Surveys Considered Harmful? Reflecting on the Use of Surveys in AI Research, Development, and Governance

Mohammad Tahaei[1, 2], **Daricia Wilkinson**[3], **Alisa Frik**[1], **Michael Muller**[4], **Ruba Abu-Salma**[5], **Lauren Wilcox**[2]

[1]International Computer Science Institute, [2]eBay, [3]Arizona State University, [4]IBM Research, [5]King's College London
mtahaei@icsi.berkeley.edu, daricia.wilkinson@asu.edu, afrik@icsi.berkeley.edu, michael_muller@us.ibm.com, ruba.abu-salma@kcl.ac.uk, lgw231@acm.org

Abstract

Calls for engagement with the public in Artificial Intelligence (AI) research, development, and governance are increasing, leading to the use of surveys to capture people's values, perceptions, and experiences related to AI. In this paper, we critically examine the state of human participant surveys associated with these topics. Through both a reflexive analysis of a survey pilot spanning six countries and a systematic literature review of 44 papers featuring public surveys related to AI, we explore prominent perspectives and methodological nuances associated with surveys to date. We find that public surveys on AI topics are vulnerable to specific Western knowledge, values, and assumptions in their design, including in their positioning of ethical concepts and societal values, lack sufficient critical discourse surrounding deployment strategies, and demonstrate inconsistent forms of transparency in their reporting. Based on our findings, we distill provocations and heuristic questions for our community, to recognize the limitations of surveys for meeting the goals of engagement, and to cultivate shared principles to design, deploy, and interpret surveys cautiously and responsibly.

1 Introduction

Artificial Intelligence (AI)[1] and Machine Learning (ML) researchers, developers, and policymakers are increasingly using surveys to capture people's values, perceptions, and experiences, to inform development and governance of AI. Surveys are used to guide the design and development of new technology directions and products (e.g., Alkhathlan et al. 2024; Sindermann et al. 2021; Persson, Laaksoharju, and Koga 2021; Othman 2023; Loefflad and Grossklags 2024; Davani et al. 2024), shape companies' technology policies (Anthropic 2023; Google 2024; OpenAI 2023), and inform national and international policies (Ada Lovelace Institute and Alan Turing Institute 2022; AI.gov 2023a; United Nations 2022). However, critical perspectives caution that if human participant research methods are poorly designed or applied (e.g., embed biases or lack context), they may fail to serve their intended purpose, possibly leading to ethics and participation washing, and other forms of harm, instead of being beneficial (Cooper et al. 2022; Groves et al. 2023).

In this paper, we critically examine the use of surveys in AI research, development, and governance, as they are recurringly used to assess people's subjective views and experiences of AI (van Berkel, Sarsenbayeva, and Goncalves 2023). Surveys, and related research instruments such as questionnaires, inherently employ abstraction and reduction as methods of knowing and understanding (Ornstein 2013), which may result in overlooking nuances that on the surface level may seem subtle, but in practice can result in amplifying biases and leading to harms (Bhopal et al. 2004; Proctor and Schiebinger 2008; Roberts 2012). The potential misrepresentation of marginalized perspectives by surveys, though unexplored in the AI domain, has been evident in other fields (Mir et al. 2012; Nazroo et al. 2007; Nierkens, de Vries, and Stronks 2006; Agyemang et al. 2009). For example, in the United Kingdom, "*nationally-representative*"[2] surveys measuring tobacco and alcohol use showed significant discrepancies in data collected from minorities (Bhopal et al. 2004). Results varied substantially for marginalized ethnic groups between different survey agencies, e.g., one set of results reporting a 1% smoking rate in Bangladeshi women, and another reporting 6% in the same group, a discrepancy not observed in the majority group self-identifying as European (Bhopal et al. 2004). Other researchers also showed that commercial and government entities have postponed or prevented action on critical public health matters for vulnerable groups as a result of poor survey research practices (Proctor and Schiebinger 2008). In parallel, researchers criticize the focus of AI research on Western, Educated, Industrial, Rich, and Democratic (WEIRD) populations, arguing that it may not accurately represent the expe-

[1]We use the term "*AI*" broadly in our Introduction and Discussion, adopting the US National Institute for Standards and Technology (NIST) definition of the term, as "*an engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy*." (NIST 2023) We acknowledge that its meaning and scope remain in flux. Our paper includes a systematic literature review, so we rely on authors' definitions of AI when presenting analyses of their work.

[2]Throughout this paper, we refer to "*representation*," "*representative*," or "*nationally-representative*" as it is used in the cited resources. The notion of representation in ML has been examined by Chasalow and Levy (2021), and we will also address limitations of representation in surveys in our discussion.

riences and concerns of diverse global populations affected by or interacting with AI (Septiandri et al. 2023; van Berkel, Sarsenbayeva, and Goncalves 2023).

Despite these limitations, the use of surveys is expanding rapidly to capture values and normative expectations, and monitor AI-related impacts on people (e.g., Jakesch et al. 2022; Scharowski et al. 2023; Ribeiro et al. 2019; Kramer et al. 2018; Arai and Matsumoto 2023; Bartneck, Yogeeswaran, and Sibley 2023; Ikkatai et al. 2023; Yigitcanlar et al. 2020). This pattern follows a trend that began in computing research in the 1980s (Fowler 2013). Surveys now directly shape a number of high-stakes AI projects, including finance (Hertzberg, Liberti, and Paravisini 2010), employment (Franken and Wattenberg 2019; Houser 2019), smart healthcare (Morley et al. 2020; Sunarti et al. 2021), transportation (Bharadiya 2023), and education (Blodgett and Madaio 2021; Zanetti, Iseppi, and Cassese 2019; Lünich and Keller 2024). They are also used by companies creating the most popular generative models—often with the embedded assumption that they can capture diverse perspectives and the contextual and cultural specificity of the subject matter, e.g., Collective Intelligence Project (Huang et al. 2024), Our Life with AI (Google 2024), and the Moral Machine experiment, which is *"developing global, socially acceptable principles for machine ethics."* (Awad et al. 2018)

This paper contributes to the research challenging the use of decontextualized, unidirectional human participant research methods merely as a means to justify technological advancement (Cooper et al. 2022; Groves et al. 2023; Sloane et al. 2022). It aligns with a broader call within the sociotechnical research community to scrutinize research methods, processes, and practices, not just artifacts or outcomes (Mann and Daly 2019; Irani et al. 2010; Dourish et al. 2020; Ali 2016; Cooper et al. 2022). Our goal is to guide people using such methods toward a critical examination of AI-related survey design processes and data collection, analysis, and interpretation methods—toward more equitable and just research practices that respect, rather than misrepresent or exploit, surveyed communities.

The research questions (RQs) motivating our work are:

- **RQ1.** How has research in the relevant body of literature positioned surveys as a way of understanding human values, perceptions, and experiences with regard to AI? What are the elements reported in the literature, and what are the cultural and methodological implications of foregrounding those elements?
- **RQ2.** Extending the body of knowledge related to survey methods and epistemology, what are the unique questions that could guide the ethical design, deployment, interpretation, and reporting of (large-scale) survey research on AI topics with human participants, at the particular intersection of survey methodology, AI, and society?

We integrate a **reflexive analysis** of an international survey pilot with a **systematic literature review**, to critically examine our assumptions and offer a set of provocations that are vital for the AI research community, as public surveys gain an increasingly stronger foothold in the field. We walk through the design and testing of a pilot survey we conducted to reflect on design decisions and findings with respect to key provocations, and complement this analysis with a review of the methods of 44 survey research papers.

Papers in our corpus aimed to include large sample sizes ($median = 607$ participants). Although 14 out of 44 reviewed papers claimed to have representative samples, inconsistent use of the term "*representation*" created an illusion of representation rather than engaging with representation in a meaningful way. This is concerning, as misrepresentation could harm marginalized communities, limit our understanding of such groups, and perpetrate incorrect narratives about AI. Only six papers included authors from the Global South (as per the definition of Finance Center for South-South Cooperation (2024)), 11 papers lacked authors from the countries where the studies were conducted, and 38 papers lacked feedback from participants during the design stage. We also reflect on our pilot survey's design and provide a set of heuristic questions related to the use of AI in the design and analysis of surveys, the transparency of research platforms (e.g., Prolific), and the potential harms of conducting surveys across different cultures to characterize perspectives without using culture-sensitive approaches. These questions aim to help the community rethink who controls the data and how it's obtained, for what purpose it is used, and how the results are interpreted and disseminated.

We argue that approaches to human participant surveys designed to reach the public to inquire about topics related to AI must be critically examined for their role in perpetuating and maintaining the potential to amplify and exacerbate worrisome power dynamics (Baeza-Yates 2018; Nicoletti and Bass 2023). Shifting from unidirectional survey designs to co-creating survey instruments with the impacted communities—ensuring that the surveys are not only *about* people but designed *with* them—could enable researchers to include multiple knowledge systems and account for power dynamics embedded in knowledge production processes (Alvarado Garcia et al. 2021; Bird 2020; Kwet 2019).

2 Related Work

A Brief History of Surveys: From Agriculture and Military to Computing and AI

Modern-day survey methods emerged from a long history of societies that sought measurements of their populations through censuses to make plans essential to core governance (e.g., managing food supplies, distributing land, and managing taxation)—dating back to ancient times (Rossi, Wright, and Anderson 2013; Midena and Yeo 2022). As the focus of studies became more specific, such as examining the economic status of households or conducting consumer research, surveys gained increased popularity over traditional small-scale experimental studies (Rossi, Wright, and Anderson 2013). Surveys also played a prominent role in the development and study of *psychometrics*, or measuring people's mental activities. Important concepts in psychometrics include correlation, personality scale and psychometric reliability, experimental designs, and increasingly sophisticated statistical analysis (Rust and Golombok 2014). Significantly abused by *racism* and *oppressive* applications in social poli-

cies (Reyes 2019; Winston 2020), this evolution is also marked by significant developments, including the emergence of literature on questionnaire design, the introduction of standardized scales like the Likert scale for attitude measurement (Groves et al. 2009; Lee et al. 2002), the establishment of state-supported institutes dedicated to survey research, and the development of technology-assisted survey-taking tools (Rossi, Wright, and Anderson 2013; Fowler 2013). A notable example of large-scale survey usage is the extensive surveying of American soldiers returning from World War II—which provided useful norms for design but systemically excluded women (Epstein et al. 2013). Over time, this movement also created a divide in the empirical research community. One school of thought regards surveys as the "*language of empirical social research*," (Ornstein 2013) while others criticize the dominance of survey-based scholarly work that is divorced from theory—coined by Mills as "*abstracted empiricism*" (Mills 2023).

According to Arnstein's ladder of citizen participation (Arnstein 1969), which describes eight levels of citizen involvement in planning processes in the US, attitude surveys and public inquiries belong to the fourth rung of citizen participation, called "*consultation*." In other words, consulting public opinions is a legitimate step toward understanding their perspectives, but it is not sufficient if used alone. Moreover, if the exchange of data is unidirectional, the providers of the data are viewed as "*sources*" of information, raising concerns about the extractive versus participatory nature of surveys when engaging with the public (Ada Lovelace Institute 2021). Arnstein (1969) encourages the use of more direct citizen participation modes like committees, partnerships, and community engagement that allow people to engage in planning, decision-making, and policy-making more actively (see also work in participatory survey design (Smith, Christopher, and McCormick 2004; Tillyard and DeGennaro Jr 2019) and in action research (Arcaya et al. 2018; Hayes 2014)).

Despite these criticisms, surveys have become widely used in contemporary research, employed for collecting structured qualitative and statistical data and gaining quantified insights into people' perspectives and attitudes—a prominent methodological cornerstone in areas such as public opinion polls and large-sample approaches to computing research disciplines like HCI and AI more broadly (Fowler 2013).[3] For example, surveys are among the primary methods used to capture user engagement with mass market user interfaces and to provide insights into users' attitudes, experiences, demographics, and psychological characteristics shaping their behavior with technology (Doherty and Doherty 2018; Müller, Sedley, and Ferrall-Nunge 2014). Several best practices have been suggested to plan, design, and conduct effective surveys in the real world (Rea and Parker 2014; Converse and Presser 1986; Dillman et al. 1978; Dillman, Tortora, and Bowker 1998; Fowler Jr and Mangione 1990; Kelley et al. 2003; Krosnick 1999; Center 2024; Tourangeau, Rips, and Rasinski 2000; Tourangeau and Smith 1996; Brown 2023), from mail and telephone surveys to web surveys, to help participants comprehend questions, retrieve the information necessary to answer questions, and judge how much information they need to provide (Cannell et al. 1977).

