focused on certain tasks and responsibilities they outlined at the start of the interview related to their jobs and organizations. We then asked them about the bias mitigation strategies

Table 1. Participant Demographics

PID	Age	Race/Ethnicity	Gender	Org Type	Job title
1	48	White	Male	Govt (contract)	Front-End Developer
2	25	White	Female	Govt (contract)	Project Manager
3	56	White	Male	Govt (contract)	Developer
4	28	White	Female	Govt (contract)	Project Manager
5	58	Other	Male	Govt (contract)	Programming Team Leader
6	55	Asian	Male	Govt (contract)	Senior Software Developer
7	42	White	Male	Govt (contract)	AI Engineer
8	48	White	Male	Govt (contract)	Full-Stack Developer
9	21	White	Male	Govt (contract)	Development Intern
10	45	Hispanic/Latino, White	Male	Private	Chief Product Officer
11	37	White	Male	Private	Chief Technology Officer
12	37	Hispanic/Latino	Male	Private	Data Scientist
13	25	White	Male	Private/Academia	Data Scientist/Postdoc. Researcher
14	69	White	Male	Private (consult)	Owner/Consultant
15	45	White	Male	Private	Vice President/Development Head
16	47	White	Male	Govt	Chief Enterprise Officer
17	25	White	Male	Private	Machine Learning Engineer
18	23	White	Male	Private	Software Engineer
19	59	White	Male	Private	Chief Scientist/ Solutions Architect
20	31	Asian	Male	Private/Academia	Director/Adjunct Professor

they employ or have seen employed (e.g., "*Please explain the types of data-driven bias mitigation strategies that you are aware of and/or have deployed in the past.*"), the role humans play related to AI technologies and bias (e.g., "*As AI technology becomes more integrated within the decision-making processes, how do you see the role of humans within the process evolving, if at all?*") and other ideal approaches to bias mitigation (e.g., "*Please tell me what you think the best way would be to make sure that future biases in an AI system can be identified and rectified, to your knowledge.*"). Per the semi-structured style, the participants and interviewers followed natural derivations from these interview questions based on the natural flow of conversation [21]. The interviews lasted, on average, 60 minutes, ranging from 45 to 150 minutes.

3.3 Data Analysis

Following the interviews, we transcribed them and removed certain filler words, such as "um" or "like," when they would not affect the meaning but would increase readability. Once we transcribed the interviews, three researchers worked to analyze the findings using an approach inspired by Grounded Theory [32]. Given our positioning in a highly interdisciplinary field related to HCI, we followed McDonald et al.'s [93] guidelines for reliability. As such, we opted to approach the coding process to explore areas of interest and relationships between ideas that support rich rigor rather than inter-rater reliability between coders. Starting with an open coding process, each researcher individually reviewed and coded the twenty transcripts grounded in the ideas presented by the participants [32]. Following these individual explorations, we explored our codes as a group, looking for overlap or contradiction and discussing these to narrow the process. Through the second round of axial coding, we better defined the codes and found new understanding grounded in the data [41, 120]. To complete our analysis, we utilized focused coding to extract quotes and create themes and sub-themes around the research questions [32].

To aid in this broader and novel analysis, this effort leveraged theory as a needed foundation for analysis. Throughout the coding process, STS theory guided the authors' framing process,

Table 2. Table Summarizing the Qualitative Findings of this Work

Findings
Theme 1: Individual Justifications for Moral Disengagement from Bias Mitigation Responsibilities
(1A) Knowledge does not Equal Action
(1B) Obligations to Objectivity Tie Hands
(1C) Bias is Inevitable & Unchangeable
Theme 2: Organizational Facilitators that Enable Moral Disengagement in AI Bias Mitigation
(2A) Intra-Organizational Power Dynamics
(2B) Organizational Value Presentations
(2C) Client Considerations

particularly for investigating how the individual actors and organizational factors become interconnected in a larger network that builds on the principle of responsible autonomy, allowing the authors to unpack the factors and constraints that influence their perceptions of their accountability toward bias mitigation [117, 133, 134]. Furthermore, literature rooted in responsible AI, AI ethics, and general considerations from ethical and moral social psychology research (i.e., moral disengagement theory) helped coders refine their findings, particularly in understanding why, despite clear recognition of the technical and social inter-workings of AI bias, the participants separated themselves from clear, actionable accountability. Through iterative discussions, the themes related to individual conceptualizations of and organizational influence on perceptions of accountability were drawn out to give a unique insight into these factors that influence capabilities toward responsible autonomy. The lead researcher refined these themes to ensure we grounded our findings in the participants' perspectives while also balancing the components important to STS theory. In composing our conclusions, we continually returned to the quotes to achieve thick descriptions that put participants' voices first.

4 FINDINGS

Our findings suggest individuals are situated in a larger socio-technical system of responsibility that must act together to address bias and biased-related harm. These AI professionals discussed their technical understanding of AI bias and mitigation and then connected their responsibilities to their conceptualizations of what this means for their practices in the AI field. These perceptions narrowed how they felt they could or should address AI bias, which is used to justify their moral disengagement in bias mitigation (**RQ1 & RQ1a**). Furthermore, these individuals felt that their organizations further restricted their culpability through organizational facilitators (i.e., organizational power dynamics, culture and values, and catering to external actors) that enable moral disengagement (**RQ2**). Results are summarized in Table 2.

4.1 Individual Justifications for Moral Disengagement from Bias Mitigation Responsibilities

In this section, we explore AI professionals' perspectives on bias in AI and their culpability for mitigating bias. We first explore participants' technical understanding of bias and bias mitigation related to their AI systems (Section 4.1.1). Then, we investigate the predominant perceptions of bias and how, given their technical competencies, they assess their accountability for mitigating it (Section 4.1.2 and 4.1.3. Together, these beliefs demonstrate how perceptions held by these individuals act as justification for their moral disengagement processes that restrict how AI professionals take accountability for their role in mitigating bias and addressing its harms.

4.1.1 Knowledge Does Not Equal Action. Participants demonstrated their technical understanding and capabilities surrounding AI bias and mitigation in their role and domain. Through this insight, we connect back to how this understanding influences individual perceptions of accountability for bias mitigation, albeit at a high level. For instance, P17 (25, white, male, private industry, Machine Learning Engineer) outlined how bias may present itself in his system and the key aspects he must consider for mitigation. Given his use of reinforcement learning, he asserted that bias is an important consideration in the design and evaluation of these systems:

> When you're analyzing state data, you're trying to see whether your system is fully trained on all the possible states, analyzing the input states and seeing if our model generalizes well. So if we're taking the states of a hundred sensors on a system, and we want to cut that down into half, how do we accurately do that without eliminating key parts of the system? With deep neural networks the input data and the readability of that output data is where the bias comes in.

P17 indicated that he would look for bias in the "*state data*" and ensure that the "*model generalizes well*", such as ensuring that changes to the related input data do not negatively affect the system function. Given the design of deep neural networks, he acknowledged that data is often the controllable source of bias and that individuals may be able to truly affect and mitigate it by understanding the nuance of the input and output data. As such, by possessing the technical knowledge of what bias is and how it affects the system, those working with these systems are accountable for analyzing this data and ensuring that systems operate correctly. While he understands the important role an individual plays technically to the system outcomes regarding bias, his role seemingly ends at interpreting data rather than actively intervening in biased data itself, limiting his actions.

Another participant, P20, summarizes his knowledge of AI bias as based on his extensive knowledge of the "*45 different definitions of bias from computational and linguistic terms,*" (31, Asian, male, private/academia, Director/Adjunct Professor). He asserted his expertise by discussing the many nuanced definitions of AI bias and acknowledged that bias identification and mitigation are challenging tasks given these complexities. Not only does he identify this complexity, but, throughout the interview, he highlighted mitigation solutions, including using "*explainability in any AI systems,*" expanding development team foci from "*the accuracy and optimizing of machine learning models to map in the overall context of equity and justice and bias metric*", utilizing abundant frameworks commercially available such as "*the AI Fairness 360 toolkit by IBM*", capitalizing "*on human involvement in things like labeling and annotation*", and incorporating "*human-in-the-loop processes with subject-matter experts in the product feedback lifecycle*". P20's numerous ideas for bias mitigation demonstrated a technical understanding of the topic. He avoided situating himself as the actor in these scenarios, limiting his ownership of the problem in his own work.

While participants often focused on bias as a quantifiable construct, others explored the societal context that contributes to the manifestation of bias. For instance, P10 discussed how his team used statistical interference when examining the outputs of their AI models, looking at the "*column that has to do with gender to understand if the distribution in there is biased or not... we know that data is already naturally biased to be insufficient in some areas*" (P10, 45, Hispanic/Latino and white, male, private industry, Chief Product Officer). P10 recognized some data, such as gender, may already be "*naturally biased*" based on how it was collected. He noted that developers can evaluate trends and locate areas where bias occurs, building from the technical foundations toward these broader societal aims. P10 further discussed the roles needed for this identification, sharing, *"if it's a gender column, one of our data scientists, an anthropologist who specializes in gender studies, would say, "Okay, this data is binary. But guess what: sex is not binary." And so she will look at the distribution and say, "Well, this is very skewed what's happening here?"*." P10 relied on his team to commence a technical review and locate instances in the data or algorithm contributing to the bias, giving them the power to institute mitigation practices. While he supports these mitigation efforts, he assigns culpability for mitigation work to others on his team to investigate the data and question the results rather than also identifying his own role in that identification and mitigation process.

4.1.2 Obligations to Objectivity Tie Hands. Given our participants' expertise in these technical operations, they often perceived bias as violations of the objectivity associated with the technical system and felt their role was to maintain objectivity in (1) the data and design of the AI model, and (2) presenting model outcomes. With these, we unpack how their perceptions limit their culpability as they narrow their focus to just the objective standards of these designs, disengaging from the ethical imperatives that they may otherwise strive for in their work.

First, the participants often asserted that when humans design AI systems, they must remain impartial in handling data. For example, P8 felt that the "*programs that we write, any code we write, any model we make, we would like it to be as objective as possible [...] You want to try and keep it as tightly coupled to what you're doing as possible, meaning you only want it to do what you need it to do, and don't try to preemptively solve other problems*" (P8, 48, white, male, government contractor, Full-Stack Developer). P8 felt his work must remain "*as objective as possible*" to prevent bias. While this perspective demonstrates accountability for select outcomes, those desired from the model, this approach can limit examinations of the model's impact, especially when an outcome appears objectively correct but makes an unintended association based on biases present in the dataset.

P6 (55, Asian, male, government contractor, Senior Software Developer) also felt that objectivity must be maintained in developing these AI systems, sharing, "*I think we should keep our feelings out of software. It isn't about how I feel. I want to build the best product that's going to do the best for the user, and not put any of what I feel about how it should work. We should limit that to build less bias*". For him, the key to preventing bias is to "*keep our feelings out of software*" or avoid inserting subjective judgments surrounding its operations. By maintaining an objective assessment of the system and its outcomes, P6 believed they are doing *"the best for the user*" and, through this inaction, avoiding building bias in their products. As a result, the weight of these outcomes falls on the system user to notice and ask for change. Blame is passed down the line to serve this objective purpose.

Second, while some participants focused on objective development, others focused on bias as violations of objectivity from data presentation and application. P2 discussed how bias is likely to occur when information is intentionally skewed to satisfy subjective desires rather than objective outcomes. She felt, "*there's bias in it [. . .] based on how they label something or based on how they zoomed in on the chart or based on any of it, they are biasing you towards their opinion, in a way, manipulating the statistics*" (P2, 25, white, female, government contractor, Project Manager). P2 asserted that data presentation is a primary source of bias from an AI as a "*manipulation*" of the output, where people make the system speak toward "*their opinion*" rather than the objective truth. As such, her role becomes to remain objective in her reporting procedures and avoid deviating from the models' findings. Furthermore, her use of "*they*" throughout this discussion, referring to the clients in this case, may suggest she views others as the culprits of this behavior more so than herself, removing herself from this biasing and reducing her accountability.

Similarly, another participant in a client-facing role believed he should remove personal viewpoints when presenting model outputs and *"not put our biases in it [...] I try to do it all based on quantitative data that I get, not anything that's intrinsically important to my value personal value system"* (P11, 37, white, male, private industry, Chief Technology Officer). P11 thought he must use objective, quantitative evaluations and avoid applying his own "*value system*" to evaluate the AI's decisions. He takes accountability for the technical components of the model, ensuring his interpretations match the data. At the same time, he avoids applying his own values to the analysis. This limits his potential to explore further the consequences of these technologies needed to identify more nuanced societal biases exacerbated by AI. As such, P11 demonstrated evident ownership of bias related to the technical system but failed to connect his work to broader societal implications.

4.1.3 Bias is Inevitable & Unchangeable. While the previous sub-theme discussed bias as ignoring objectivity, other participants viewed bias as an inevitable outcome of human-designed systems. Because these participants consider bias unavoidable, they often do not feel obligated to or capable of acting against these biased outcomes. As such, they do not feel they are tasked with intervening in the presence of bias from the systems they develop. One such participant, P9 (21, white, male, government contractor, Development Intern), shared that "*it's kind of human nature in itself to kind of have bias [from AI decisions] no matter how hard we try because we're not computers ourselves*". For P9, human involvement with these systems guarantees that bias will result from their outputs. As such, AI systems will always perpetuate the biases coming from inherently subjective humans. Thus, individuals are better off stepping away from these technologies rather than becoming more involved with them, allowing the "*computer*" to operate without further human interference.

Similarly, P5 (58, other, male, government contractor, Programming Team Leader) asserted that AI is biased because "*it's a human system, humans are biased, where you make decisions on those biases at all levels*". Because the system operates on human bias, the AI itself will always be biased Unlike P9, P5 specifically labels the system as biased but also views this as an extension of the human creators.Expanding this discussion, he shared that "*in the model development, certainly the key technical modelers have a bias for similar experiences and how things have worked in the past when they've modeled other human organizational systems like this could come into play. The hopeful check on that is that the validation process, where it's got to line up with the actual data.*" P5 felt that human bias is present through the creation process from the "*technical modelers*" who have a bias toward previous experiences working on "*human organizational systems".* Fixing these biases relies then on the "*validation process*", leaving the system, not the humans, culpable to "*check*" for the proliferation of bias.

In contrast, others believed that, while bias is unavoidable, it is a positive feature. For P10 (45, Hispanic/Latino and white, male, private industry, Chief Product Officer), the bias makes his systems functional. While he does not support harms derived from bias, he also acknowledged that "*bias is something that you cannot escape, period. Because the models that we build need some previous knowledge so that they're able to construct a model in themselves [...] no model can give you meaning without the bias that goes into its training*". P10 recognized that bias is an inescapable byproduct of the system, but it is necessary to interpret data and provide meaningful decisions to users. As such, biases are not a problem but a solution; his role is to capitalize on the bias to serve his system and customers.

In sum, these three strands demonstrate how AI professionals' perceptions of accountability for AI bias mitigation largely stem from their technical understanding of the system and how their perceptions of bias are either constrained by their disengagement from affecting the objectivity of the model or the belief that their involvement will have little impact due to the inevitable nature of bias from human-designed systems. As such, their perceptions of bias detach them from their obligations to accountability and limit their actions toward the larger socio-technical system of responsibility.

4.2 Organizational Facilitators that Enable Moral Disengagement in AI Bias Mitigation

While our participants' perspectives on bias limit their ability to mitigate it, additional organizational factors further inhibited their accountability. Indeed, building off the literature on moral disengagement, namely displacement and diffusion of responsibility [62, 111], the standards set by these organizations further encourage inaction and underplaying of ethical dilemmas stemming from AI bias that make AI professionals disconnected from the larger socio-technical system of responsibility. For one, several participants focused on organizational power dynamics that affect their culpability (Section 4.2.1). Participants also outlined how company culture and values, especially in everyday practices, contribute to their feelings of accountability (Section 4.2.2). Finally, they assessed how those factors outside their organization, namely clients, compound these issues and pigeonhole their ability to address bias (Section 4.2.3).

4.2.1 Intra-Organizational Power Dynamics. Because participants worked in roles across the technology development lifecycle, we gained insight into how their perceived positionality and power in their organization affect their accountability to mitigate bias. Participants expressed a generalized belief that they hold an ethical imperative for mitigating bias, such as one participant who shared, "*I think it's everybody's role to speak up anytime you see something you don't like [...] but, in my experience, not everyone will*" (P7, 42, white, male, govt, AI Engineer). This thread of collective culpability was evident throughout these interviews, as all viewed humans as generally accountable for the tools they create. However, in practice, he also realized that "*not everyone will*" follow through, especially as they contend with their own perceived power and the factors that limit it. Namely, participants highlight experiences in their career and role, and direct knowledge of the system dictates their accountability, often leading to a displacement or diffusion of responsibility toward others perceived to be more powerful or in control of the mitigation process.

The youngest and lowest-ranking employees in the interview pool believed that those in charge are accountable, removing themselves from acting. P9 (21, white, male, government contractor, Development Intern), felt that when the user discovers bias-related issues, the response depends on the following:

> If the owner of the software wanted to correct the issue and wanted to come up with a solution, to work with the user to find a way to solve the issue, because it's an issue that should be solved. When they find a conclusion, they then come to the developers and tell them, hey, this is the problem, this is how we get to solve and make it happen.

P9 placed the impetus to fix bias on the software owner, who should interface with the direct users to find a solution. He acknowledged that these issues should be solved, just not by lower-level employees like himself; only once the leadership of these companies "*find a conclusion*" will they ask developers to fix it. As such, ownership towards mitigation becomes narrowly assigned to those in high-ranking positions at companies and requires that they see "*the problem*" as truly a problem. Similarly, P18 (23, white, male, private industry, Software Engineer), another new developer, felt "*that developers, especially if you have developers at a large company, will not be doing exactly what they want to do. They will be doing what has been passed down from the highest level.*" Like P9, P18 notes that developers are at the bottom of the chain of command, restricting them from acting on what they want or believe is right. They receive directives that tell them exactly what to do, leaving little room for them to work outside those parameters. These restrictions make them believe that they are not responsible for acting and adopting their own feelings of accountability for bias. Thus, the organizational hierarchy and limitations on positional agency prevent greater accountability adoption by junior employees.

In comparison, a more seasoned employee, P6 (55, Asian, male, government contractor, Senior Software Developer), discussed his career growth and his role in preventing bias, explaining:

> You're more impressionable when you're younger. Depending on what role a person has, say your boss comes to you, and you're a new developer, and they're pushing this

> now [...] I would feel like they would be more pressured into doing it, versus where if someone came to me, it wouldn't matter if they're my boss, the CEO of the company. I just feel like I'm just not going to let it happen.