Following the trend of using surveys in empirical human participant studies, researchers at AIES and other sociotechnical AI research communities have been using surveys to investigate public perceptions of AI, including the societal risks and expectations associated with AI. For example, van Berkel, Sarsenbayeva, and Goncalves (2023) found that most papers (published at prominent venues) documenting studies with human participants (65%, 130 out of 200 papers included in the review) have used surveys to capture perceptions of broadly-defined AI fairness. **This abundance of survey research underscores the importance of empirically investigating the impact of the use of this method to explore AI-related societal issues (Said et al. 2023).**

Rapid Expansion of Survey Use in AI Research, Development, and Governance

As more empirical and irrefutable evidence emerges, it becomes clear that understanding AI's impact requires a multi-stakeholder effort (Aragon et al. 2022; Delgado et al. 2023; Himmelreich 2023; Torkamaan et al. 2024; Havens et al. 2020). Yet, given the prevailing power asymmetries in this space, AI development is predominantly shaped by industry, research, and policy (Moloi and Marwala 2021). This dynamic has only recently begun to shift toward public involvement (Birhane et al. 2022a; QueerInAI 2023; Sloane et al. 2022; Dennler et al. 2023), though scholars have long advocated for increased public involvement in science and technology as a means to "*foster greater accountability, better decision outcomes, and increased trust*" (Holdren, Sunstein, and Siddiqui 2011; Bao et al. 2022).

As such, multiple studies have engaged with the public to investigate AI's impact. A plethora of surveys aiming to be representative have been deployed over the years with the goal of examining public awareness of, perceived challenges with, and trust in AI (Zhang and Dafoe 2020) and how the public understands AI (Selwyn et al. 2020; Kieslich and Lünich 2024), as well as distinguishing between utopian and dystopian narratives surrounding AI (Cave, Coughlan, and Dihal 2019).[4] A "*nationally representative*" (i.e., "*results are weighted to be representative of the US adult population*" (Zhang and Dafoe 2020)) survey in 2018 with 2,000

[3] Human-Computer Interaction (HCI) is referred to as the discipline that provides foundational knowledge for both industry and academic technology research for studying interactions between humans and computers (e.g., from user interaction techniques to sociotechnical systems). User experience is a term commonly used to include, in part, the application of select HCI methods in industry and product settings.

[4] Survey methods may have very different uptake among some marginalized populations that have histories of exploitation by quantitatively-oriented majoritarian researchers (Chilisa 2019; Denzin, Lincoln, and Smith 2008; Kovach 2021; Smith 2021). Thereby, we urge caution in interpreting claims that a sample has been properly weighted. Even with a *statistical* approach to weighting, discriminatory questions may systematically reduce participation by intimidating some marginalized groups (Berry-James, Gooden, and Johnson III 2020).

Americans showed that most Americans supported AI development but also expressed deep concerns about its future impact (Zhang and Dafoe 2020).

Similarly, a "nationally representative" survey (i.e., "*weighted sample by main demographic characteristics*" (Selwyn et al. 2020)) with 2019 respondents from Australia supported the development of AI in healthcare but exhibited mixed views on its professional integration (Selwyn et al. 2020). These findings resonate with similar survey studies on public perceptions of AI in Russia, India, and the UK (Fast and Horvitz 2017; Bao et al. 2022; Kapania et al. 2022; Ada Lovelace Institute and Alan Turing Institute 2022). Survey results also influence national policies; for example, the frequent citation of survey results in US government policy and strategy documents related to AI (e.g., Science and Council 2023; AI.gov 2023a,b) and surveys' influence on the UK's AI policy (Ada Lovelace Institute and Alan Turing Institute 2022). Additionally, periodic public surveys to monitor the evolving AI landscape have been proposed, as evidenced by the National Artificial Intelligence Advisory Committee (NAIAC) and other government documents focused on strengthening AI capabilities (AI.gov 2023a,b). Public surveys are also influencing corporate strategies and future products of companies like Anthropic (Anthropic 2023) and OpenAI (OpenAI 2023).

Recent reflections on participation and inclusion in AI research have contributed to studies extending beyond WEIRD populations to build knowledge representing a broader spectrum of lived experiences (Boyon 2022; Linxen et al. 2021; Epstein et al. 2023). For example, Kelley et al. (2021)'s survey with 10,000 respondents from eight countries showed their widespread support for AI development and their hesitations because of its associated risks. Participants were hopeful about the potential for AI to improve healthcare, while their fears revolved around concerns over job loss, social isolation, and significant threats to humanity. The data reflected differences in how people in non-Western societies perceived specific risks that media discourses and various predispositional values could shape.

These examples highlight the profound impact that survey research can have on the future of AI and society. Yet, despite the recent surge in the use of surveys, prevailing disciplinary norms and best practices for their design and use for AI research are inadequate.[5] As a community, we have yet to discuss essential criteria and summarize principles for the design and use of surveys to understand people's values, perceptions, and experiences with regard to AI. **We argue that survey research on *AI-related topics* introduces unique methodological challenges and considerations that warrant field-wide attention, further methodological research inquiry, and collaborative debate.**

[5]There are already well-documented limitations and caveats in survey research to date. Examples include vulnerabilities to biases related to social desirability, order effects, and sampling. In Appendix included in our arXiv extended version, we detail some of these known issues with surveys.

Critically Reflexive Practices in AI Research

"*Critically reflexive practice embraces subjective understandings of reality as a basis for thinking more critically about the impact of our assumptions, values, and actions on others.*" (Cunliffe 2004)

Reflexivity has a history, in scholarly inquiry, of generating understandings of how underlying scholarly structures and systems influence and are influenced by actors, including scholars themselves (Bourdieu 1990; Bourdieu and Wacquant 1992; Jamieson, Govaart, and Pownall 2023). Despite the recent establishment of sociotechnical AI research communities, a body of work that critically examines the communities' methods and results is rapidly emerging (Groves et al. 2023; Laufer et al. 2022; Chasalow and Levy 2021; Miceli et al. 2021; Constantinides et al. 2024b,a; Havens et al. 2020). Laufer et al. (2022) leaned on reflexivity to provide a nuanced exploration of the "*presuppositions, underlying values and assumptions*" guiding the direction of such scholarship since its inception. Similarly, Young, Katell, and Krafft (2022) reflected on emerging conflicts of interest that inadvertently and inherently impact models of participation with audiences who contend with the negative impacts of algorithmic systems. Researchers have questioned the use of human research methods within the AI industry and their role in driving narratives and redefining norms around empirical research in AI. Groves et al. (2023) critique the approaches of commercial AI labs, such as OpenAI and Anthropic, in their public engagement strategies. They observe that business interests are often prioritized over societal needs, characterized by a lack of context, clarity in methods, and rigor, stemming from the fast-paced nature of the industry and the conflicting interests involved. Moreover, a reflexive study reveals that the majority of FAccT papers involving human participants predominantly focus on WEIRD populations, particularly those from the US (Septiandri et al. 2023). This indicates that much of AI research is influenced by American values and perspectives, potentially widening the gap in understanding marginalized communities or exacerbating existing disparities.

Building on the critical self-reflective literature in AI research, our perspective is informed by critical computing (Comber et al. 2020; Ko et al. 2023), a body of work that includes titles with "*considered harmful*" (starting in 1968 to challenge research norms and structures (Dijkstra 1968)). This resonates across various computing disciplines such as computer security (Singer and Bishop 2021) and HCI (Aragon et al. 2022; Comber et al. 2020; Greenberg and Buxton 2008; Crabtree et al. 2009). Similarly, the field of critical data studies examines how data are not *given* or even *captured* (Muller et al. 2019), but rather *designed* (Feinberg 2017, 2022) and *created* (Muller et al. 2021; Muller and Strohmayer 2022) as human-made components of larger sociotechnical assemblages of privilege and power (Iliadis and Russo 2016; Kitchin and Lauriault 2014). **This paper builds on the trove of existing and emerging research to critically examine the often-overlooked assumptions embedded in the use of surveys within AI scholarship, and to identify opportunities where the AIES community is uniquely positioned to take a leading role.**

3 Methods

Our approach and perspectives are informed by a critically reflexive stance, rooted in the self-critical perspectives of the AI research community (Section 2) and our positionality (Section 7). The rationale behind such a critical stance is to initiate discourse within the community, especially as surveys are increasingly becoming a "go-to" method for capturing public perceptions of AI. To understand the pitfalls of using public surveys in the AI domain, we employ two methods: (1) **a pilot survey as the basis for critical reflection using reflexivity** and (2) **a systematic literature review of public surveys in AI research**.

Pilot Survey of Perceived Benefits and Risks of AI

To facilitate reflexivity in survey methods, we conducted a pilot survey in six countries (one in each of six continents) with 282 participants to explore the complexities of survey research associated with the challenging topic of capturing perceptions of AI's benefits and risks. Despite adhering to *known* survey research best practices (see Appendix in the extended arXiv version), we observed that there are both *unknown knowns* and *unknown unknowns* that require further attention and are often overlooked in AI research practices.

Survey Design. The pilot survey explored perceptions of the *benefits* and *risks* of *existing* AI systems via two separate open-ended questions: (1) How do you think existing AI systems could benefit you?; (2) How do you think existing AI systems could put you at risk? We also used two storytelling questions inspired by the computer security domain (Rader, Wash, and Brooks 2012; Pfeffer et al. 2022) to capture how stories and memories about benefits and risks of AI spread within the society: (3) Write down a story that you heard from someone about benefits of AI; (4) Write down a story that you heard from someone about risks of AI. To understand participants' views on the *future* of AI systems, we used probing strategies grounded in speculative design (Marenko 2018; Auger 2013; Wong and Khovanskaya 2018) and asked: (5) If you had a magic wand that could create an AI system, what would you want that AI system to do for you? (6) How could the AI system that you just described put you (or someone else) at risk? Finally, to explore perceptions of the *trustworthy* development of AI systems, we asked: (7) What characteristics should an AI system have to be trustworthy? For this question, we used a slightly modified version of NIST's definition of AI systems (NIST 2023). More details about the survey design are provided below and in Appendix.

Refining the Pilot Survey. We employed five strategies to mitigate known issues with survey research in our pilot survey (Gilovich, Keltner, and Nisbett 2006; Albert, Tullis, and Tedesco 2009; Colton and Covert 2007; Groves et al. 2011): (1) We positioned demographic questions toward the end of our survey to mitigate potential priming and sensitivity concerns that could result from stereotype threat; (2) we incorporated two attention-check questions to identify low-quality responses; (3) we conducted expert reviews with five domain experts before deploying the survey to improve clarity; (4) we did a walk-through with people who have lived most of their lives in countries included in the survey but where authors have not resided, to capture their views on the survey design (e.g., in Australia, people may consider the benefits of AI for various aspects of their lives differently compared to other countries, with a particular emphasis on the importance of indigenous identity recognition in their region; or in Japan, the use of "stories" could reflect factual events or rumors/gossips); and (5) we conducted small survey pilots in two rounds (with six and five participants, respectively) to pinpoint areas for further clarification, before conducting the larger pilot.

Pilot Deployment. We hosted the survey on Qualtrics (Qualtrics 2023) and used Prolific (Prolific 2023), a crowdsourcing platform, for participant recruitment, between July and August 2023. Using Prolific's screening tool and its "gender-balanced" sampling,[6] we selected participants who were at least 18 years old and fluent in English, had a minimum approval rate of 95%, and resided in one of the six countries we recruited our study participants from, including Australia (AU), Chile (CL), Israel (IL), the United Kingdom (UK), the United States (US), and South Africa (ZA). These countries were chosen based on access to participants through our recruitment platform to cover one country per continent.

Recruitment. We recruited 50 participants from each of the six countries, totaling 300 participants. Six responses were removed due to failed attention checks, and 12 additional responses were discarded as we classified them as either AI-generated or copied from the Internet (based on discussions among authors). The resulting dataset is comprised of 282 responses. Participants were paid $5 USD via Prolific for completing the study. The survey's average completion time was 22 minutes ($std = 11$ minutes), with responses averaging 211 words in length.

Participant Demographics. Our sample returned an almost equal number of participants in each country (see Appendix for a summary of participant demographics). We acknowledge that our sample did not attempt to measure *internal* diversity or sample sub-populations within each country, a choice that we discuss in Appendix.

Systematic Literature Review

We conducted a systematic literature review centered on papers related to the themes of *public*, *AI*, *surveys*, and *perceptions* described in a variety of ways (see Appendix for search terms). We sourced our material from the ACM DL (conference publications) as well as from Springer's AI & Ethics and AI & Society journals (journal publications). Our search was constrained to a two-year period (01/2022–01/2024), except for papers from AIES and FAccT, for which we did not impose any date restrictions. The exclusion criteria were as follows: (1) surveys of literature, policies, or guidelines, rather than respondents; (2) use of past surveys or datasets;

[6]Prior work shows a "gender-balance" sample on Prolific is similar to its "representative" sample but costs less (Tang, Birrell, and Lerner 2022). Therefore, due to budget limits, we opted in for a "gender-balanced" sample. We discuss potential harms of this framing in Appendix.

(3) surveys not intended to include representative samples or, if purposive, not intended to have large reach; and (4) abstracts or short papers lacking detailed methods (see details in Appendix). Our final dataset consists of **44 papers**, including 15 papers from ACM conferences, 6 from AI & Ethics, and 23 from AI & Society. A spreadsheet with a list of all the papers we reviewed, along with our analysis, is available upon request.

Limitations. Our keywords span a wide range within each topic of interest, but our search may not encompass all available literature in the domain. Nevertheless, we believe our queries sufficiently capture recent trends in the use of surveys within the AI research community. We acknowledge that our understanding of the literature is influenced by our positionality and academic backgrounds (see Section 7). Future work could expand on our research to include non-academic literature, such as studies conducted by corporate research platforms like Pew or Gallup, or other agencies worldwide. Our focus was on the recent surge in AI, particularly post-generative AI, with conferences like AIES established to address the ethical implications of AI. Thus, we concentrated on literature from the past two years. Future research could expand this time frame to identify long-term patterns in the use of surveys in computing and AI.