With time, "*impressionable*" new developers may gain the faculties to stand up to those in power, though this is placed down the line and displaces these individuals from their duties and responsibilities rather than supporting greater investment. At the same time, P6 felt that he had a greater capacity to stand up to those in power. Thus, he thought his culpability toward bias mitigation was greater because his position affords him greater leverage and job security than less experienced employees.

Others did not attribute age and experience as the driving factors for their bias mitigation culpability but their knowledge about the system due to their role. For instance, P17 (25, white, male, private industry, Machine Learning Engineer), shared that, "*I feel that the designers of AI systems definitely take on some of the role of mitigating bias because most of the time if you're working for a customer, they're not going to notice bias unless you point it out [...] But at the same time, it puts the burden on me*". Given his role, P17 possesses greater knowledge about the system and its potential impact than most customers. However, this responsibility to educate users places "*the burden*" on him. Thus, ethical duty becomes an undesirable aspect of his job, even if he knows he is uniquely positioned "*to notice bias*". As such, he accepted accountability as a facet of his technical knowledge of the system, but he only has "*some of the role of mitigating bias*" because the clients are ultimately the ones deploying this system.

Unlike P17, P1 (48, white, male, government contractor, Front-End Developer), took a hands-off approach to accountability because his role, designing interfaces, does not give him direct knowledge of the model development process He shared that, "*as far as my role in it all, by the time the data gets to me, what's done is done. I'm just the unbiased journalist*". P1 believed he could not change anything about the model, even if he noted bias in that data. Instead, he is an "*unbiased journalist*" who must report everything without inserting his perspective or opinions into the data. Based on his role, he perceived that he lacked the positioning to investigate data further for bias. Thus, he does not feel accountable because the hierarchy of the developmental decision-making process binds him.

4.2.2 Organizational Value Presentations. The participants also highlighted that the values exhibited by their organizations toward AI bias and ethics dictate how they feel they must operate toward the issue. While extra-organizational efforts affected their perceptions, internal actions in pursuit of their aims also influenced how culpable they felt.

Participants expressed how their company leadership discussed AI bias and ethics, which influenced their desire to look for harmful effects of their AI. If their company valued taking social responsibility for their outputs, it made them more inclined to feel accountable for AI bias. For instance, P12 (37, Hispanic/Latino, male, private industry, Data Scientist) shared how his CEO said they would not "*work for super right-wing communications divisions or anything like that. That's not where we stand as a company. We're not going to lend our services to that. So, I think that's one way to affect bias*". P12's discussion illuminates that when organizational leadership emphasizes social responsibility, the employees are inspired to embrace accountability in their own workplace actions. Thus, organizational social responsibility practices, such as taking a "*stand as a company*", can imbue all parts of the organization with greater feelings of culpability and push employees to see how they can "*affect bias.*"

Relatedly, P4 (28, white, female, government contractor work, Project Manager) maintained that the overarching accountability for these AI systems stems from the organization's policies. She shared that when developing broad AI decision-making tools:

> There has to be sort of strict [guidelines], and this is more maybe in like an ethical question than a scientific or developer question. There have to be ethical practices in place, or reviews and meticulous kinds of documentation of the decisions and review of the decisions that the tool is making to make sure that that's in line with policy or in line with the actual data that you rely on and not biased.

P4 posits that an organization developing these tools must institute "*ethical practices*", including clear policies and review mechanisms for biased outcomes, that must come from an organizational and administrative "*policy*" level rather than individual "*scientific or developer*" perspectives. However, the ones working on these systems also have a duty to support the "*review*" processes. As such, while leadership is responsible for setting these policies, they place culpability on in-the-weeds employees to actively engage with this bias mitigation review process.

Other participants felt that when social responsibility was not a company focus, employees are not incentivized to address bias. For instance, P13 (25, white, male, academia, Data Scientist/Postdoctoral Researcher) stated that there is:

> A public and private belief about it [bias]: I think the public belief about it is such an important problem. This is why we're going to hire all of these people to kind of deal with this. I think the private belief is that it tends to be a waste of time because it's relatively low impact [...] That's not the problem that we need to focus on right now.

P13 observed organizations act in service of their public image, where companies will publicly take steps to "*deal*" with bias but privately deprioritize mitigation. Because of this, if the company does not place internal value on bias mitigation, there is minimal support for employees to mitigate bias. As such, individuals and their organizations often see little reason to take accountability for bias because this it is not worth investing time and resources to accomplish.

Like P13, P18 (23, white, male, 23, private industry, Software Engineer) felt that organizations have a public and private persona around AI bias. When discussing bias-related harms and company cultural values, he asserted companies are "*going to be pushing their bias to whatever they want it to be on [...] they're definitely trying to serve their own purpose. They're trying to make money*". While P18 observed these companies stating that they have ethical imperatives to prevent harm from AI, underneath this facade, they are just driven by making money. Consequently, while organizations depict their public values as focused on reducing adverse AI bias outcomes, these companies serve their own priorities, ethical or not. Indeed, there is an inherent belief that these companies will not change. P18 added that, because of this, those on project teams must then discuss potential AI harm behind "*closed doors with the developers I think you have a moral obligation to at least lead it away as much as you can without it being so ghastly obvious to your company.*" Because P18 assumes companies will ignore obligations internally to serve their own purposes, even if bias mitigation is the more ethical approach, it often falls on the project leaders and developers to navigate the hurdles to both the company's demands and their ethical obligations, even when they are in conflict. Thus, when a company does not practice its values, developers are stuck in a liminal space that does not support nor empower their ability to mitigate AI bias.

Finally, in talking about company and industry failures to address bias, another participant believed that, "*companies and society are not as progressive as some of the people that exist within those systems. And people tend to be the driving force behind making changes*" (P11, 37, white, male, private industry, Chief Technology Officer). Companies often hide behind "*progressive*" values but fail to live up to the promises they assert. However, individuals possess unique capabilities to be the "*driving force*" behind instituting necessary changes. As such, while there are clear constraints on individuals from their organizations, individuals are still culpable for supporting ethical outcomes.

4.2.3 Client Considerations. Participants also acknowledged how commitments outside the organization inhibit their feelings of individual accountability. Namely, the participants focused on how their obligations to client needs and society often conflicted. This mediated their accountability for mitigating AI bias, leading to indecision from conflicting pressures and limited perceived agency.

For one, participants felt that client budgets dictate responsible AI practices. P11 (37, white, male, private industry, Chief Technology Officer) and P2 (25, white, female, 25, government contractor, Project Manager), who represent private and public organizations, stated:

> When it comes to budgeting something for a client, it would be quite difficult to defend a budgeting decision to add more people to help prevent bias. It's not that you can't, but it's just more difficult to make that a defensible thing (P11).
> The emphasis is always going to be on budget, and then they're going to sacrifice manpower for budget. Because that's what I've seen in the past. Everybody always talks about budget and the bottom line (P2).

P11 and P2 have intimate relationships with clients and understand how their budgets shape companies' abilities to mitigate bias. Their roles place their power to act at the whims of the client's needs and budget resources. P2, for instance, claims that clients will care little about bias regarding cost and are much more interested in the development projects' performance and "*bottom line.*" Furthermore, due to client budget sensitivity, companies will opt to cut costs, including the manpower necessary for bias mitigation efforts. Thus, emphasis on bias mitigation is limited, discouraging employees from taking extra steps toward AI bias accountability. Moreover, leadership may even stop them from taking these steps as it will incur additional costs to the client.

Given this emphasis on client wants, we asked these participants about their feelings about clients who may want to develop AI that produces harm or induces bias. Several participants expressed that they would feel uncomfortable with these requests, but organizations' emphasis on client desires affects their abilities to act. Indeed, participants often expressed hesitancy about these ethical conflicts and how they may act in these scenarios, such as P6, who shared, *"I'm sure you could say something, like maybe a company hires you and wants bias there. They want to force that kind of bias on there because maybe they want a different result. I mean, I don't know... It just would depend... It would depend on the situation"* (P6, 55, Asian, male, government contractor, Senior Software Developer). In this statement, P6 went back and forth about his requirement to act when a client wants potentially harmful bias in the model. As a team leader who works directly with clients, his role dictates that he satisfies the customer's needs for his organization. He acknowledges skewing results to match client desires is not unheard of in his field. By that logic, he would consider intentionally biasing results if the situation called for it. P6 found difficulty taking a strong stance on his responsibility, given his internal conflict. As such, the client's demands and his company's reliance on these client needs create ambiguity toward his responsibility to act in the face of unethical issues. Interestingly, he previously asserted that he would have no issue speaking up about ethical problems, given his position and power within the organization. This divergence indicates that he perceives responsibility for bias grows as one gains power in a company, but their power is capped for client desires and represents another way that responsibility toward mitigation can be displaced for AI professionals via an organization's priorities.

Furthermore, other participants took stronger stances on their detachment from accountability and responsibility, such as P4 (28, white, female, government contractor, Project Manager), who shared that when evaluating and fixing a model that they suspect may be biased, the AI development organization should take the stance that addressing this bias is:

> (1) not our jobs and (2) not our place to put in anything into the model that doesn't come directly from a client. I mean, that's sort of the nature of our jobs. I know that we will not be able to model the performance of the organization faithfully if we are too far away from what the client tells us.

P4 aims to serve the client's needs above all else, believing that it is not her "job" to step outside of the client's wishes, even if she may think differently. In her mind, her role is to represent "*what the client tells*" them in the model and avoid inserting any outside assumptions. As such, the client is the one who dictates if and how the organization and its employees address bias from AI.

Participants also emphasized that organizational leadership, in serving client needs, set the stage for project outcomes. P20 (31, Asian, male, private/academia, Director/Adjunct Professor) pointed out that, in terms of implementing these bias mitigation practices:

> Developers have that agency to define what goes within a particular segment of the code. But I think it also has to come down to the business owners, and the product owners will actually define what needs to be built. So I would say education and awareness why we are working on a particular use case and what does that mean in the context of the user, that's really important to business owners and developers.

P20 acknowledged that developers have the power and "*agency to define*" what goes into "*the code*" and are, therefore, accountable for the outcomes of their actions within the system. However, business and product owners are the ones who "*define what needs to be built*" and offer guidance for the project overall, thus shaping what approaches will be incorporated into the project and constraining the individual actions as a result. As such, P20 felt that all AI professionals need "*education and awareness*" about their role in this development process and the context of where their work is deployed. Thus, all participants on these teams must understand their role in the proliferation of and accountability for bias-related harm from the AI they create rather than just serving the client or business case.

In summary, in our interviews, AI professionals wrestled with accountability for bias in their daily practice. Conceptually, our participants understood that their actions influence AI bias, yet little research explores why individuals cannot or do not take accountability for bias mitigation. This research takes a step toward this aim, identifying barriers that can prevent individual culpability and action, including individual justifications for moral disengagement (i.e., an individual mismatch between their technical abilities and their perceptions of bias as a lack of objectivity or as a subjective, but unavoidable, problem); intra-organizational power structures that inhibit accountability adoption, particularly for those who lack authority or feel restricted by company practices; and, inter-organizational factors, namely client wishes and budgetary constraints in conflict with moral imperatives. As such, their insight illuminates how these individual perceptions lead to moral disengagement, and organizational facilitators further enable this disengagement that leads to AI professionals failing to take accountability for their role in the socio-technical systems of responsibility.

5 DISCUSSION

The following section discusses our contributions to responsible AI literature surrounding individuals' needs and organizations' roles in supporting responsible AI practices. Namely, we highlight the need for greater individual empowerment of AI professionals who partake in these development projects, even tangentially, to overcome their moral disengagement and take responsibility for reducing AI bias's harmful outcomes. In doing so, we acknowledge the role that organizational practices play in shaping these responsible AI practices and discuss how the future of AI development organizations must center these ethical conceptions as practices rather than offloading the blame. We conclude this section by offering high-level recommendations to support individual accountability, which helps collective bias mitigation efforts.

5.1 Individual Empowerment for Accountability

Our participants, even those not in direct development roles, were knowledgeable about their systems' technical workings and biases. Based on how they discussed the technical operations of their systems, participants are evidently well-versed in how bias may occur and can identify solutions to avoid it. While the literature acknowledges the philosophical and moral imperatives of humans to support ethical practices and responsibility gaps that hinder these systems [121], these responsibility gaps are discussed vaguely, making it difficult to pinpoint the problem. Furthermore, the empirical research in this area focuses only on the gaps that exist in developers' technical knowledge [108] likely due to their limited exploration into individual experiences related to responsible AI practices [97]. Even acknowledging the complex nature of AI bias [94], our participants' plentiful discussions about the manifestations of bias and mitigation indicate that AI professionals' technical knowledge is not their central hurdle. However, their moral disengagement and passing of the blame demonstrates a need to connect these general perceptions around accountability with the actions these individuals undertake, aligning their technical knowledge with direct action that is not limited to those deeply enmeshed in the responsible AI world [115] but extending to the entire field.

As such, AI professionals demonstrate that the actual problem rests in their perceptions of individual agency and the imperative to take accountability as shaped by their understanding of bias. The responsible AI literature focuses too much on a lack of technical skills as a hurdle to greater accountability adoption when they should also be examining how these professionals' understandings of bias constrain their perceived and actionable accountability, such as examining how these conceptions act as justifications for moral disengagement behaviors [6, 7]. Our conversations depict a collection of individuals across industries and organizations who acknowledge a broader need for accountability but feel morally disconnected from it and they lack the agency to persuade larger change. Looking at this from Ropohl's hierarchy of action systems assertion in STS, if certain individuals or groups fail to effectively play their role in the action (i.e., in bias mitigation) then this will contribute to the proliferation of suboptimal outcomes for the group, especially when it is not uniquely occurring in one unit of the system. This, paired with responsible autonomy, suggests that accountability failures occur across individuals and groups, further encouraging moral disengagement mechanisms.

Furthermore, these feelings of limited agency reflect much of the arguments in other social responsibility literature where people view individual actions as disconnected from and unimportant to collective efforts, such as in climate change responsibility [72, 106] and general corporate social responsibility research [88]. Responsible AI literature has failed to look at the needs of these individuals and how to leverage their knowledge and capabilities toward these more extensive efforts [97]. Indeed, we found that direct personal agency is often overlooked as a powerful force for encouraging more significant collective efficacy [9]. This has led to unfavorable conditions for supporting responsible autonomy for bias mitigation that leaves individuals paralyzed by their perceptions of limited agency, power, and responsibility for the problem. Thus, there is a breakdown in this system that fails to instill a belief in AI professionals that their actions can influence change despite these individuals occupying roles that have the most direct impact on the creation of AI technology.

5.2 Organizational Values Must Build, Not Diffuse, Responsibility

Participants also highlighted how their organization obfuscates their feelings toward AI bias mitigation and general ethical practices. These values, such as creating safe spaces to speak up, made them comfortable to take accountability for preventing harm from these technologies. However, beyond identifying these values, participants desired that their organization and work groups actively and intentionally enact these values through their daily business practices. In this sense, the participants expressed the need for responsible autonomy as these "continuous, redundant and recursive interactions" [137], akin to daily practices outlined by participants, support mutual awareness, making it easier to act toward favorable, ethical outcomes. At the same time, AI professionals expressed that when these values are not routinely demonstrated, ethical apathy from the organizational level breeds indifference to individual accountability, as there was little motivation for responsible autonomy, encouraging moral disengagement behaviors.

Indeed, participants felt that organizations expect them to bend their ethics, deprioritizing bias and restricting the individual's actions, a pervasive problem across technology development fields [40, 87, 138], but especially in AI development [58, 67, 112]. While researchers have pushed for cultural shifts toward responsibility-sharing across the development lifestyle to support performance and satisfaction [141], they often ignore this for building an ethical culture. However, popular media has drawn attention to this issue where AI companies' ethical principles are often "just talk" and not followed in practice [59, 61]. Our present research adds empirical support to these failures and indicates that the disconnect between what is valued and what is practiced in these AI development companies breeds apathy and displacement of responsibility among those employed across the AI development lifecycle. Further, our work extends explorations of responsible AI work practices [115], adding how externalities, namely client needs and wants, compound upon internal practices to constrict individual perceptions of accountability, creating a dichotomy between understanding the problem and acting against it. Thus, our work demonstrates that ethics and accountability are largely ignored in these AI development organizations and that these poison individuals' perceptions of accountability, hindering their work.

Building from this perspective, our participants' discussions about culpability demonstrate an overarching culture of passing the buck and pointing the finger, an idiom for a failure to take ownership of the problems associated with AI bias associated with moral disengagement theory. Despite their understanding of bias, AI professionals rarely emphasize themselves as accountable for the product outcomes. Blame avoidance is common in human organizations, especially when organizational culture does not emphasize accountability [65]. However, blame avoidance in AI organizations has been examined as creators blaming agents due to their autonomous capabilities [143] or from administrative blame avoidance [67, 121], rather than the effects on the individual conceptions of accountability. In contrast, our participants offloaded their culpability for AI bias based on their beliefs about their positionality and power in the organization to execute change.

5.3 Future Directions for Supporting Responsible AI

As Kling and Star [80] assert, workable technology, that is, technology that supports humanity is not designed in a vacuum but "supported by a strong socio-technical infrastructure." Thus, if the responsible AI field aims to enhance society through better AI systems [33], the human systems that develop these technologies must take accountability for their work. As such, we offer three principled guidelines for supporting accountability for AI bias. While specific policies and actions will be context-dependent, our guidelines can help navigate future directions for individuals and organizations alike.

5.3.1 Guideline 1: Bias Mitigation Training Should Directly Call Individuals to Act. Our findings indicate that AI professionals understand the technical infrastructure in which they operate, but their perceptions of what bias involves and their disengagement from accountability for bias often limit their actions toward mitigation. As such, we recommend that those educating and employing

AI professionals capitalize on this clear expertise while expanding their understanding of their actionable abilities to mitigate harm from bias. Indeed, participant responses were heavily rooted in traditional machine learning concepts or focused on objectivity, which does not always bode well with social forms of bias. Expanding their understanding of bias through exposure to the many definitions of bias [94] and their relationships to fairness [105] could support a more holistic view of the issue. Furthermore, greater emphasis on FATE and other responsible AI materials [54, 139] can help connect their abundant technical knowledge with accountability, demonstrating avenues where their direct action can reduce the impact of bias from these systems.