4 Large-Scale Surveys of AI in the Literature

After analyzing our corpus of 44 papers, we noticed inconsistent practices in the reporting of research procedures (e.g., ethical review approvals, informed consents, and concerns related to cultural sensitivity and congruence of research practices) and research transparency practices (e.g., lack of information about funding sources, positionality statements, recruitment strategies, and socio-demographic characteristics of participants or their compensation).

Research Methods. Across the 44 papers, there were 58 primarily quantitative studies (some papers reported results of more than one study) conducted (e.g., surveys and online experiments, with qualitative analysis of open-ended responses), nine of which were accompanied by qualitative studies with human participants (e.g., interviews and focus groups, sometimes with quantitative insights into occurrence counting or other descriptive data). The median sample size for the quantitative studies was $n = 607$ participants ($mean = 2,668$, $min = 57$, $max = 47,951$, $std = 8,259$). The median sample size for the qualitative studies was $n = 12$ participants ($mean = 20$, $min = 8$, $max = 45$, $std = 14$).

Recruitment Strategies. Ten out of 44 papers did not report where they recruited their participants from, among those that reported it, the most common recruitment strategy was using a market research agency ($n = 11$) and online or crowdsourcing platforms ($n = 8$) (e.g., Prolific (Prolific 2023), MTurk (Amazon 2023)), or using the census panels ($n = 2$) or electoral poll databases ($n = 2$). Other recruitment strategies included snowball sampling or word-of-mouth ($n = 4$), social media ($n = 3$), internal databases, for example, in universities ($n = 3$), direct emails ($n = 3$), convenience samples ($n = 3$), online websites advertising the study ($n = 2$), and physical posters ($n = 1$). Many of these methods use online channels to recruit participants, which may exclude certain marginalized populations that do not have consistent access to the Internet or lack knowledge, skills, experience, or physical abilities to engage with those online channels. Online panels (especially, MTurk) have been previously criticized for having non-diverse or non-representative user bases (Posch et al. 2018), for data quality issues, associated with dishonesty, or low-effort responses especially among most experienced survey respondents (Peer et al. 2022; Douglas, Ewell, and Brauer 2023), and more recently, for using AI for completing tasks (Veselovsky, Ribeiro, and West 2023).

Reporting of Demographics. The reporting of socio-demographic characteristics was not consistent across the papers. Except for gender ($n = 36$) and age ($n = 34$) that were reported in the majority of papers, and education levels ($n = 20$) that were reported in slightly less than half of the papers, other characteristics like race/ethnicity, income, and employment status were reported only in a handful of papers. Fourteen papers examined samples that were census- or nationally-"representative"—one reported using a quota-stratified sample for age and gender, and one recruited participants from the general public but put effort into including marginalized groups such as people of color, gender minorities, and those with mental illnesses. The remaining papers ($n = 28$) either used random sampling or did not report sampling strategies; some of these papers reported achieving balanced samples in terms of age and gender, while others had skewed samples.

Geographic Diversity Among Authors. Only six papers had authors from the Global South, and 11 papers did not include authors from the countries where these studies were conducted (specifically, two of these 11 studies included populations from the Global South but did not include authors who lived or worked in these countries). These findings suggest trends of limited geographic diversity among authors and raise concerns that authors may not always have the sufficient cultural context about the populations they study. Two of the 44 papers reported authors' Western points of view or limitations of their stance as computing educators. There is a chance that the authors were born or have lived in countries other than the countries of authors' current affiliations and, therefore, have sufficient cultural context. However, without a positionality statement, it is hard to understand if this is actually the case. Therefore, these findings highlight both the need to strongly encourage positionality statement disclosures when studying human participants and the importance of further in-depth research to address the apparent misalignment between the geographic distribution of authors and the populations they study.

Funding Sources. Five papers had the opposite issue; these studies had authors from countries that did not include participants from those countries (e.g., authors from Japan ran a study with US participants, but not with participants from Japan). Many papers acknowledged funding support either from non-profit organizations ($n = 14$) like foundations, philanthropic funds, trusts, which are mostly funded by contributions from individual donors and organizations, or from university-sponsored ($n = 5$) and government-

provided grants ($n = 13$), a significant portion of which comes from tax payers' money and other public sources. Five papers acknowledged support from private companies, which brings to mind concerns about potential conflicts of interest (as surfaced by Birhane et al. (2022b), and issues extensively discussed in Young, Katell, and Krafft (2022)). Finally, 15 papers did not disclose their funding sources.

Continuity of Research. Most studies in our corpus were cross-sectional (i.e., the data were collected from participants at a single point in time, without periodically repeated measurements). Only one paper reported several (four) rounds of a survey; another reported findings before and after the 2020 US general elections. While one other paper ran a survey similar to another earlier one conducted by the same authors, it did not report any direct comparisons. Two other papers mentioned that their questions were part of an annual survey but did not specify if the same questions were asked again or whether they observed any shift in responses over time, and no precedent or follow-up papers for those annual surveys came up in our literature search. These findings suggest a lack of continuity of existing survey research in AI and a lack of comparability in both repeated surveys and related surveys in the literature.

Research Ethics. Many papers ($n = 20$) did not mention Institutional Review Board (IRB) or other ethical committee approvals, nor did they indicate that informed consent was obtained from participants before the studies, echoing recent findings in AI research with human participants (McKee 2023). However, these results may require a more critical view. Researchers report diversity in IRB practices around the world (Patel et al. 2013; PE and MD 2019). While IRBs are common in the US, there may be different structures in Europe, Asia, and Africa (e.g., Orimadegun 2020). The practicalities of review and consent may depend on prior experiences of marginalization and exploitation (Angal et al. 2016; Kuhn, Parker, and Lefthand-Begay 2020; Norton and Manson 1996; Tapaha 2017). Schrag (2010) summarized some of these risks in the term "*ethical imperialism*"; i.e., the imposition of review board practices from the Global North—particularly from the US—on other countries and cultures, despite differences in local values and practices.

Testing and Feedback. Most of the papers ($n = 38$) did not report any testing of the study materials before data collection. Only a few papers mentioned conducting pilot surveys ($n = 4$) or walk-through interviews or focus groups ($n = 6$) prior to launching the primary study. In addition, one paper mentioned expert consultation to review study materials, and another paper mentioned pre-registering the study on the Open Science Framework's website in order to increase research transparency by specifying analysis methods before analyzing the data. Pre-testing and feedback can enable representation of pluralistic perspectives and review of question language in instrument design before deployment (Schurgin et al. 2021), though we acknowledge that an over-emphasis on pre-testing and pre-registration may suggest a rigid, positivist approach to research, potentially overlooking the value of exploratory surveys and the interpretive role of researchers. So while these practices *could* contribute to greater rigor and transparency of research, they alone do not "pre-guarantee" validity of survey knowledge.

Evaluation and Replication. Some papers included study materials such as survey instruments and interview guides ($n = 9$), full replication packages ($n = 2$), or study data ($n = 4$), or made the study data available upon request ($n = 9$). However, 16 papers did not include additional artifacts, contributing to ongoing concerns about the replication challenge, e.g., lack of replication packages, analysis codes, or datasets (Dreber and Johannesson 2019; Freese and Peterson 2017; Echtler and Häußler 2018).

5 Discussion

Pitfalls of Surveys as Enablers of "Participation"

A recent trend in critical AI research advocates for *participation* in AI (e.g., Feffer et al. 2023; Sloane et al. 2022; Delgado et al. 2023; Bondi et al. 2021), with researchers using different terminologies and viewpoints to describe levels of participation, with frequent reference to the ladder of participation (Arnstein 1969). For example, Sloane et al. (2022) categorize participation in AI into three levels: *work*, *consultation*, and *justice*, whereas Delgado et al. (2023) propose a four-level framework: *consult*, *include*, *collaborate*, and *own*. Given that surveys are often a paid, one-time transaction, they most likely fall under the categories of participation as *work* and *consultation*—the minimum level of participation in both frameworks. Other frameworks for assessing participation in AI propose questions for researchers to consider. Birhane et al. (2022a) provide a framework focusing on *empowerment*, *reciprocity*, and *reflexivity* for critically evaluating participatory approaches in AI. Feffer et al. (2023) offer a more detailed ten-axis framework that encompasses Birhane et al.'s dimensions as well. We adopt the ten-axis framework by Feffer et al. (2023) to assess surveys as a method for participation in AI given its comprehensive integration into prior scholarship:

1. Representation. As a term, representation has often been misused in public survey literature regarding opinions on AI, which can create an *illusion* of rigor, act as a veneer for bias reduction, and give misleading perceptions of achievements in sampling. In our pilot survey, we could have used terms like "cross-cultural" or "across six continents" in our title or abstract to attract reader's attention, while the results did not truly speak to these terms.

2. Stage. The current literature primarily focuses on early stages of design, development, and governance of AI to capture attitudes and perceptions, to inform either products or policies. Surveys, like ours, often focus on capturing what people think about AI technologies, such as autonomous driving. (Etienne and Cova 2024; Awad et al. 2020), which is intended to inform product design or policies, or to help with "*identifying citizens' expectations.*" (Awad et al. 2020)

3. Setting. In reviewed papers, the research setting was situated online, with unknown, paid, or free compensation structures. We had to decide how to fairly compensate participants across countries (e.g., $10 per hour does not provide the same level of compensation in the US and Chile, due to differences in living costs and minimum wages).

4. Resources. While prior research has often failed to

consistently report the details, in our survey, the primary resources needed to conduct high-quality research were detailed, including the reliable channels used to recruit participants, digital consent and survey instruments, and the budget to cover research expenses (such as participant compensation, platform fees, researcher salaries, and costs of analysis software licenses). Additionally, hosting shareable versions of survey results for respondents and other communities also necessitates resources, as outlined below.

5. Communication. Communication between researchers and participants is often, unidirectional, with a transactional or task-oriented tone. There is no way for participants to know about the results without actively seeking and searching for results of their participation. In Prolific, the platform we used for our pilot survey, there is a feature that allows for messaging between participants and researchers. However, it is not intended for discussing the results of the survey; rather, it is designed for addressing any issues related to responses or payments. Recent work by Do et al. (2024) suggests that Prolific is like other gig work platforms, which similarly limit communications among workers and employers to transactional matters, and which discourage or prohibit communications from one gig worker to another.

6. Elicitation. Surveys are generally not structured to be longitudinal and are seldom designed to engage participants through multiple methods. In online surveys, participants are often given a limited amount of time to complete the survey. In particular, if they are professional participants, they may take multiple surveys in a short period of time (Hara et al. 2018). This scenario can lead to fatigue and diminish the level of meaningful engagement with the survey.

7. Conflict resolution. Survey data is usually aggregated and converted into numbers and statistics, representing a form of knowledge. Surveys are not typically seen as collaborative methods that involve ongoing discussions or conflict resolution processes. Consequently, they do not incorporate participant feedback or clarifications after data collection, nor do they include participants' input on the researchers' analytical summaries of results. Although our pilot survey featured multiple open-ended questions requiring thematic analysis, which necessitated reflection and collaboration among researchers, this approach still did not allow participants to contribute to the analysis.

8. Feedback. Although surveys may include a final question asking for participants' overall feedback or comments about the study, or provide an option to message or email the researcher, there is typically no mechanism for participants to access the results, offer feedback, or influence the final report or its implications. In our consent form, we provided an email address for participants to ask questions; however, no one reached out to us regarding the study.

9. Empowerment. The potential outcomes for participants and how the results might benefit or pose risks to them are often unclear. This information could be conveyed in the consent form, but many papers we reviewed lacked these details. In our pilot study, which focused mainly on understanding people's perceptions, there is no apparent direct benefit to the participants. While researchers may gain from publications that cover large segments of certain populations, and companies may benefit from extensive datasets reflecting attitudes toward their AI technologies, it remains unclear how participants are empowered through the collection and use of survey data.

10. Evaluation. Transparency regarding the details of methods and analyses, or the availability of full replication packages, is necessary to evaluate research practices. However, these elements were often missing from the papers we reviewed. Additionally, making data publicly available poses challenges from a researcher's perspective, as IRBs are frequently cautious about data-sharing practices, particularly concerning qualitative data. Open data are vulnerable to privacy violations (Borgesius, Gray, and Van Eechoud 2015; Crawford and Schultz 2014)—particularly in medical domains—and to cultural misappropriation, extraction, and exploitation (Duarte 2021; Karsgaard 2023; Palacios Abad et al. 2022). Our pilot survey's consent form and IRB application indicated that anonymized data would be made public to ensure transparency and facilitate future evaluation. We also planned to include the details of our qualitative coding in the replication package.