Furthermore, more work, particularly in applied research, needs to focus on individual agency and power as contributing to systemic solutions to AI bias. Building on the agentic perspectives of SCT from which moral disengagement is derived, there is a need to foster personal direct agency and greater collective efficacy between AI professionals by supporting responsible autonomy that can naturally scale up the hierarchies within socio-technical systems. Engaging Lowland's expansion of STS into organizational design (STSL) can offer recommendations for increasing the leanness of AI development organizations, establishing cross-functional teams that, as a result, allow greater power to support individuals' agency such as via cross-training methods [35]. However, until individuals are more uniformly supported to act on their faculties, the system of socio-technical responsibility will be significantly hindered.

5.3.2 Guideline 2: Prioritize Ethics at the Individual Actor Level. Our findings demonstrate that AI professionals feel that their organization's ethical failures and power structures silo their individual roles and limit their culpability. As such, these practices allow individuals to displace their responsibility in the AI bias mitigation process, especially when it goes beyond their position. As previous research asserts, tech companies often manipulate ethical policies to allow them to avoid responsibility, and tech workers are often complacent in these practices [67]. Our work further asserts that lived, daily practices, even more so than those broader ethical AI statements and values, are what truly matter for AI professionals to align themselves with their accountability for AI bias and their value toward systemic change. Given the often opaque aims of these tech companies and the tendency to manipulate ethical policies, these changes will often need to be both systemically enforced and supported from the ground up. When properly equipped, governmental policies and regulatory bodies may offer the necessary external pressure to force culture change and ethical accountability for AI bias, especially by offering support and guidance to the individual AI professionals that make up these organizations [73]. Regardless, AI development codes of conduct need to be implemented in ways that motivate individual engagement while also being enforced from outside the biased prerogatives of the organization, addressing both the internal and external sources that may lead to displacement and diffusion of responsibility.

At the same time, participants also highlighted a belief that their individual role did not influence the collective problem, feeling as though they lacked the power to take accountability or to make a real difference in the bias-related outcomes of this project. We recommend practices that support better role delineation while supporting individuals' awareness of their roles' connections and influence on others in the bias mitigation process, such as suggested by STSL theory's emphasis on lean organizational approaches to demystify other roles and responsibilities in the organization. By utilizing this approach, individuals can see how they fit into this larger system and how their actions impact their decisions' outcomes to take accountability. Learning lessons from other empowerment practices [72], these strategies should also indicate how an individual actor can support the overall system to reduce decision paralysis. Returning to the ethical guidelines, individuals should have a role in outlining the specific actions and responsibilities that contribute to these ethical guidelines, offering a necessary bottom-up perspective on AI bias accountability that encourages ownership [52] and overcomes helpless or disconnected feelings that were especially pervasive for less experienced AI professionals. Thus, organizations should work with employees to build and regularly refine the company's ethical practices, ensuring that individual perspectives can contribute to these collective aims.

5.3.3 Guideline 3: Prioritize Bias Awareness to All Stakeholders. Finally, our findings also indicate that external pressures on an organization, which are ostensibly out of the scope of control for individuals, nonetheless influence their perceptions of accountability. AI professionals suggested that budgets outlined by clients impact their responsibilities. In light of this, as P20 recommended, in the cases where the client needs and budgetary concerns are defining the project aims, education and awareness on the particular use case and context that acknowledges every level, from business owners to developers, can support responsible AI practices. Thus, we recommend company-wide professional development focused on bias mitigation and stakeholder research that enhances understanding of social forces that shape outcomes from their products and highlights how actors across the development lifecycle can positively affect these outcomes. Such ways to achieve this could be through constructing and sharing in collaborative organizational case studies, utilizing data statements on development projects, and other immersive approaches to understanding the potential ramifications of their work validated in the AI ethics literature [19, 26, 99]. However, we believe these approaches should go further and place particular emphasis on how their roles influence the technology development lifecycle and contribute to de-biasing outcomes. This demonstrates how even when certain decisions are outside AI professionals' scope of work, they can still influence the overall system through their decisions given the ripple effects actions can have through the hierarchies in socio-technical systems.

Additionally, building on this education, when working with clients, development companies should discuss issues of bias and prioritize these as part of their development costs. Mitigation efforts take significant human and computational resources, and organizations must argue for these costs with clients. While arguing for these increased costs may seem difficult, preemptive ethical practices can be beneficial in the long term; as research suggests, perceptions of biased AI products can hurt the brand reputation, tank revenue, and lose customers [48], making these investments worthwhile overall. Social, financial, governmental, and organizational-peer pressures may all offer incentives for organizations to prioritize mitigation in all their work, which in turn will set a standard for those AI professionals to see the impact in their daily work.

While all of these recommendations promise to support better bottom-up ethical practices related to AI bias, we also acknowledge that there must be continued support from the government and globally responsible AI communities to enforce these practices. Indeed, the responsible AI literature's focus on these large-scale approaches to responsibility and enforcement speaks to their importance [28, 49, 50, 71]. However, we believe that emerging governance efforts must enforce organizational adherence to these practices and protect the rights of individuals to act on their accountability when it conflicts with their organization. This approach requires governments to become more agile in their approaches to technology regulation and establish guidelines that protect citizens from harm while allowing for continued innovation and investigation into the positive possibilities of AI technologies. As such, bottom-up pressure from these entities, matched with continued calls from citizens, popular media, researchers, and beyond, can further encourage the change needed to support responsible AI governance from the top-down as well.

5.4 Limitations and Future Work

This study comes with several limitations, which we outline here. First, our participant sample heavily skews toward white/Caucasian and male individuals. We recognize that because our study pulls heavily from these demographics, we can create our own forms of bias with this research. However, given the lack of representation in the overall AI development field and our focus on diversifying other components of their work, we felt that this demographic pool was sufficient for the present research. In future studies, we aim to recruit more diverse populations within the field to understand the nuances of AI bias accountability. Second, despite the global AI development field, our work also takes a heavily U.S.-centric perspective. We recognize that these issues take on various forms when seen from a global perspective, and growing regulations around AI's effects could encourage greater awareness and culpability toward these aims. Future studies should include global perspectives and/or focus on specific geographic regions that provide useful points of comparison that can highlight and even address the flaws seen in the U.S. approaches to AI responsibility. Third, while we did recruit many different roles within these AI development organizations, there was not an evident balance between the varying power levels in the sample. Indeed, we often only had one person representing a specific role in their organization, which may come with its unique experiences of accountability and agency. Thus, more research into the differing dynamics across entry-level, middle-level, and executive-level roles can better elucidate the factors that strain AI professionals' culpability based on their power. Finally, given that the ultimate goal of this work is to assist workers, organizations, and the AI community in addressing bias, this analysis prioritized the creation of codes and themes that bridge organization and job type differences. Consequently, this work did not extensively differentiate and compare these varying factors to offer comprehensive recommendations. Future research should explore these differences to provide more context-specific advice, building upon the foundational exploration presented here.

6 CONCLUSION

The urgent need to address bias in AI systems to prevent large-scale societal harms has driven governmental bodies (e.g., the *Blueprint for an AI Bill of Rights* [66]) and responsible AI researchers to explore how failures in socio-technical systems, including due to moral disengagement from responsibility, contribute to this problem [28, 49, 50, 54, 66, 71, 139]. Through in-depth interviews with 20 AI professionals from various industries and developmental positions, our work furthers this agenda by uncovering how these individuals are constrained in their motivations and perceived agency to prevent bias despite social and ethical imperatives for those designing, developing, and deploying these systems. In doing so, we explore how an individual defines what it means to have bias in an AI system, as either failing objectivity or inevitable subjectivity, actually prevents them from feeling accountable for bias mitigation *even when they have the technical ability to do so*. Our work also extends prior work that sets up ideal organizational structures for responsible AI practices [115], and inductively discovers how an individual's position within a company's internal structure and culture, and external pressures can constrain their motivation and perceived agency to confront and mitigate bias. As a result, we present three critical guidelines to support accountability for bias mitigation among individual AI professionals. Thus, our work offers fundamental building blocks necessary for further exploration into how the larger socio-technical system of responsibility, from the individual to societal level, may operate toward addressing AI bias and wider social and ethical considerations of this technology.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Wen Duan for her initial input on this work and continued support for the project.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 60–69. ISSN: 2640-3498.

[2] Alekh Agarwal, Miroslav Dudik, and Zhiwei Steven Wu. 2019. Fair regression: Quantitative definitions and reduction-based algorithms. In *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 120–129. ISSN: 2640-3498.

[3] Safinah Ali, Blakeley H. Payne, Randi Williams, Hae Won Park, and Cynthia Breazeal. 2019. Constructionism, ethics, and creativity: Developing primary and middle school artificial intelligence education. In *International Workshop on Education in Artificial Intelligence K-12 (EDUAI'19)*, Vol. 2. 1–4.

[4] Duncan N. Angwin, Kamel Mellahi, Emanuel Gomes, and Emmanuel Peter. 2016. How communication approaches impact mergers and acquisitions outcomes. *The International Journal of Human Resource Management* 27, 20 (Nov. 2016), 2370–2397. https://doi.org/10.1080/09585192.2014.985330

[5] Amanda Askell, Miles Brundage, and Gillian Hadfield. 2019. The role of cooperation in responsible AI development. *arXiv preprint arXiv:1907.04534* (2019).

[6] Albert Bandura. 1986. Social foundations of thought and action. *Englewood Cliffs, NJ* 1986, 23–28 (1986).