Current research practices and platforms have a significant impact on how surveys are assessed. To maintain anonymity, researchers and participants are often completely disconnected, leaving participants with no way to receive feedback on their contributions or learn about the results. Even if there were a way for participants to access the outcomes, traditional research publications would not be the most effective means of dissemination. These publications are often lengthy, use scientific jargon, require a high reading level, and are frequently behind paywalls. Additionally, the AI research community often emphasizes technological impacts—such as creating new datasets or enhancing efficiency—over the societal impacts of their research (Birhane et al. 2022b). Therefore, establishing a bidirectional communication and feedback loop in survey research necessitates a fundamental shift in how research impact is perceived and evaluated within the community.

Researchers have the discretion to use, report, or disregard parts of the results. What happens if researchers disagree with participants' responses, or if the outcomes do not align with the research goals or hypotheses? How much influence do participants have over the results and their eventual impact? Under current research practices, especially in online settings, participants have no power to affect or control the implications of their contributions. Such large-scale surveys are employed to *inform* rather than *empower*, treating survey takers as "*subjects in an investigation*" (Himmelreich 2023).

Representation Issues in Surveys

True Representation or an Illusion of Representation? Dataset attributes and annotation practices are known to introduce biases into AI, potentially resulting in poor representation of the perspectives of marginalized groups (Bergman et al. 2023; D'Ignazio and Klein 2020; Sambasivan et al. 2021; Lu et al. 2024). A model trained on a "representative" dataset might still under-perform when making decisions for marginalized communities (Aragon et al. 2022; Bergman et al. 2023). Our pilot survey, which involved 282 partic-

ipants, was methodologically acceptable, as we continued recruiting until data saturation—an acceptable practice in qualitative research (Francis et al. 2010; Guest, Bunce, and Johnson 2006). However, it does not fully encompass the entire spectrum of people in the studied countries, which are likely to be internally heterogeneous, potentially including marginalized or minoritized groups—a situation described by Trefzer et al. (2014, p. 1) as "*the global south within the global north.*" In our literature review, 14 out of 44 papers (32%) claimed to have a "nationally-representative" sample. The remaining papers typically included over 300 participants ($mean = 607$), aligning with the recommended sample size of 385 participants for a large population with 95% confidence and a 5% margin of error (SurveyMonkey 2024).

While survey methods can be effective for specific studies with targeted design and population, such as testing a new app feature, they may not be adequate for complex and impactful topics like AI, where results could affect large populations. Our work extends the discussion of representation issues to survey methods. We critique how surveys influence our research, development, and governance of AI, highlighting issues that extend beyond datasets and model development to encompass how AI should *be* and *behave*. Companies like Anthropic base their AI models' behavior on survey responses (Anthropic 2023), and Google glorifies public optimism about AI using their large-scale survey in partnership with Ipsos (Google 2024). Conversely, other surveys by Ipsos show increasing public nervousness about AI (Boyon 2023), with opinions varying depending on a country's level of economic development (Boyon 2022). In this line of research, Feffer et al. (2023) critique the Moral Machine experiment, which claims to represent global perspectives on the ethics of machines by encompassing 233 countries and 40 million decisions. However, an examination of the survey's geographical coverage reveals that it is largely dominated by contributions from North America, Europe, and some parts of South America, with minimal input from the African continent. These discrepancies in "representative" surveys and claims regarding "representative" results reinforce our argument about the true meaning of representation and underscore the need for our community to critically assess the current use of survey methods.

Cross-Cultural Surveys Considered Helpful or Harmful? The complexity increases when conducting surveys across different cultures and regions (Duarte 2021; Karsgaard 2023; Palacios Abad et al. 2022; Davani et al. 2024). In our pilot, we aimed to capture a broad range of perspectives from six continents without being fully immersed in the target populations and experiencing what it means to live in countries like Chile or South Africa. Or, conducting surveys exclusively in English in countries, where this is not the official language, limited our insights and was exclusionary by design. We received responses in languages other than English, indicating a desire to participate irrespective of English proficiency, driven by interest or compensation. Similarly, 11 out of 44 papers (25%) in our dataset studied countries without having an author affiliated with those countries.

The literature review also validates prior concerns about research focusing predominantly on Western populations (Septiandri et al. 2023)—only six papers (14%) featured authors from the Global South. This lack of representation could create an echo chamber effect, where the voices of certain populations dominate the discourse in AI (including responsible AI, safety of AI, and ethics of using AI), potentially eclipsing perspectives from underrepresented regions or populations (Duarte 2021; Trefzer et al. 2014). While the intention to include more countries might be well-meaning, the way *how* this inclusivity is approached remains a topic that requires attention and care. Without meaningfully engaging with the target populations, conducting cross-cultural studies might be motivated by the desire for a large sample size rather than an understanding of cross-cultural differences, potentially causing harm rather than benefit due to misinterpretations and a lack of contextual understanding, including different conceptions of the meaning of *consent* based on cultural roles of decision-makers and/or histories of exploitation (Munteanu and Sadownik 2019).

Another concern is the use of standardized questions across different cultures. Employing a uniform set of questions, even with thorough translations, suggests that we, as researchers, have not adequately adapted our inquiries to specific cultures and communities. This approach risks missing key cultural insights and misinterpreting responses rooted in specific cultural contexts. While standard questions aid comparative analysis, the extent to which they overlook contextual nuances and local values is an open question. This highlights the need not only for localization of study materials but also for more flexible and culturally-sensitive research methodologies in general, especially when studying diverse populations. When we sought feedback on our pilot survey, we found that people had diverse interpretations of the questions. For example, the term *story* can have various meanings in Japanese, and a specific demographic question about ethnic origins might be crucial for indigenous populations in Australia. However, as researchers without lived experiences in the target countries, we would not have been able to understand these cultural details without conducting walk-through interviews with representatives from those countries—a practice that was absent in 85% of the papers we reviewed.

Value Tensions in Surveys: Heuristic Questions

In this paper, we discussed many positions and concerns, which often did not converge toward simple outcomes. Survey research in AI vexes us with many choices and decisions. We summarize the major issues as tensions that are currently unresolved. In view of the many reasons and motivations for using surveys in AI research, we propose that "*heuristic questions*" (Muller 1997) may be more valuable than advice. Asking "*big questions*" (Beck and Stolterman 2017, ms. p. 1), (Reiser et al. 2017; Schaeffer and Presser 2003) or "*the right question*" (Mao et al. 2019, ms. p. 1) has been deemed valuable when facing newly-problematic research challenges (Phillips, Watkins, and Hammer 2018). Thus, we propose to use the following reflexive questions (e.g., inspired by Laufer et al. 2022; Septiandri et al. 2023) when planning survey-based studies about AI:

(A) Breadth and Depth:

- **Standardization and Customization.** Do we attempt to standardize certain survey content through invariant questions addressed to all persons in all geographical locations and cultural backgrounds (Ornstein 2013)? How can we address the tendency for a survey to primarily reflect the cultural perspective of the Global North and its associated dominance (Septiandri et al. 2023)?
- **Languages.** Do we adapt the survey for regional languages (Kelley et al. 2021)? When does openness and accommodation shade into cultural differences in inquiry, leading to incommensurable outcomes?
- **Sampling.** Survey research is often constrained by time and budget. In settings with diverse cultures, how great an effort should be spent on recruiting a balanced or weighted sample across cultures (Selwyn et al. 2020; Zhang and Dafoe 2020)? Where is the "*stopping point*", and is this a question that requires members of the survey population to help answer? How much sampling stratification is needed? Who defines cultural boundaries?

(B) Manual and Automated Approaches. If we use generative AI in question generation, do we risk a bland, so-called "*universal*", tone that reflects the worldview of the AI-provider (Paxton 2023)? Or if we rely on humans in the research process, e.g., in data collection and analysis, how can we account for the limited views of research teams and their biases? What are the specific choices related to use of AI that need to be disclosed as part of consent (e.g., Wilcox, Brewer, and Diaz 2023; Andreotta, Kirkham, and Rizzi 2022; Gomez Ortega et al. 2023) and transparency in publication (e.g., Wacharamanotham et al. 2020; Hosseini, Resnik, and Holmes 2023)?

(C) Who and What Influences Survey Designs? Survey design is often considered to be the domain of specialists (e.g., Fink 2003; Spector 2013). While it is true that the design of questions and response-scales requires professional knowledge, the selection of topics may be informed by members of stakeholder groups or affected classes (Baeza-Yates 2018; Nicoletti and Bass 2023; Alvarado Garcia et al. 2021; Bird 2020; Kwet 2019). Mindful of the different meanings of "*participation*" (e.g., Hansen, Fourie, and Meyer 2021; Muller and Kuhn 1993; Schuler and Namioka 1993; Simonsen and Robertson 2012), we ask: What are the opportunities for involving community members or leaders in participatory or co-design processes to select and refine the survey topics and the way they are framed in questions (e.g., Arnstein 1969; Dugan et al. 2021; Flicker et al. 2010; Schulz et al. 2005)?

(D) Trust and Research Engagement. Trust among many communities that might participate in survey research has been eroded due to past mistreatment in research more broadly, along with experiences of racism and various forms of prejudice by different research institutions (Scharff et al. 2010). This historical context influences a community's decisions regarding whether and how to engage in such activities. (Wilcox et al. 2023). To what extent should survey design, deployment, and reporting directly address issues of trust? What measures (e.g., establishing publication standards that include sharing data with participants) should we expect our community to implement to ensure responsible and transparent research practices?

(E) Mixed Methods and Balanced Inquiry. As discussed in Section 1, surveys tend to decontextualize responses and isolate respondents (Ornstein 2013). Is it feasible to integrate large-scale quantitative survey methods with smaller-scale, rigorous qualitative analyses involving strategically selected groups of informants (e.g., Baumer et al. 2017; Greenberg and Buxton 2008; Muller et al. 2016)?

(F) Transparency and Research Practices. We should discuss whether and how to establish consistent reporting methods for surveys on AI topics. For instance, what transparency artifacts (e.g., Crisan et al. 2022; Mitchell et al. 2019; Chmielinski et al. 2022; Díaz et al. 2022; Rostamzadeh et al. 2022; Srinivasan et al. 2021) might inspire new forms of methodological transparency? How should we determine which types of data to include in these artifacts? Selection criteria to consider may include the contingent and potentially sensitive nature of AI topics addressed in surveys, the diverse communities that survey research serves, the fact that cross-cultural perspectives may require varied forms of transparency, the need for participant anonymization, and the interpretive traditions associated with qualitative analyses (Soden, Toombs, and Thomas 2024).

(G) Researcher and Participant Empowerment. Researchers should consider early in their study what survey participants stand to gain from the research (Oldendick 2012; Jamieson, Govaart, and Pownall 2023). This consideration becomes particularly important when studying hard-to-reach populations or when using public funds (or considering conflicts of interest when using private funds). Adopting a reflexive, critical approach to the implications of their research can significantly benefit researchers in understanding and improving the value of their work for participants. In reflecting, we ask: Are there alternative avenues to improve the value exchange for the populations being studied?

6 Conclusion

In this paper, we combine epistemic approaches grounded in critical reflexivity with a systemic literature review to examine the state of large-scale surveys in AI scholarship. The study reveals a range of performative and misleading practices with a method that has garnered adoption, informing research, shaping publicly-facing narratives, and justifying trajectories in AI development—highlighting the need for urgent intervention. The stakes are high in AI research, and some of these issues cannot be adequately addressed or dismissed by merely tucking challenges within the limitations section of research papers. As such, we aim to spark reflexive engagement with research processes that shape how surveys could be used responsibly and offer a list of heuristic questions to prompt more thorough acknowledgments of bias and subjectivity.

7 Research Ethics and Social Impact

Ethical Considerations Statement

In addition to the ethics considerations described in our paper body, our pilot survey obtained approval from the Research Ethics Office at King's College London. We implemented strict measures to ensure the confidentiality, anonymity, and privacy of our participants. No personally-identifiable information was collected, and participation was voluntary and anonymous. Participants were provided with an informed consent form in English, detailing the study's purpose and the intended use of the data collected.

Researcher Positionality Statement

The authors come from varied research backgrounds that shape their perspectives. The study was funded by an academic institution in the Global North, and funding for the study was restricted to respondent incentives and vendor survey services. Authors were employed by their institutions and were not explicitly paid to conduct this research. One author is employed by an academic institution, and one is employed at a non-governmental organization. Four authors are employed in industry research roles, though this study was not part of their company research. The research team have experiences living in two of the six countries surveyed (United Kingdom and United States). All six authors have extensive experience with survey methods, with four having experience with international and cross-cultural survey approaches. Authors were born in, currently live in, or had previously lived in, nine different countries collectively. Our race/ethnicity is collectively White (European) ($n = 3$), Middle Eastern ($n = 2$), and Afro-Caribbean ($n = 1$). All authors identify as having some experience with marginalization in computing, either through years of conducting computing research with marginalized groups or as members of a marginalized group themselves.

Our positionality is influenced by our backgrounds and experiences; as researchers trained and working in predominantly Western institutions, we acknowledge that complementary scholarship related to our research questions is needed, to further the understandings presented in this paper. Our positionality has also influenced the subjectivity inherent in framing our paper approach, research questions, study pilot design, literature review, and data interpretation and analysis, as we elaborate on throughout the paper.