[7] Albert Bandura. 1990. Selective activation and disengagement of moral control. *Journal of Social Issues* 46, 1 (1990), 27–46.

[8] Albert Bandura. 1999. Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review* 3, 3 (1999), 193–209.

[9] Albert Bandura. 2001. Social cognitive theory: An agentic perspective. *Annual Review of Psychology* 52, 1 (2001), 1–26.

[10] Albert Bandura, Claudio Barbaranelli, Gian Vittorio Caprara, and Concetta Pastorelli. 1996. Mechanisms of moral disengagement in the exercise of moral agency. *Journal of Personality and Social Psychology* 71, 2 (1996), 364.

[11] Iain Barclay and Will Abramson. 2021. Identifying roles, requirements and responsibilities in trustworthy AI systems. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*. ACM, Virtual USA, 264–271. https://doi.org/10.1145/3460418.3479344

[12] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org

[13] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (June 2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[14] Genie Barton, Paul Resnick, and Nicol Turner Lee. 2019. *Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms*. Technical Report. The Brookings Institution. https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

[15] France Bélanger, Mary Beth Watson-Manheim, and Bret R. Swan. 2013. A multi-level socio-technical systems telecommuting framework. *Behaviour & Information Technology* 32, 12 (2013), 1257–1279.

[16] Haydn Belfield. 2020. Activism by the AI community: Analysing recent achievements and future prospects. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 15–21.

[17] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943* (2018).

[18] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2019. Think your artificial intelligence software is fair? Think again. *IEEE Software* 36, 4 (July 2019), 76–80. https://doi.org/10.1109/MS.2019.2908514

[19] Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604. https://doi.org/10.1162/tacl_a_00041

[20] Arpita Biswas, Marta Kolczynska, Saana Rantanen, and Polina Rozenshtein. 2020. The role of in-group bias and balanced data: A comparison of human and machine recidivism risk predictions. In *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies*. 97–104.

[21] Ann Blandford. 2013. *Semi-Structured Qualitative Studies* (2nd ed.). The Interaction Design Foundation, Aarhus, Denmark. https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/semi-structured-qualitative-studies

[22] William Boag, Harini Suresh, Bianca Lepe, and Catherine D'Ignazio. 2022. Tech worker organizing for power and accountability. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 452–463.

[23] Veronika Bogina, Alan Hartman, Tsvi Kuflik, and Avital Shulner-Tal. 2022. Educating software and AI stakeholders about algorithmic fairness, accountability, transparency and ethics. *International Journal of Artificial Intelligence in Education* 32, 3 (2022), 808–833.

[24] Jason Borenstein and Ayanna Howard. 2021. Emerging challenges in AI and the need for AI ethics education. *AI and Ethics* 1, 1 (Feb. 2021), 61–65. https://doi.org/10.1007/s43681-020-00002-7

[25] Stephen Brammer and Andrew Millington. 2005. Corporate reputation and philanthropy: An empirical analysis. *Journal of Business Ethics* 61 (2005), 29–44.

[26] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, Hyrum Anderson, Heather Roff, Gregory C. Allen, Jacob Steinhardt, Carrick Flynn, Seán Ó hÉigeartaigh, Simon Beard, Haydn Belfield, Sebastian Farquhar, Clare Lyle, Rebecca Crootof, Owain Evans, Michael Page, Joanna Bryson, Roman Yampolskiy, and Dario Amodei. 2018. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv:1802.07228 [cs]* (Feb. 2018). http://arxiv.org/abs/1802.07228 arXiv: 1802.07228.

[27] Francois Buet-Golfouse and Islam Utyagulov. 2022. Towards fair unsupervised learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 1399–1409. https://doi.org/10.1145/3531146.3533197

[28] Miriam C. Buiten. 2019. Towards intelligent regulation of artificial intelligence. *European Journal of Risk Regulation* 10, 1 (March 2019), 41–59. https://doi.org/10.1017/err.2019.8

[29] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. 2017. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/hash/9a49a25d845a483fae4be7e341368e36-Abstract.html

[30] Davide Castelvecchi. 2020. Is facial recognition too biased to be let loose? *Nature* 587, 7834 (2020), 347–350.

[31] Corinne Cath. 2018. Governing artificial intelligence: Ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 2133 (2018), 20180080.

[32] Kathy Charmaz. 2006. *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*. SAGE. Google-Books-ID: v1qP1KbXz1AC.

[33] Lu Cheng, Kush R. Varshney, and Huan Liu. 2021. Socially responsible AI algorithms: Issues, purposes, and challenges. *Journal of Artificial Intelligence Research* 71 (Aug. 2021), 1137–1181. https://doi.org/10.1613/jair.1.12814

[34] Amit K. Chopra and Munindar P. Singh. 2018. Sociotechnical systems and ethics in the large. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 48–53.

[35] Jac Christis and Erik Soepenberg. 2015. Lowlands sociotechnical design theory and lean production. In *Co-Creating Humane and Innovative Communities of Work: Evolutions in the Practice of Socio-Technical System Design*. Global STS-D Network Press.

[36] Peter Cihon, Matthijs M. Maas, and Luke Kemp. 2020. Should artificial intelligence governance be centralised?: Design lessons from history. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York NY USA, 228–234. https://doi.org/10.1145/3375627.3375857

[37] Mark Coeckelbergh. 2020. Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics* 26, 4 (Aug. 2020), 2051–2068. https://doi.org/10.1007/s11948-019-00146-8

[38] European Commission. 2019. *Ethics Guidelines for Trustworthy AI | Shaping Europe's Digital Future*. Technical Report. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[39] Mihaela Constantinescu, Cristina Voinea, Radu Uszkai, and Constantin Vică. 2021. Understanding responsibility in responsible AI. Dianoetic virtues and the hard problem of context. *Ethics and Information Technology* 23, 4 (Dec. 2021), 803–814. https://doi.org/10.1007/s10676-021-09616-9

[40] Katy Cook. 2020. *The Psychology of Silicon Valley: Ethical Threats and Emotional Unintelligence in the Tech Industry*. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-27364-4

[41] Juliet Corbin and Anselm Strauss. 2008. *Basics of Qualitative Research (3rd ed.): Techniques and Procedures for Developing Grounded Theory*. SAGE Publications, Inc., Thousand Oaks, California. https://doi.org/10.4135/9781452230153

[42] Lorrie Faith Cranor. 2008. A framework for reasoning about the human in the loop. In *Proceedings of the 1st Conference on Usability, Psychology, and Security* (San Francisco, California) *(UPSEC'08)*. USENIX Association, USA, Article 1, 15 pages.

[43] Paulo V. R. de Carvalho. 2006. Ergonomic field studies in a nuclear power plant control room. *Progress in Nuclear Energy* 48, 1 (2006), 51–69.

[44] Advait Deshpande and Helen Sharp. 2022. Responsible AI systems: Who are the stakeholders?. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Oxford United Kingdom, 227–236. https://doi.org/10.1145/3514094.3534187

[45] Virginia Dignum. 2017. Responsible autonomy. *arXiv preprint arXiv:1706.02513* (2017).

[46] Virginia Dignum. 2019. Ensuring responsible AI in practice. In *Responsible Artificial Intelligence*. Springer International Publishing, Cham, 93–105. https://doi.org/10.1007/978-3-030-30371-6_6 Series Title: Artificial Intelligence: Foundations, Theory, and Algorithms.

[47] Jessica Maria Echterhoff, Matin Yarmand, and Julian McAuley. 2022. AI-moderated decision-making: Capturing and balancing anchoring bias in sequential decision tasks. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–9. https://doi.org/10.1145/3491102.3517443

[48] Ray Eitel-Porter. 2021. Beyond the promise: Implementing ethical AI. *AI and Ethics* 1, 1 (Feb. 2021), 73–80. https://doi.org/10.1007/s43681-020-00011-6

[49] Joshua Ellul, Gordon Pace, Stephen McCarthy, Trevor Sammut, Juanita Brockdorff, and Matthew Scerri. 2021. Regulating artificial intelligence: A technology regulator's perspective. *São Paulo* (2021), 5.

[50] Olivia J. Erdélyi and Judy Goldsmith. 2018. Regulating artificial intelligence: Proposal for a global solution. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New Orleans LA USA, 95–101. https://doi.org/10.1145/3278721.3278731

[51] Leon Festinger. 1957. *A Theory of Cognitive Dissonance*. Stanford University Press.

[52] Kenneth Fleischmann and William Wallace. 2010. Value conflicts in computational modeling. *Computer* 43, 7 (July 2010), 57–63. https://doi.org/10.1109/MC.2010.120

[53] Kenneth R. Fleischmann, Sherri R. Greenberg, Danna Gurari, Abigale Stangl, Nitin Verma, Jaxsen R. Day, Rachel N. Simons, and Tom Yeh. 2019. Good systems: Ethical AI for CSCW. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. 461–467.

[54] Sorelle A. Friedler and Christo Wilson. 2018. Preface. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 1–2. https://proceedings.mlr.press/v81/friedler18a.html ISSN: 2640-3498.

[55] Batya Friedman, Peter Kahn, and Alan Borning. 2002. Value Sensitive Design: Theory and Methods. *University of Washington Technical Report* 2 (2002), 12.

[56] Runshan Fu, Yan Huang, and Param Vir Singh. 2020. Artificial intelligence and algorithmic bias: Source, detection, mitigation, and implications. In *Pushing the Boundaries: Frontiers in Impactful OR/OM Research*. INFORMS, 39–63. https://doi.org/10.1287/educ.2020.0215

[57] Bhavya Ghai and Klaus Mueller. 2022. D-bias: A causality-based human-in-the-loop system for tackling algorithmic bias. *IEEE Transactions on Visualization and Computer Graphics* (2022), 1–10. https://doi.org/10.1109/TVCG.2022.3209484

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

[59] Karen Hao. 2019. In 2020, let's stop AI ethics-washing and actually do something. *MIT Technology Review* (2019). https://www.technologyreview.com/2019/12/27/57/ai-ethics-washing-time-to-act/

[60] Maria Hedlund and Erik Persson. 2022. Expert responsibility in AI development. *AI & Society* (June 2022). https://doi.org/10.1007/s00146-022-01498-9

[61] Merve Hickok. 2021. Lessons learned from AI ethics principles for future actions. *AI and Ethics* 1, 1 (Feb. 2021), 41–47. https://doi.org/10.1007/s43681-020-00008-1

[62] Minna-Maaria Hiekkataipale and Anna-Maija Lämsä. 2019. (A) moral agents in organisations? The significance of ethical organisation culture for middle managers' exercise of moral agency in ethical problems. *Journal of Business Ethics* 155, 1 (2019), 147–161.

[63] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv:1805.03677 [cs]* (May 2018). http://arxiv.org/abs/1805.03677 arXiv: 1805.03677.

[64] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/3290605.3300830

[65] Christopher Hood. 2010. *The Blame Game: Spin, Bureaucracy, and Self-Preservation in Government*. Princeton University Press. https://doi.org/10.1515/9781400836819

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

[67] Lily Hu. 2021. Tech ethics: Speaking ethics to power, or power speaking ethics? *Journal of Social Computing* 2, 3 (2021), 238–248.

[68] Dietmar Hübner. 2021. Two kinds of discrimination in AI-based penal decision-making. *ACM SIGKDD Explorations Newsletter* 23, 1 (May 2021), 4–13. https://doi.org/10.1145/3468507.3468510

[69] IBM. 2021. What is Human-Centered AI? https://research.ibm.com/blog/what-is-human-centered-ai

[70] Zahoor Ul Islam. 2021. Software engineering methods for responsible artificial intelligence. (2021), 2.

[71] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (Sept. 2019), 389–399. https://doi.org/10.1038/s42256-019-0088-2

[72] Charlotte A. Jones and Aidan Davison. 2021. Disempowering emotions: The role of educational experiences in social responses to climate change. *Geoforum* 118 (Jan. 2021), 190–200. https://doi.org/10.1016/j.geoforum.2020.11.006

[73] Robert A. Kagan, Neil Gunningham, and Dorothy Thornton. 2003. Explaining corporate environmental performance: how does regulation matter? *Law & Society Review* 37, 1 (2003), 51–90.

[74] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*. 1–6. https://doi.org/10.1109/IC4.2009.4909197

[75] Faisal Kamiran, Sameen Mansha, Asim Karim, and Xiangliang Zhang. 2018. Exploiting reject option in classification for social discrimination control. *Inf. Sci. (Ny)* 425 (Jan. 2018), 18–33.

[76] Constance E. Kampf. 2018. Connecting corporate and consumer social responsibility through social media activism. *Social Media + Society* 4, 1 (Jan. 2018), 205630511774635. https://doi.org/10.1177/2056305117746357

[77] Michael Kane. 2010. Validity and fairness. *Language Testing* 27, 2 (2010), 177–182.

[78] Fariba Karimi, Mathieu Génois, Claudia Wagner, Philipp Singer, and Markus Strohmaier. 2018. Homophily influences ranking of minorities in social networks. *Sci. Rep.* 8, 1 (July 2018), 11077.

[79] Thomas Kehrenberg, Zexun Chen, and Novi Quadrianto. 2020. Tuning fairness by balancing target labels. *Frontiers in Artificial Intelligence* 3 (May 2020), 33. https://doi.org/10.3389/frai.2020.00033

[80] Rob Kling and Susan Leigh Star. 1998. Human centered systems in the perspective of organizational and social informatics. *ACM SIGCAS Computers and Society* 28, 1 (March 1998), 22–29. https://doi.org/10.1145/277351.277356

[81] Alina Köchling and Marius Claus Wehner. 2020. Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research* 13, 3 (2020), 795–848.

[82] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18* (Lyon, France). ACM Press, New York, New York, USA.

[83] Anastassia Lauterbach and A Bonim. 2016. Artificial intelligence: A strategic business and governance imperative. *NACD Directorship* (2016), 54–57.

[84] Po-Ming Law, Sana Malik, Fan Du, and Moumita Sinha. 2020. The impact of presentation style on human-in-the-loop detection of algorithmic bias. *arXiv:2004.12388 [cs]* (May 2020). http://arxiv.org/abs/2004.12388 arXiv: 2004.12388.

[85] 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.

[86] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.

[87] Andrew Limbong. 2022. Judge approves Activision Blizzard $18 million settlement in sexual harassment suit: NPR. *NPR* (March 2022). https://www.npr.org/2022/03/29/1089577389/judge-activision-blizzard-settlement-sexual-harassment

[88] Peter Lund-Thomsen and Khalid Nadvi. 2010. Global value chains, local collective action and corporate social responsibility: A review of empirical evidence. *Business Strategy and the Environment* 19, 1 (Jan. 2010), 1–13. https://doi.org/10.1002/bse.670

[89] Kurt Luther, Andrea Kavanaugh, Jacob Thebault-Spieker, and Judd Antin. 2020. Introduction to the Special Issue on Negotiating Truth and Trust in Socio-Technical Systems. 1 page.

[90] Trisha Mahoney, Kush R. Varshney, and Michael Hind. 2020. *AI Fairness: How to Measure and Reduce Unwanted Bias in Machine Learning*. Technical Report. IBM. 21 pages.

[91] James Manyika, Jake Silberg, and Brittany Presten. 2019. What do we do about the biases in AI? *Harvard Business Review* (Oct. 2019). https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai

[92] Kirsten Martin. 2022. *Ethics of Data and Analytics*. Auerbach Publications. https://doi.org/10.1201/9781003278290

[93] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–23. https://doi.org/10.1145/3359174

[94] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[95] Debra E. Meyerson. 2004. The tempered radicals. *Stanford Social Innovation Review* 2 (2004), 1423.

[96] Stuart E. Middleton, Emmanuel Letouzé, Ali Hossaini, and Adriane Chapman. 2022. Trust, regulation, and human-in-the-loop AI. *Commun. ACM* 65, 4 (April 2022), 64–68. https://doi.org/10.1145/3511597

[97] Patrick Mikalef, Kieran Conboy, Jenny Eriksson Lundström, and Aleš Popovič. 2022. Thinking responsibly about responsible AI and 'the dark side' of AI. *European Journal of Information Systems* 31, 3 (May 2022), 257–268. https://doi.org/10.1080/0960085X.2022.2026621

[98] Katherine L. Milkman, Dolly Chugh, and Max H. Bazerman. 2009. How can decision making be improved? *Perspectives on Psychological Science* 4, 4 (July 2009), 379–383. https://doi.org/10.1111/j.1745-6924.2009.01142.x

[99] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application* 8 (2021), 141–163.

[100] Aparna Moitra, Dennis Wagenaar, Manveer Kalirai, Syed Ishtiaque Ahmed, and Robert Soden. 2022. AI and disaster risk: A practitioner perspective. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 1–20. https://doi.org/10.1145/3555163

[101] Dena F. Mujtaba and Nihar R. Mahapatra. 2019. Ethical considerations in AI-based recruitment. In *2019 IEEE International Symposium on Technology and Society (ISTAS)*. IEEE, 1–7.

[102] Deirdre K. Mulligan, Joshua A. Kroll, Nitin Kohli, and Richmond Y. Wong. 2019. This thing called fairness: Disciplinary confusion realizing a value in technology. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–36. https://doi.org/10.1145/3359221

[103] Nataliya Nedzhvetskaya and J. S. Tan. 2021. In Oxford handbook on AI governance: The role of workers in AI ethics and governance. *arXiv preprint arXiv:2108.07700* (2021).

[104] David T. Newman, Nathanael J. Fast, and Derek J. Harmon. 2020. When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes* 160 (2020), 149–167.

[105] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, Ioannis Kompatsiaris, Katharina Kinder-Kurlanda, Claudia Wagner, Fariba Karimi, Miriam Fernandez, Harith Alani, Bettina Berendt, Tina Kruegel, Christian Heinze, Klaus Broelemann, Gjergji Kasneci, Thanassis Tiropanis, and Steffen Staab. 2020. Bias in data-driven artificial intelligence systems–An introductory survey. *WIREs Data Mining and Knowledge Discovery* 10, 3 (May 2020). https://doi.org/10.1002/widm.1356

[106] Nick Obradovich and Scott M. Guenther. 2016. Collective responsibility amplifies mitigation behaviors. *Climatic Change* 137, 1-2 (July 2016), 307–319. https://doi.org/10.1007/s10584-016-1670-9

[107] National Research Foundation: Government of Singapore. 2021. *AI Singapore*. Technical Report. National Research Foundation: Government of Singapore. https://www.nrf.gov.sg/programmes

[108] Will Orr and Jenny L. Davis. 2020. Attributions of ethical responsibility by Artificial Intelligence practitioners. *Information, Communication & Society* 23, 5 (2020), 719–735.