Adverse Impact Statement

Our research aims to promote critical thinking within the AIES community about survey methods in AI, but they could be interpreted as an outright dismissal of these methods without full engagement with the nuances we present in our paper. Our intention is not to entirely discourage the use of surveys in AI and responsible AI research. Instead, our goal is to foster thoughtful and critical engagement within the AIES community to develop perspectives on the principles associated with the *who, what, when, where, why,* and *how* of human survey methods in AI research. Finally, we do not intend to suggest that designing the "*perfect*" survey will address the systemic issues that surround survey research in AI. Power relationships and broader structural concerns cannot be resolved simply by a survey designed and deployed in ways that uphold our community's principles; though such a survey may meet specific research goals, it would not address issues surrounding its application (e.g., the actual impact that survey results have on the practices of powerful institutions). We encourage future research to continue utilizing reflexive approaches and to further develop best practices and standards for conducting high-quality, inclusive, reliable, and impactful survey research in AI.

References

Ada Lovelace Institute. 2021. Participatory data stewardship: A framework for involving people in the use of data.

Ada Lovelace Institute and Alan Turing Institute. 2022. How do people feel about AI?

Agyemang, C.; Addo, J.; Bhopal, R.; de Graft Aikins, A.; and Stronks, K. 2009. Cardiovascular disease, diabetes and established risk factors among populations of sub-Saharan African descent in Europe: a literature review. *Globalization and health*.

AI.gov. 2023a. National Artificial Intelligence Advisory Committee (NAIAC).

AI.gov. 2023b. Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem — An Implementation Plan for a National Artificial Intelligence Research Resource.

Albert, B.; Tullis, T.; and Tedesco, D. 2009. *Beyond the usability lab: Conducting large-scale online user experience studies*. Morgan Kaufmann.

Alexander, L.; and Moore, M. 2021. Deontological Ethics. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2021 edition.

Ali, S. M. 2016. A Brief Introduction to Decolonial Computing. *XRDS*.

Alkhathlan, M.; Cachel, K.; Shrestha, H.; Harrison, L.; and Rundensteiner, E. 2024. Balancing Act: Evaluating People's Perceptions of Fair Ranking Metrics. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*.

Alvarado Garcia, A.; Maestre, J. F.; Barcham, M.; Iriarte, M.; Wong-Villacres, M.; Lemus, O. A.; Dudani, P.; Reynolds-Cuéllar, P.; Wang, R.; and Cerratto Pargman, T. 2021. Decolonial Pathways: Our Manifesto for a Decolonizing Agenda in HCI Research and Design. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM.

Amazon. 2023. Amazon Mechanical Turk.

Andreotta, A. J.; Kirkham, N.; and Rizzi, M. 2022. AI, big data, and the future of consent. *Ai & Society*.

Angal, J.; Petersen, J. M.; Tobacco, D.; Elliott, A. J.; in SIDS, P. A.; and Network, S. 2016. Ethics review for a multi-site project involving Tribal Nations in the Northern Plains. *Journal of Empirical Research on Human Research Ethics*.

Anthropic. 2023. Collective Constitutional AI: Aligning a Language Model with Public Input.

Aragon, C.; Guha, S.; Kogan, M.; Muller, M.; and Neff, G. 2022. *Human-centered data science: An introduction*. MIT Press.

Arai, K.; and Matsumoto, M. 2023. Public perceptions of autonomous lethal weapons systems. *AI and Ethics*.

Arcaya, M. C.; Schnake-Mahl, A.; Binet, A.; Simpson, S.; Church, M. S.; Gavin, V.; Coleman, B.; Levine, S.; Nielsen, A.; Carroll, L.; et al. 2018. Community change and resident needs: designing a participatory action research study in metropolitan Boston. *Health & place*.

Arnstein, S. R. 1969. A ladder of citizen participation. *Journal of the American Institute of planners*.

Auger, J. 2013. Speculative design: crafting the speculation. *Digital Creativity*.

Awad, E.; Dsouza, S.; Bonnefon, J.-F.; Shariff, A.; and Rahwan, I. 2020. Crowdsourcing moral machines. *Commun. ACM*.

Awad, E.; Dsouza, S.; Kim, R.; Schulz, J.; Henrich, J.; Shariff, A.; Bonnefon, J.-F.; and Rahwan, I. 2018. The Moral Machine experiment. *Nature*.

Baeza-Yates, R. 2018. Bias on the Web. *Commun. ACM*.

Bao, L.; Krause, N. M.; Calice, M. N.; Scheufele, D. A.; Wirz, C. D.; Brossard, D.; Newman, T. P.; and Xenos, M. A. 2022. Whose AI? How different publics think about AI and its social impacts. *Computers in Human Behavior*.

Bartneck, C.; Yogeeswaran, K.; and Sibley, C. G. 2023. Personality and demographic correlates of support for regulating artificial intelligence. *AI and Ethics*.

Baumer, E. P.; Mimno, D.; Guha, S.; Quan, E.; and Gay, G. K. 2017. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology*.

Beck, J.; and Stolterman, E. 2017. Reviewing the big questions literature; or, should HCI have big questions? In *Proceedings of the 2017 Conference on Designing Interactive Systems*.

Beck, U. 1992. *Risk society: Towards a new modernity*. sage.

Beck, U. 1996. World risk society as cosmopolitan society? Ecological questions in a framework of manufactured uncertainties. *Theory, culture & society*.

Bergman, A. S.; Hendricks, L. A.; Rauh, M.; Wu, B.; Agnew, W.; Kunesch, M.; Duan, I.; Gabriel, I.; and Isaac, W. 2023. Representation in AI Evaluations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM.

Bernstein, P. L.; and Bernstein, P. L. 1996. *Against the gods: The remarkable story of risk*. Wiley New York.

Berry-James, R. M.; Gooden, S. T.; and Johnson III, R. G. 2020. Civil Rights, Social Equity, and Census 2020. *Public Administration Review*.

Bharadiya, J. 2023. Artificial Intelligence in Transportation Systems A Critical Review. *American Journal of Computing and Engineering*.

Bhopal, R.; Vettini, A.; Hunt, S.; Wiebe, S.; Hanna, L.; and Amos, A. 2004. Review of prevalence data in, and evaluation of methods for cross cultural adaptation of, UK surveys on tobacco and alcohol in ethnic minority groups. *BMJ*.

Bird, S. 2020. Decolonising Speech and Language Technology. In Scott, D.; Bel, N.; and Zong, C., eds., *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics.

Birhane, A.; Isaac, W.; Prabhakaran, V.; Diaz, M.; Elish, M. C.; Gabriel, I.; and Mohamed, S. 2022a. Power to the People? Opportunities and Challenges for Participatory AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM.

Birhane, A.; Kalluri, P.; Card, D.; Agnew, W.; Dotan, R.; and Bao, M. 2022b. The Values Encoded in Machine Learning Research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM.

Blodgett, S. L.; and Madaio, M. 2021. Risks of AI Foundation Models in Education.

Boholm, Å. 1996. Risk perception and social anthropology: Critique of cultural theory. *Ethnos*.

Bondi, E.; Xu, L.; Acosta-Navas, D.; and Killian, J. A. 2021. Envisioning Communities: A Participatory Approach Towards AI for Social Good. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery.

Borgesius, F. Z.; Gray, J.; and Van Eechoud, M. 2015. Open data, privacy, and fair information principles: Towards a balancing framework. *Berkeley Technology Law Journal*.

Bourdieu, P. 1990. *In other words: Essays toward a reflexive sociology*. Stanford University Press.

Bourdieu, P.; and Wacquant, L. J. 1992. *An invitation to reflexive sociology*. University of Chicago press.

Boyon, N. 2022. Opinions about AI vary depending on countries' level of economic development.

Boyon, N. 2023. AI is making the world more nervous.

Brown, M. 2023. How to Run Surveys at Every Stage of the Design Cycle.

Byun, C.; Vasicek, P.; and Seppi, K. 2023. Dispensing with Humans in Human-Computer Interaction Research. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM.

Cannell, C. F.; Marquis, K. H.; Laurent, A.; et al. 1977. *A summary of studies of interviewing methodology*. US Government Printing Office, Washington, DC 20402.

Cave, S.; Coughlan, K.; and Dihal, K. 2019. "Scary Robots": Examining Public Responses to AI. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM.

Center, P. R. 2024. Writing Survey Questions.

Chasalow, K.; and Levy, K. 2021. Representativeness in Statistics, Politics, and Machine Learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM.

Chilisa, B. 2019. *Indigenous research methodologies*. Sage publications.

Chmielinski, K. S.; Newman, S.; Taylor, M.; Joseph, J.; Thomas, K.; Yurkofsky, J.; and Qiu, Y. C. 2022. The Dataset Nutrition Label (2nd Gen): Leveraging Context to Mitigate Harms in Artificial Intelligence.

Choi, B. C. K.; and Pak, A. W. P. 2005. A catalog of biases in questionnaires. *Prev. Chronic Dis.*

Colton, D.; and Covert, R. W. 2007. *Designing and constructing instruments for social research and evaluation*. John Wiley & Sons.

Comber, R.; Bardzell, S.; Bardzell, J.; Hazas, M.; and Muller, M. 2020. Announcing a new CHI subcommittee: critical and sustainable computing. *Interactions*.

Constantinides, M.; Bogucka, E.; Quercia, D.; Kallio, S.; and Tahaei, M. 2024a. A Method for Generating Dynamic Responsible AI Guidelines for Collaborative Action. *Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*.

Constantinides, M.; Tahaei, M.; Quercia, D.; Stumpf, S.; Madaio, M.; Kennedy, S.; Wilcox, L.; Vitak, J.; Cramer, H.; Bogucka, E. P.; Baeza-Yates, R.; Luger, E.; Holbrook, J.; Muller, M.; Blumenfeld, I. G.; and Pistilli, G. 2024b. Implications of Regulations on the Use of AI and Generative AI for Human-Centered Responsible Artificial Intelligence. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*. ACM.

Converse, J. M.; and Presser, S. 1986. *Survey questions: Handcrafting the standardized questionnaire*. Sage.

Cooper, N.; Horne, T.; Hayes, G. R.; Heldreth, C.; Lahav, M.; Holbrook, J.; and Wilcox, L. 2022. A Systematic Review and Thematic Analysis of Community-Collaborative Approaches to Computing Research. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM.

Crabtree, A.; Rodden, T.; Tolmie, P.; and Button, G. 2009. Ethnography Considered Harmful. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM.

Crawford, K.; and Schultz, J. 2014. Big data and due process: Toward a framework to redress predictive privacy harms. *BCL Rev.*

Crisan, A.; Drouhard, M.; Vig, J.; and Rajani, N. 2022. Interactive Model Cards: A Human-Centered Approach to Model Documentation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM.

Cunliffe, A. L. 2004. On Becoming a Critically Reflexive Practitioner. *Journal of Management Education*.

Davani, A.; Díaz, M.; Baker, D.; and Prabhakaran, V. 2024. Disentangling Perceptions of Offensiveness: Cultural and Moral Correlates. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*.

Delgado, F.; Yang, S.; Madaio, M.; and Yang, Q. 2023. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM.

Dennler, N.; Ovalle, A.; Singh, A.; Soldaini, L.; Subramonian, A.; Tu, H.; Agnew, W.; Ghosh, A.; Yee, K.; Peradejordi, I. F.; Talat, Z.; Russo, M.; and Pinhal, J. D. J. D. P. 2023. Bound by the Bounty: Collaboratively Shaping Evaluation Processes for Queer AI Harms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. ACM.

Denzin, N. K.; Lincoln, Y. S.; and Smith, L. T. 2008. *Handbook of critical and indigenous methodologies*. Sage.

Díaz, M.; Kivlichan, I.; Rosen, R.; Baker, D.; Amironesei, R.; Prabhakaran, V.; and Denton, E. 2022. CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery.

D'Ignazio, C.; and Klein, L. F. 2020. *Data feminism*. MIT press.

Dijkstra, E. W. 1968. Letters to the Editor: Go to Statement Considered Harmful. *Commun. ACM*.

Dillman, D. A.; Tortora, R. D.; and Bowker, D. 1998. Principles for constructing web surveys. In *Joint Meetings of the American Statistical Association*.

Dillman, D. A.; et al. 1978. *Mail and telephone surveys: The total design method*. Wiley New York.

Do, K.; De Los Santos, M.; Muller, M.; and Savage, S. 2024. GigSousveillance: Designing Gig Worker Centric Sousveillance Tools. In *ACM CHI Conference on Human Factors in Computing Systems*.

Doherty, K.; and Doherty, G. 2018. Engagement in HCI: Conception, Theory and Measurement. *ACM Comput. Surv.*

Douglas, B. D.; Ewell, P. J.; and Brauer, M. 2023. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *Plos one*.

Douglas, M.; and Wildavsky, A. 1983. *Risk and culture: An essay on the selection of technological and environmental dangers*. Univ of California Press.

Dourish, P.; Lawrence, C.; Leong, T. W.; and Wadley, G. 2020. On Being Iterated: The Affective Demands of Design Participation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM.

Dreber, A.; and Johannesson, M. 2019. Statistical significance and the replication crisis in the social sciences. In *Oxford research encyclopedia of economics and finance*.

Duarte, M. E. 2021. Native and indigenous women's cyberdefense of lands and peoples. *Networked Feminisms: Activist Assemblies and Digital Practices*.

Dugan, A. G.; Namazi, S.; Cavallari, J. M.; Rinker, R. D.; Preston, J. C.; Steele, V. L.; and Cherniack, M. G. 2021. Participatory survey design of a workforce health needs assessment for correctional supervisors. *American journal of industrial medicine*.

Echtler, F.; and Häußler, M. 2018. Open Source, Open Science, and the Replication Crisis in HCI. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM.

Epstein, R.; Bordyug, M.; Chen, Y.-H.; Chen, Y.; Ginther, A.; Kirkish, G.; and Stead, H. 2023. Toward the search for the perfect blade runner: a large-scale, international assessment of a test that screens for "humanness sensitivity". *AI & SOCIETY*.

Epstein, Y.; Yanovich, R.; Moran, D. S.; and Heled, Y. 2013. Physiological employment standards IV: integration of women in combat units physiological and medical considerations. *European journal of applied physiology*.

Etienne, H.; and Cova, F. 2024. The more they think, the less they want: studying people's attitudes about autonomous vehicles could also contribute to shaping them. *AI and Ethics*.

Fast, E.; and Horvitz, E. 2017. Long-Term Trends in the Public Perception of Artificial Intelligence. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI Press.

Feffer, M.; Skirpan, M.; Lipton, Z.; and Heidari, H. 2023. From Preference Elicitation to Participatory ML: A Critical Survey & Guidelines for Future Research. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. ACM.

Feinberg, M. 2017. A Design Perspective on Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM.

Feinberg, M. 2022. *Everyday Adventures with Unruly Data*. MIT Press.

Finance Center for South-South Cooperation. 2024. Global South Countries.

Fink, A. 2003. *How to design survey studies*. Sage.

Flicker, S.; Guta, A.; Larkin, J.; Flynn, S.; Fridkin, A.; Travers, R.; Pole, J. D.; and Layne, C. 2010. Survey design from the ground up: Collaboratively creating the Toronto Teen Survey. *Health Promotion Practice*.

Fowler, F. J. 2013. *Survey Research Methods*. SAGE Publications.

Fowler Jr, F. J.; and Mangione, T. W. 1990. *Standardized survey interviewing: Minimizing interviewer-related error*. Sage.

Francis, J. J.; Johnston, M.; Robertson, C.; Glidewell, L.; Entwistle, V.; Eccles, M. P.; and Grimshaw, J. M. 2010. What is an adequate sample size? Operationalising data saturation for theory-based interview studies. *Psychology & Health*.

Franken, S.; and Wattenberg, M. 2019. The impact of AI on employment and organisation in the industrial working environment of the future. In *ECIAIR 2019 European Conference on the Impact of Artificial Intelligence and Robotics*. Academic Conferences and publishing limited.

Freese, J.; and Peterson, D. 2017. Replication in social science. *Annual Review of Sociology*.

Gilovich, T.; Keltner, D.; and Nisbett, R. E. 2006. Being a member of a stigmatized group: stereotype threat. *Gilovich, Thomas; Keltner, Dacher; Nisbett, Richard E., Social psychology, New York: WW Norton*.

Gomez Ortega, A.; Bourgeois, J.; Hutiri, W. T.; and Kortuem, G. 2023. Beyond data transactions: a framework for meaningfully informed data donation. *AI & SOCIETY*.

Google. 2024. Global Study Shows Optimism About AI's Potential.

Greenberg, S.; and Buxton, B. 2008. Usability Evaluation Considered Harmful (Some of the Time). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM.

Groves, L.; Peppin, A.; Strait, A.; and Brennan, J. 2023. Going Public: The Role of Public Participation Approaches in Commercial AI Labs. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM.

Groves, R. M.; Fowler Jr, F. J.; Couper, M. P.; Lepkowski, J. M.; Singer, E.; and Tourangeau, R. 2009. *Survey methodology*. John Wiley & Sons.

Groves, R. M.; Fowler Jr, F. J.; Couper, M. P.; Lepkowski, J. M.; Singer, E.; and Tourangeau, R. 2011. *Survey methodology*. John Wiley & Sons.

Guest, G.; Bunce, A.; and Johnson, L. 2006. How Many Interviews Are Enough?: An Experiment with Data Saturation and Variability. *Field Methods*.

Hansen, P.; Fourie, I.; and Meyer, A. 2021. *Third space, information sharing, and participatory design*. Springer.

Hara, K.; Adams, A.; Milland, K.; Savage, S.; Callison-Burch, C.; and Bigham, J. P. 2018. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM.

Havens, L.; Terras, M.; Bach, B.; and Alex, B. 2020. Situated Data, Situated Systems: A Methodology to Engage with Power Relations in Natural Language Processing Research. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics.

Hayes, G. R. 2014. Knowing by doing: action research as an approach to HCI. In *Ways of Knowing in HCI*. Springer.

Hertzberg, A.; Liberti, J. M.; and Paravisini, D. 2010. Information and incentives inside the firm: Evidence from loan officer rotation. *The Journal of Finance*.

Himmelreich, J. 2023. Against "Democratizing AI". *AI & SOCIETY*.

Holdren, J. P.; Sunstein, C. R.; and Siddiqui, I. A. 2011. *Principles for regulation and oversight of emerging technologies*. Office of Science and Technology Policy.

Hosseini, M.; Resnik, D. B.; and Holmes, K. 2023. The ethics of disclosing the use of artificial intelligence tools in writing scholarly manuscripts. *Research Ethics*.

Houser, K. A. 2019. Can AI solve the diversity problem in the tech industry: Mitigating noise and bias in employment decision-making. *Stan. Tech. L. Rev.*

Huang, S.; Siddarth, D.; Lovitt, L.; Liao, T. I.; Durmus, E.; Tamkin, A.; and Ganguli, D. 2024. Collective Constitutional AI: Aligning a Language Model with Public Input.

In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*.

Hursthouse, R.; and Pettigrove, G. 2023. Virtue Ethics. In Zalta, E. N.; and Nodelman, U., eds., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2023 edition.

Ikkatai, Y.; Hartwig, T.; Takanashi, N.; and Yokoyama, H. M. 2023. Segmentation of ethics, legal, and social issues (ELSI) related to AI in Japan, the United States, and Germany. *AI and Ethics*.

Iliadis, A.; and Russo, F. 2016. Critical data studies: An introduction. *Big Data & Society*.

Irani, L.; Vertesi, J.; Dourish, P.; Philip, K.; and Grinter, R. E. 2010. Postcolonial Computing: A Lens on Design and Development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM.

Jakesch, M.; Buçinca, Z.; Amershi, S.; and Olteanu, A. 2022. How Different Groups Prioritize Ethical Values for Responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM.

Jamieson, M. K.; Govaart, G. H.; and Pownall, M. 2023. Reflexivity in quantitative research: A rationale and beginner's guide. *Social and Personality Psychology Compass*.

Kapania, S.; Siy, O.; Clapper, G.; SP, A. M.; and Sambasivan, N. 2022. "Because AI is 100% Right and Safe": User Attitudes and Sources of AI Authority in India. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM.

Karsgaard, C. 2023. New Visions for Anti-colonial Digital Methods. In *Instagram as Public Pedagogy: Online Activism and the Trans Mountain Pipeline*. Springer.

Kaufmann, N.; Schulze, T.; and Veit, D. 2011. More than fun and money. worker motivation in crowdsourcing–a study on mechanical turk.

Kelley, K.; Clark, B.; Brown, V.; and Sitzia, J. 2003. Good practice in the conduct and reporting of survey research. *International Journal for Quality in health care*.

Kelley, P. G.; Yang, Y.; Heldreth, C.; Moessner, C.; Sedley, A.; Kramm, A.; Newman, D. T.; and Woodruff, A. 2021. Exciting, Useful, Worrying, Futuristic: Public Perception of Artificial Intelligence in 8 Countries. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM.

Kieslich, K.; and Lünich, M. 2024. Regulating AI-Based Remote Biometric Identification. Investigating the Public Demand for Bans, Audits, and Public Database Registrations. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*.

Kitchin, R.; and Lauriault, T. 2014. Towards critical data studies: Charting and unpacking data assemblages and their work.

Ko, A. J.; Beitlers, A.; Wortzman, B.; Davidson, M.; Oleson, A.; Kirdani-Ryan, M.; Druga, S.; and Everson, J. 2023. *Critically Conscious Computing: Methods for Secondary Education*.

Kovach, M. 2021. *Indigenous methodologies: Characteristics, conversations, and contexts*. University of Toronto press.

Kramer, M. F.; Schaich Borg, J.; Conitzer, V.; and Sinnott-Armstrong, W. 2018. When Do People Want AI to Make Decisions? In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery.

Krosnick, J. A. 1999. Survey Research. *Annual Review of Psychology*.

Kuhn, N. S.; Parker, M.; and Lefthand-Begay, C. 2020. Indigenous research ethics requirements: an examination of six tribal institutional review board applications and processes in the United States. *Journal of empirical research on human research ethics*.

Kwet, M. 2019. Digital colonialism: US empire and the new imperialism in the Global South. *Race & Class*.

Laufer, B.; Jain, S.; Cooper, A. F.; Kleinberg, J.; and Heidari, H. 2022. Four Years of FAccT: A Reflexive, Mixed-Methods Analysis of Research Contributions, Shortcomings, and Future Prospects. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM.

Lee, J. W.; Jones, P. S.; Mineyama, Y.; and Zhang, X. E. 2002. Cultural differences in responses to a Likert scale. *Research in nursing & health*.

Linxen, S.; Sturm, C.; Brühlmann, F.; Cassau, V.; Opwis, K.; and Reinecke, K. 2021. How WEIRD is CHI? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM.

Loefflad, C.; and Grossklags, J. 2024. How the Types of Consequences in Social Scoring Systems Shape People's Perceptions and Behavioral Reactions. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*.

Lovett, M.; Bajaba, S.; Lovett, M.; and Simmering, M. J. 2018. Data quality from crowdsourced surveys: A mixed method inquiry into perceptions of Amazon's Mechanical Turk Masters. *Applied Psychology*.

Lu, A. J.; Moy, C.; Ackerman, M. S.; Morenoff, J.; and Dillahunt, T. R. 2024. Perceptions of Policing Surveillance Technologies in Detroit: Moving Beyond "Better than Nothing". In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery.

Lünich, M.; and Keller, B. 2024. Explainable Artificial Intelligence for Academic Performance Prediction. An Experimental Study on the Impact of Accuracy and Simplicity of Decision Trees on Causability and Fairness Perceptions. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*.

Mann, M.; and Daly, A. 2019. (Big) Data and the North-in-South: Australia's Informational Imperialism and Digital Colonialism. *Television & New Media*.

Mao, Y.; Wang, D.; Muller, M.; Varshney, K. R.; Baldini, I.; Dugan, C.; and Mojsilović, A. 2019. How data scientistswork together with domain experts in scientific collaborations: To find the right answer or to ask the right question? *Proceedings of the ACM on Human-Computer Interaction*.

Marenko, B. 2018. Futurecrafting. A speculative method for an imaginative AI. *AAAI Spring Symposium Series*.

McKee, K. R. 2023. Human participants in AI research: Ethics and transparency in practice.

Miceli, M.; Yang, T.; Naudts, L.; Schuessler, M.; Serbanescu, D.; and Hanna, A. 2021. Documenting Computer Vision Datasets: An Invitation to Reflexive Data Practices. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM.

Midena, D.; and Yeo, R. 2022. Towards a history of the questionnaire.

Mills, C. W. 2023. The sociological imagination. In *Social Work*. Routledge.

Mir, G.; Salway, S.; Kai, J.; Karlsen, S.; Bhopal, R.; Ellison, G. T.; and Sheikh, A. 2012. Principles for research on ethnicity and health: the Leeds Consensus Statement. *European Journal of Public Health*.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM.

Moloi, T.; and Marwala, T. 2021. *Artificial Intelligence and the Changing Nature of Corporations: How Technologies Shape Strategy and Operations*. Springer Nature.

Morley, J.; Machado, C. C.; Burr, C.; Cowls, J.; Joshi, I.; Taddeo, M.; and Floridi, L. 2020. The ethics of AI in health care: a mapping review. *Social Science & Medicine*.

Muller, M.; Guha, S.; Baumer, E. P.; Mimno, D.; and Shami, N. S. 2016. Machine learning and grounded theory method: convergence, divergence, and combination. In *Proceedings of the 2016 ACM International Conference on Supporting Group Work*.

Muller, M.; Lange, I.; Wang, D.; Piorkowski, D.; Tsay, J.; Liao, Q. V.; Dugan, C.; and Erickson, T. 2019. How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM.

Muller, M.; and Strohmayer, A. 2022. Forgetting Practices in the Data Sciences. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM.

Muller, M.; Wolf, C. T.; Andres, J.; Desmond, M.; Joshi, N. N.; Ashktorab, Z.; Sharma, A.; Brimijoin, K.; Pan, Q.; Duesterwald, E.; and Dugan, C. 2021. Designing Ground Truth and the Social Life of Labels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM.

Muller, M. J. 1997. Ethnocritical heuristics for reflecting on work with users and other interested parties. In *Computers and design in context*.