[109] Emmanouil Papagiannidis, Patrick Mikalef, John Krogstie, and Kieran Conboy. 2022. From responsible AI governance to competitive performance: The mediating role of knowledge management capabilities. In *The Role of Digital Technologies in Shaping the Post-Pandemic World (Lecture Notes in Computer Science)*, Savvas Papagiannidis, Eleftherios Alamanos, Suraksha Gupta, Yogesh K. Dwivedi, Matti Mäntymäki, and Ilias O. Pappas (Eds.). Springer International Publishing, Cham, 58–69. https://doi.org/10.1007/978-3-031-15342-6_5

[110] Raja Parasuraman and Dietrich H. Manzey. 2010. Complacency and bias in human use of automation: An attentional integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 52, 3 (June 2010), 381–410. https://doi.org/10.1177/0018720810376055

[111] Laura Petitta, Tahira M. Probst, and Claudio Barbaranelli. 2017. Safety culture, moral disengagement, and accident underreporting. *Journal of Business Ethics* 141 (2017), 489–504.

[112] Thao Phan, Jake Goldenfein, Monique Mann, and Declan Kuch. 2022. Economies of virtue: The circulation of 'ethics' in Big Tech. *Science as Culture* 31, 1 (2022), 121–135.

[113] 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 (April 2021), 1–25. https://doi.org/10.1145/3449205

[114] Arun Rai. 2020. Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science* 48, 1 (Jan. 2020), 137–141. https://doi.org/10.1007/s11747-019-00710-5

[115] 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 (April 2021), 1–23. https://doi.org/10.1145/3449081

[116] Dremliuga Roman and Prisekina Natalia. 2019. Artificial intelligence legal policy: Limits of use of some kinds of AI. In *Proceedings of the 2019 8th International Conference on Software and Computer Applications*. ACM, Penang Malaysia, 343–346. https://doi.org/10.1145/3316615.3316627

[117] Günter Ropohl. 1999. Philosophy of socio-technical systems. *Society for Philosophy and Technology Quarterly Electronic Journal* 4, 3 (1999), 186–194.

[118] Drew Roselli, Jeanna Matthews, and Nisha Talagala. 2019. Managing bias in AI. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco USA). ACM, New York, NY, USA.

[119] Eliane Röösli, Brian Rice, and Tina Hernandez-Boussard. 2021. Bias at warp speed: How AI may contribute to the disparities gap in the time of COVID-19. *Journal of the American Medical Informatics Association* 28, 1 (Jan. 2021), 190–192. https://doi.org/10.1093/jamia/ocaa210

[120] Johnny Saldana. 2015. *The Coding Manual for Qualitative Researchers*. SAGE. Google-Books-ID: jh1iCgAAQBAJ.

[121] Filippo Santoni de Sio and Giulio Mecacci. 2021. Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology* 34, 4 (Dec. 2021), 1057–1084. https://doi.org/10.1007/s13347-021-00450-x

[122] Anton Saveliev and Denis Zhurenkov. 2020. Artificial intelligence and social responsibility: The case of the artificial intelligence strategies in the United States, Russia, and China. *Kybernetes* 50, 3 (Jan. 2020), 656–675. https://doi.org/10.1108/K-01-2020-0060

[123] Devansh Saxena, Erhardt Graeff, Shion Guha, EunJeong Cheon, Pedro Reynolds-Cuéllar, Dawn Walker, Christoph Becker, and Kenneth R. Fleischmann. 2020. Collective organizing and social responsibility at CSCW. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. 503–509.

[124] Ulf Schaefer and Onno Bouwmeester. 2021. Reconceptualizing moral disengagement as a process: Transcending overly liberal and overly conservative practice in the field. *Journal of Business Ethics* 172 (2021), 525–543.

[125] Lewin Schmitt. 2022. Mapping global AI governance: A nascent regime in a fragmented landscape. *AI and Ethics* 2, 2 (May 2022), 303–314. https://doi.org/10.1007/s43681-021-00083-y

[126] Klaus Schwab. 2018. *The Global Gender Gap Report 2018*. Technical Report. World Economic Forum, Geneva, Switzerland. 355 pages. OCLC: 1126007340.

[127] Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. 2022. *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*. Technical Report. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.1270

[128] Genevieve Smith and Ishita Rustagi. 2020. *Mitigating Bias in Artificial Intelligence: An Equity Fluent Leadership Playbook*. Technical Report. Berkeley Haas Center for Equity, Gender and Leadership. https://haas.berkeley.edu/wp-content/uploads/UCB_Playbook_R10_V2_spreads2.pdf

[129] Ramya Srinivasan and Ajay Chander. 2021. Biases in AI systems. *Commun. ACM* 64, 8 (Aug. 2021), 44–49. https://doi.org/10.1145/3464903

[130] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1630–1640. https://doi.org/10.18653/v1/P19-1159

[131] Mariarosaria Taddeo and Luciano Floridi. 2018. How AI can be a force for good. *Science* 361, 6404 (Aug. 2018), 751–752. https://doi.org/10.1126/science.aat5991

[132] Daniel W. Tigard. 2021. Responsible AI and moral responsibility: A common appreciation. *AI and Ethics* 1, 2 (May 2021), 113–117. https://doi.org/10.1007/s43681-020-00009-0

[133] ERIC Trist and FRED Emery. 2005. Socio-technical systems theory. *Organizational Behavior 2: Essential Theories of Process and Structure* 169 (2005).

[134] Eric Lansdown Trist and Kenneth W. Bamforth. 1951. Some social and psychological consequences of the Longwall method of coal-getting: An examination of the psychological situation and defences of a work group in relation to the social structure and technological content of the work system. *Human Relations* 4, 1 (1951), 3–38.

[135] Alexis Tsoukias. 2021. Social responsibility of algorithms: An overview. *EURO Working Group on DSS* (2021), 153–166.

[136] Diederick van Thiel and Willem Frederik (Fred) van Raaij. 2019. Artificial intelligence credit risk prediction: An empirical study of analytical artificial intelligence tools for credit risk prediction in a digital era. *Journal of Risk Management in Financial Institutions* 12, 3 (2019), 268–286.

[137] Guy H. Walker, Neville A. Stanton, Paul M. Salmon, and Daniel P. Jenkins. 2008. A review of sociotechnical systems theory: A classic concept for new command and control paradigms. *Theoretical Issues in Ergonomics Science* 9, 6 (2008), 479–499.

[138] Cale Guthrie Weissman. 2017. This is what caused Uber's Broken Company Culture. *Fast Company* (Feb. 2017). https://www.fastcompany.com/3068475/this-is-what-caused-ubers-broken-company-culture

[139] Karl Werder, Balasubramaniam Ramesh, and Rongen (Sophia) Zhang. 2022. Establishing data provenance for responsible artificial intelligence systems. *ACM Transactions on Management Information Systems* 13, 2 (June 2022), 1–23. https://doi.org/10.1145/3503488

[140] Christopher D. Wickens, Benjamin A. Clegg, Alex Z. Vieane, and Angelia L. Sebok. 2015. Complacency and automation bias in the use of imperfect automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57, 5 (Aug. 2015), 728–739. https://doi.org/10.1177/0018720815581940

[141] Anna Wiedemann, Nicole Forsgren, Manuel Wiesche, Heiko Gewald, and Helmut Krcmar. 2019. Research for practice: The DevOps phenomenon. *Commun. ACM* 62, 8 (July 2019), 44–49. https://doi.org/10.1145/3331138

[142] 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*. 377–384.

[143] Karen Yeung. 2019. *Responsibility and AI: A Study of the Implications of Advanced Digital Technologies (including AI systems) for the Concept of Responsibility within a Human Rights Framework*. Technical Report DGI (2019) 05. The Expert Committee on human rights dimensions of automated data processing and different forms of artificial intelligence (MSI-AUT). 83 pages. https://rm.coe.int/responsability-and-ai-en/168097d9c5

[144] Richard Zemel. 2013. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, Vol. 28. JMLR, Atlanta, Georgia, 9.

[145] Baobao Zhang, Markus Anderljung, Lauren Kahn, Noemi Dreksler, Michael C. Horowitz, and Allan Dafoe. 2021. Ethics and governance of artificial intelligence: Evidence from a survey of machine learning researchers. *Journal of Artificial Intelligence Research* 71 (Aug. 2021). https://doi.org/10.1613/jair.1.12895

[146] Daniel Zhang, Nestor Maslej, Erik Brynjolfsson, John Etchemendy, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Michael Sellitto, Ellie Sakhaee, Yoav Shoham, Jack Clark, and Raymond Perrault. 2022. *The AI Index 2022 Annual Report*. Technical Report. AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University. 229 pages.

[147] Lu Zhang, Yongkai Wu, and Xintao Wu. 2016. A causal framework for discovering and removing direct and indirect discrimination. *arXiv:1611.07509 [cs]* (Nov. 2016). http://arxiv.org/abs/1611.07509 arXiv: 1611.07509.

[148] Yunfeng Zhang, Rachel Bellamy, Q. Vera Liao, and Moninder Singh. 2021. Introduction to AI fairness. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–3. https://doi.org/10.1145/3411763.3444998

Received 21 March 2023; accepted 22 March 2023