Muller, M. J.; and Kuhn, S. 1993. Participatory design. *Communications of the ACM*.

Munteanu, C.; and Sadownik, S. 2019. Field Studies of Interactive Technologies for Marginalized Users: A Canadian Ethics Policy Perspective. *Ageing and Digital Technology: Designing and Evaluating Emerging Technologies for Older Adults*.

Müller, H.; Sedley, A.; and Ferrall-Nunge, E. 2014. *Survey Research in HCI*. Springer. In J. Olson & W. Kellogg (Eds.).

Nazroo, J.; Jackson, J.; Karlsen, S.; and Torres, M. 2007. The Black diaspora and health inequalities in the US and England: does where you go and how you get there make a difference? *Sociology of Health & Illness*.

Nicoletti, L.; and Bass, D. 2023. Humans Are Biased. Generative AI Is Even Worse.

Nierkens, V.; de Vries, H.; and Stronks, K. 2006. Smoking in immigrants: do socioeconomic gradients follow the pattern expected from the tobacco epidemic? *Tobacco control*.

NIST. 2023. AI Risk Management Framework.

Norton, I. M.; and Manson, S. M. 1996. Research in American Indian and Alaska Native communities: navigating the cultural universe of values and process. *Journal of consulting and clinical psychology*.

Oldendick, R. W. 2012. Survey research ethics. *Handbook of survey methodology for the social sciences*.

OpenAI. 2023. Official ChatGPT survey - Shape the future of ChatGPT.

Orimadegun, A. 2020. Protocol and Researcher's Relationship with Institutional Review Board. *African Journal of Biomedical Research*.

Ornstein, M. 2013. *A Companion to Survey Research*. SAGE Publications.

Othman, K. 2023. Understanding how moral decisions are affected by accidents of autonomous vehicles, prior knowledge, and perspective-taking: a continental analysis of a global survey. *AI and Ethics*.

Palacios Abad, B.; Belding, E.; Vigil-Hayes, M.; and Zegura, E. 2022. Note: Towards Community-Empowered Network Data Action. In *Proceedings of the 5th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*.

Patel, D. I.; Stevens, K. R.; Puga, F.; et al. 2013. Variations in institutional review board approval in the implementation of an improvement research study. *Nursing Research and Practice*.

Paxton, A. 2023. What Is It Like to Sound Like a Bot? *Discourse and Writing/Rédactologie*.

PE, E.; and MD, G. 2019. Inconsistencies in institutional review board decisions: A proposal to regulate the decision-making process. *Bratislava Medical Journal/Bratislavske Lekarske Listy*.

Peer, E.; Rothschild, D.; Gordon, A.; Evernden, Z.; and Damer, E. 2022. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*.

Persson, A.; Laaksoharju, M.; and Koga, H. 2021. We Mostly Think Alike: Individual Differences in Attitude Towards AI in Sweden and Japan. *The Review of Socionetwork Strategies*.

Pfeffer, K.; Mai, A.; Weippl, E.; Rader, E.; and Krombholz, K. 2022. Replication: Stories as Informal Lessons about Security. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*. USENIX Association.

Phillips, A. M.; Watkins, J.; and Hammer, D. 2018. Beyond "asking questions": Problematizing as a disciplinary activity. *Journal of Research in Science Teaching*.

Posch, L.; Bleier, A.; Flöck, F.; Lechner, C. M.; Kinder-Kurlanda, K.; Helic, D.; and Strohmaier, M. 2018. Characterizing the global crowd workforce: A cross-country comparison of crowdworker demographics. *arXiv preprint arXiv:1812.05948*.

Proctor, R. N.; and Schiebinger, L. 2008. Agnotology: The making and unmaking of ignorance.

Prolific. 2023. Online participant recruitment for surveys and market research.

Qualtrics. 2023. Qualtrics XM - The Leading Experience Management Software.

QueerInAI, O. O. 2023. Queer In AI: A Case Study in Community-Led Participatory AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM.

Rader, E.; Wash, R.; and Brooks, B. 2012. Stories as Informal Lessons about Security. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*. ACM.

Rastogi, C.; Tulio Ribeiro, M.; King, N.; Nori, H.; and Amershi, S. 2023. Supporting Human-AI Collaboration in Auditing LLMs with LLMs. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. ACM.

Rea, L. M.; and Parker, R. A. 2014. *Designing and conducting survey research: A comprehensive guide*. John Wiley & Sons.

Reiser, B. J.; Brody, L.; Novak, M.; Tipton, K.; and Adams, L. 2017. Asking questions. *Helping students make sense of the world using next generation science and engineering practices*.

Reyes, A. 2019. Eugenic Visuality: Racist Epistemologies from Galton to "The Bell Curve". *Amerikastudien/American Studies*.

Ribeiro, F. N.; Saha, K.; Babaei, M.; Henrique, L.; Messias, J.; Benevenuto, F.; Goga, O.; Gummadi, K. P.; and Redmiles, E. M. 2019. On Microtargeting Socially Divisive Ads: A Case Study of Russia-Linked Ad Campaigns on Facebook. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM.

Rismani, S.; Shelby, R.; Smart, A.; Jatho, E.; Kroll, J.; Moon, A.; and Rostamzadeh, N. 2023. From Plane Crashes to Algorithmic Harm: Applicability of Safety Engineering Frameworks for Responsible ML. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM.

Roberts, D. 2012. *Fatal Invention*.

Rossi, P.; Wright, J.; and Anderson, A. 2013. *Handbook of Survey Research*. Elsevier Science.

Rostamzadeh, N.; Mincu, D.; Roy, S.; Smart, A.; Wilcox, L.; Pushkarna, M.; Schrouff, J.; Amironesei, R.; Moorosi, N.; and Heller, K. 2022. Healthsheet: Development of a Transparency Artifact for Health Datasets. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery.

Rust, J.; and Golombok, S. 2014. *Modern psychometrics: The science of psychological assessment*. Routledge.

Said, N.; Potinteu, A. E.; Brich, I.; Buder, J.; Schumm, H.; and Huff, M. 2023. An artificial intelligence perspective: How knowledge and confidence shape risk and benefit perception. *Computers in Human Behavior*.

Sambasivan, N.; Kapania, S.; Highfill, H.; Akrong, D.; Paritosh, P.; and Aroyo, L. M. 2021. "Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM.

Schaeffer, N. C.; and Presser, S. 2003. The science of asking questions. *Annual review of sociology*.

Scharff, D. P.; Mathews, K. J.; Jackson, P.; Hoffsuemmer, J.; Martin, E.; and Edwards, D. 2010. More than Tuskegee: understanding mistrust about research participation. *Journal of health care for the poor and underserved*.

Scharowski, N.; Benk, M.; Kühne, S. J.; Wettstein, L.; and Brühlmann, F. 2023. Certification Labels for Trustworthy AI: Insights From an Empirical Mixed-Method Study. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery.

Schrag, Z. M. 2010. *Ethical imperialism: Institutional review boards and the social sciences, 1965–2009*. JHU Press.

Schuler, D.; and Namioka, A. 1993. *Participatory design: Principles and practices*. CRC Press.

Schulz, A. J.; Zenk, S. N.; Kannan, S.; Israel, B. A.; Koch, M. A.; and Stokes, C. A. 2005. CBPR approaches to survey design and implementation. *Methods communitybased Particip Res Heal San Fr JosseyBass*.

Schurgin, M.; Schlager, M.; Vardoulakis, L.; Pina, L. R.; and Wilcox, L. 2021. Isolation in Coordination: Challenges of Caregivers in the USA. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380966.

Science, N.; and Council, T. 2023. National Artificial Intelligence Research and Development Strategic Plan, 2023 Update.

Selwyn, N.; Cordoba, B. G.; Andrejevic, M.; and Campbell, L. 2020. AI for social good: Australian public attitudes toward AI and society.

Septiandri, A. A.; Constantinides, M.; Tahaei, M.; and Quercia, D. 2023. WEIRD FAccTs: How Western, Educated, Industrialized, Rich, and Democratic is FAccT? In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM.

Simonsen, J.; and Robertson, T. 2012. *Routledge international handbook of participatory design*. Routledge.

Sindermann, C.; Sha, P.; Zhou, M.; Wernicke, J.; Schmitt, H. S.; Li, M.; Sariyska, R.; Stavrou, M.; Becker, B.; and Montag, C. 2021. Assessing the Attitude Towards Artificial Intelligence: Introduction of a Short Measure in German, Chinese, and English Language. *KI - Künstliche Intelligenz*.

Singer, A.; and Bishop, M. 2021. Trust-Based Security; Or, Trust Considered Harmful. In *Proceedings of the New Security Paradigms Workshop 2020*. ACM.

Sinnott-Armstrong, W. 2023. Consequentialism. In Zalta, E. N.; and Nodelman, U., eds., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2023 edition.

Sloane, M.; Moss, E.; Awomolo, O.; and Forlano, L. 2022. Participation Is Not a Design Fix for Machine Learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM.

Smith, A.; Christopher, S.; and McCormick, A. K. H. G. 2004. Development and implementation of a culturally sensitive cervical health survey: a community-based participatory approach. *Women & health*.

Smith, L. T. 2021. *Decolonizing methodologies: Research and indigenous peoples*. Bloomsbury Publishing.

Soden, R.; Toombs, A.; and Thomas, M. 2024. Evaluating Interpretive Research in HCI. *Interactions*.

Spector, P. E. 2013. Survey design and measure development. *The Oxford handbook of quantitative methods*.

Srinivasan, R.; Denton, E.; Famularo, J.; Rostamzadeh, N.; Diaz, F.; and Coleman, B. 2021. Artsheets for art datasets. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 2)*.

Sunarti, S.; Rahman, F. F.; Naufal, M.; Risky, M.; Febriyanto, K.; and Masnina, R. 2021. Artificial intelligence in healthcare: opportunities and risk for future. *Gaceta Sanitaria*.

SurveyMonkey. 2024. Calculate your sample size.

Tahaei, M.; Constantinides, M.; Quercia, D.; and Muller, M. 2023. A Systematic Literature Review of Human-Centered, Ethical, and Responsible AI.

Tan, R.; and Cabato, R. 2023. Behind the AI boom, an army of overseas workers in 'digital sweatshops'.

Tang, J.; Birrell, E.; and Lerner, A. 2022. Replication: How Well Do My Results Generalize Now? The External Validity of Online Privacy and Security Surveys. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*. USENIX Association.

Tapaha, O. G. 2017. " We Lived It": Stories of Cultural Resilience, Dinék'ehgo Nanitiin (Diné-Based Instruction), and Navigating between University and Tribal Institutional Review Boards.

Tillyard, G.; and DeGennaro Jr, V. 2019. New methodologies for global health research: Improving the knowledge, attitude, and practice survey model through participatory research in Haiti. *Qualitative Health Research*.

Torkamaan, H.; Tahaei, M.; Buijsman, S.; Xiao, Z.; Wilkinson, D.; and Knijnenburg, B. P. 2024. *The Role of Human-Centered AI in User Modeling, Adaptation, and Personalization—Models, Frameworks, and Paradigms*. Springer Nature Switzerland.

Tourangeau, R.; Rips, L. J.; and Rasinski, K. 2000. The psychology of survey response.

Tourangeau, R.; and Smith, T. W. 1996. Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public opinion quarterly*.

Trefzer, A.; Jackson, J. T.; McKee, K.; and Dellinger, K. 2014. Introduction: The global south and/in the global north: Interdisciplinary investigations. *The Global South*.

United Nations. 2022. A Future with AI: Voices of Global Youth Report Launched.

van Berkel, N.; Sarsenbayeva, Z.; and Goncalves, J. 2023. The methodology of studying fairness perceptions in Artificial Intelligence: Contrasting CHI and FAccT. *International Journal of Human-Computer Studies*.

Veselovsky, V.; Ribeiro, M. H.; and West, R. 2023. Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks.

Wacharamanotham, C.; Eisenring, L.; Haroz, S.; and Echtler, F. 2020. Transparency of CHI Research Artifacts: Results of a Self-Reported Survey. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM.

Wang, D.; Prabhat, S.; and Sambasivan, N. 2022. Whose AI Dream? In Search of the Aspiration in Data Annotation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM.

Wilcox, L.; Brewer, R.; and Diaz, F. 2023. AI Consent Futures: A Case Study on Voice Data Collection with Clinicians. *Proc. ACM Hum.-Comput. Interact.*

Wilcox, L.; Shelby, R.; Veeraraghavan, R.; Haimson, O. L.; Erickson, G. C.; Turken, M.; and Gulotta, R. 2023. Infrastructuring Care: How Trans and Non-Binary People Meet Health and Well-Being Needs through Technology. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.

Winston, A. S. 2020. Scientific racism and North American psychology. In *Oxford research encyclopedia of psychology*.

Wong, R. Y.; and Khovanskaya, V. 2018. *Speculative design in HCI: from corporate imaginations to critical orientations*. Springer.

Yigitcanlar, T.; Desouza, K. C.; Butler, L.; and Roozkhosh, F. 2020. Contributions and Risks of Artificial Intelligence (AI) in Building Smarter Cities: Insights from a Systematic Review of the Literature. *Energies*.

Young, M.; Katell, M.; and Krafft, P. 2022. Confronting Power and Corporate Capture at the FAccT Conference. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM.

Zanetti, M.; Iseppi, G.; and Cassese, F. P. 2019. A "psychopathic" Artificial Intelligence: the possible risks of a deviating AI in Education. *Research on Education and Media*.

Zhang, B.; and Dafoe, A. 2020. U.S. Public Opinion on the Governance of Artificial Intelligence. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. ACM.

A Known Limitations of Surveys

There are several well-known issues with surveys that are often acknowledged as limitations in studies. Below, we list several common biases; however, this is not an exhaustive list. For a more comprehensive discussion, we refer to survey design papers and textbooks such as (Müller, Sedley, and Ferrall-Nunge 2014; Krosnick 1999; Choi and Pak 2005).

- **Acquiescence bias and experimenter effect**: Survey respondents may agree with a question or statement regardless of their actual feelings or attitudes. This can stem from a desire to be agreeable, a lack of motivation in answering questions, due to the influence of the researcher or the institution conducting the survey, or due to desire to satisfy what respondents think the researchers' expectations from the study are.
- **Satisficing:** Respondents may opt for answers that merely satisfy the survey's requirements, rather than seeking the most accurate or optimal response. This behavior could be due to the effort involved, distractions, or a lack of interest in the survey's outcomes.
- **Social desirability:** In response to sensitive questions, survey participants may answer in a manner they believe will be viewed favorably by others. For instance, questions about sexual orientation or taboo subjects might elicit responses that do not accurately reflect the respondent's true stance.
- **Question and response order bias:** The sequencing of questions or answer options can influence survey results. The order in which questions are presented can prime respondents' views as they progress through the survey. Similarly, non-randomized answer choices may not have an equal chance of being selected.
- **Framing effects:** The choice of wording in questions may affect survey responses. For example, users' agreement or disagreement with specific statements may depend on whether those statements are positively or negatively framed. The choice of specific words (even among synonyms, such as concern vs. worry vs. fear vs. discomfort) may also lead to different outcomes. Nuances in translation may further complicate the interpretation of questions and answers in multi-lingual studies. Other wording choices and associated mistakes (e.g., double negation, double-barreled questions, leading questions, etc.) can decrease respondents' comprehension of the questions and introduce biases in the analysis, compromising the objectivity and accuracy of the survey results. More broadly, framing effects may also emerge from the overall narrative of the survey, for example, suggesting a dichotomous trade-off between benefits and risks, without considering other nuances of the discourse or other potential factors affecting respondents' perspectives.
- **Sampling bias:** Collecting survey responses from a non-representative sample reduces the generalizability of the results to the broader target population, which may differ in socio-demographics and experiences. Similarly, by focusing on AI users, researchers exclude the perspectives of those who do not use AI or lack the expertise, experience, motivation, or resources to use it. Conducting research in a single country or language further limits the ability to capture cross-cultural differences. Even the data collection format (e.g., online vs. pen-and-paper surveys, accessibility features) can impact results by excluding individuals who lack the necessary knowledge, skills, access, or physical or cognitive abilities to complete the survey

Despite the acknowledged limitations and known issues, surveys remain a frequently employed method for gathering data on opinions, behaviors, attitudes, knowledge, personal characteristics, and motivations. This trend extends to AI research as well, with surveys playing a significant role in this field (for examples, refer to the paper).

B Additional Materials for Pilot Survey

Survey Instrument

[Answer options to close-ended questions were randomized where appropriate.]

- Have you heard of the term "Artificial Intelligence" (or "AI")?

  – Yes, I have heard of the term "Artificial Intelligence" (or "AI"), and I feel confident explaining what it means to an expert.

  – Yes, I have heard of the term "Artificial Intelligence" (or "AI"). However, I do not feel confident explaining what it means to an expert.

  – No, I have not heard of the term "Artificial Intelligence" (or "AI").

  The following questions are based on your understanding of *existing* AI systems.
- How do you think existing AI systems could *benefit* you? Please give details (at least 100 characters).
- We want to know what you have learned from others about the *benefits* of *existing* AI systems. Specifically, we are interested in stories you have heard about the benefits of existing AI systems from *OTHER PEOPLE*, such as friends, coworkers, social media sites, TV shows, news websites, blogs, or any other sources—NOT experiences that happened to you personally. Describe in detail the most memorable story (at least 100 characters).
- How did you hear about that story?

- How do you think existing AI systems could put you *at risk*? Please give details (at least 100 characters).
- We want to know what you have learned from others about the *risks* of *existing* AI systems. Specifically, we are interested in stories you have heard about the risks of existing AI systems from *OTHER PEOPLE*, such as friends, coworkers, social media sites, TV shows, news websites, blogs, or any other sources—NOT experiences

that happened to you personally. Describe in detail the most memorable story (at least 100 characters).

- How did you hear about that story?

- Please select Excellent to show you are paying attention to this question [Very Poor, Poor, Fair, Good, Excellent].

In this section, we want to learn about the different ways you *envision* AI systems working.

- If you had a magic wand that could create an AI system, what would you want that AI system to do for you? Please give details (at least 100 characters).
- How could the AI system that you just described put you (or someone else) *at risk*? Please give details (at least 100 characters).

Please answer the question below given the following definition of an AI system:
"An AI system is a technology that can generate outputs such as predictions, recommendations, or decisions influencing real or online spaces. AI systems are designed to work with different levels of independence, meaning some might need more human guidance, while others can handle tasks on their own."

- Based on this definition, in your opinion, what characteristics should an AI system have to be *trustworthy*? Please describe "your" understanding using your own words. Please give details (at least 100 characters).

- Please select Rarely to show you are paying attention to this question [Always, Never, Rarely].

[Demographics]

- What best describes your employment status over the last three months?
  - Working full-time
  - Working part-time
  - Unemployed and looking for work
  - A homemaker or stay-at-home parent
  - Student
  - Retired
  - Other
- What ethnic group describes you the best? [Open-ended]
- What is the highest level of education you have completed?
  - Some high school or less
  - High school diploma or GED
  - Some college, but no degree
  - Associates or technical degree
  - Bachelor's degree
  - Graduate or professional degree (MA, MS, MBA, PhD, JD, MD, DDS, etc.)
  - Prefer not to say
- Please select the option that best describes your personal income relative to others in your age group and location.
  - Below average
  - Average
  - Above average
  - Unsure
  - Prefer not to say
- Please answer the following questions [Yes, Sort of, No].
  - I know how to program in at least one programming language.
  - My family members or friends often ask me for computing-related advice.
  - I study or work in IT or a computing-related field.

Participant Demographics

Table 1: Participant demographics (N=282).

Country of residence	
Australia	50 (18%)
Israel	48 (17%)
Chile	47 (17%)
United Kingdom	47 (17%)
United States	46 (16%)
South Africa	44 (16%)
Ethnicity	
White	161 (57%)
Black	48 (17%)
Mixed	34 (12%)
Other	19 (7%)
Asian	18 (6%)
Not available	2 (1%)
Employment status	
Working full-time	126 (45%)
Working part-time	54 (19%)
Student	54 (19%)
Unemployed and looking for work	34 (12%)
Other	5 (2%)
Retired	5 (2%)
A homemaker or stay-at-home parent	4 (1%)
Gender	
Female	143 (51%)
Male	139 (49%)
Income relative to age group and location	
Average	116 (41%)
Below average	79 (28%)
Above average	66 (23%)
Prefer not to say	11 (4%)
Unsure	10 (4%)
Familiarity with AI	
Heard but can't explain to an expert	169 (60%)
Heard and can explain to an expert	113 (40%)
Technical background	62 (22%)

The final dataset consisted of participants from six countries. Table 1 summarizes the demographics of our participants. We achieved a balanced sample in terms of gender.

Racially and ethnically, many participants (57%) identified as White. In terms of employment status, several participants were employed full-time (45%), and described their income as average (41%).

Concerning familiarity with AI, 169 participants (60%) expressed that they had heard of the term but did not feel confident explaining its meaning to an expert, and 113 participants (40%) were familiar with AI and felt confident explaining its meaning to an expert. No participant claimed to be unfamiliar with the term. Based on our criteria for assessing the technical background of participants, 62 participants (22%) had some technical background.

Reflections: Impact of Researcher Tools, Practices, and Choices

In this section, we revisit the decisions made during the design, deployment, and analysis of our pilot survey that are not necessarily covered in the literature review but were crucial considerations. Each subsection addresses a specific question we encountered and had to deliberate upon. A more concise set of heuristic questions can be found in the paper.

How to Frame Questions? Our pilot survey was rooted in prior research with public perception of AI and echoed their consequentialist ethics framing—"*the view that normative properties depend only on consequences*" (Sinnott-Armstrong 2023). In our case, the framing of the questions with "How do you think existing AI systems could *benefit* you?" explicitly emphasizes the consequences instead of anything else. From an alternative standpoint such as deontologist—"*normative theories regarding which choices are morally required, forbidden, or permitted . . . In contrast to consequentialist theories, deontological theories judge the morality of choices by criteria different from the states of affairs those choices bring about*" (Alexander and Moore 2021)—, we could have instead asked about what rules should AI models or systems follow or not follow, instead of asking about specific desired and risky consequences: "What rules should a beneficial AI technology follow?" Alternatively, we could have focused on virtue ethics—"*. . . the one that emphasizes the virtues, or moral character, in contrast to the approach that emphasizes duties or rules (deontology) or that emphasizes the consequences of actions (consequentialism)*" (Hursthouse and Pettigrove 2023)—and asked more about the traits or virtues that AI should have: "How would you describe a generous AI?"

As for deployment of terms such as "risk" and "benefit" in the survey, it is in fact imperative to understand, in a much more localized way, the unique cultural and philosophical underpinnings of concepts like "risk" and "benefit" in order to write reliable survey questions on these concepts, given their cultural meaning and situation, and the high degree of influence that their surrounding social values, norms, and beliefs play in people's actual risk assessments (Douglas and Wildavsky 1983; Boholm 1996; Beck 1992; Bernstein and Bernstein 1996; Beck 1996).

All these viewpoints are contingent on definitions that vary cross-culturally. Without transparent reporting and the availability of artifacts, readers are unable to determine precisely what questions were asked, making it impossible to know whether there was a downplaying or overlooking of potential concerns raised by participants. Participants may express concerns or exhibit excitement, but this does not necessarily mean that their concerns are outweighed by the perceived benefits, especially when considering the actual net benefit or harm to others. However, results are sometimes simplistically reported as statements like "Global Study Shows Optimism About AI's Potential" (Google 2024) or "AI is making the world more nervous," (Boyon 2023), failing to capture these nuances and trade-offs.

Should We Translate the Survey? The practice of translating questions from English into other languages for comparative analysis warrants attention. While such translations, which may be conducted meticulously and carefully (e.g., Kelley et al. 2021), aim to include more countries or cultures, they may not accurately reflect other cultures' perspectives. Concluding that direct experiences with AI alone drive excitement and alleviate concerns is an oversimplification. Attitudes toward AI are shaped by various factors, including media representation, personal experiences and beliefs, and societal narratives, and not solely by direct interactions with AI. Individuals may lack meta-awareness of these other factors influencing their perceptions. Direct interactions with AI could be detrimental to some individuals, and those most vulnerable to being marginalized by AI may be individuals who have not experienced it firsthand. Furthermore, survey results indicating a pronounced positive orientation in certain regions and emerging markets (e.g., Google 2024) could result from different stages of AI adoption, economic factors, or cultural attitudes toward technology, which may not be adequately captured in a survey.

We considered using AI for (1) generating translations of questions, (2) translating participant responses, and (3) analyzing responses. However, given the ongoing debates about employing Large Language Models (LLMs) in research (e.g., Paxton 2023; Rastogi et al. 2023; Hosseini, Resnik, and Holmes 2023; Byun, Vasicek, and Seppi 2023), we realized that using AI without informing participants would be inappropriate. We took into account the following: (1) the risk that participants' data might be absorbed into corporate-owned LLMs, creating a permanent trace of their participation and ideas, (2) the accuracy of AI in conveying participants' true intentions and responses, and (3) the potential use of these responses for future model training, and the impact of over-surveyed populations on these models.

In any of these use cases, we were uncertain about how to seek consent from our participants for using AI to analyze their data. Typically, consent for participation in research covers privacy, data security, and its use by the research team. Moving forward, however, researchers, including those in UX, may need to offer participants the option to have their data analyzed by AI, with full disclosure of what this entails. Under such an approach, would researchers need to provide participants with the analysis summary for their final approval? Should participants be given the chance to review the summary if AI is used, to ensure it aligns with their perspective, considering that the analysis is no longer

conducted solely by individuals trained in positionality and research ethics?

Where to Find Participants? Our team's diversity, spanning various affiliations, faced limitations in participant recruitment due to being constrained by choosing appropriate platforms. Differing experiences with platforms such as MTurk and Prolific among team members added to these challenges. For instance, institutional policies prevented one member from using Prolific, despite their willingness to fund the project. Consequently, another team member, with more flexibility in platform choice and budget, took responsibility for project funding. The literature review also indicates a wide range of recruitment methods, from crowdsourcing platforms to emails and professional research companies.

Prolific's reach was limited in terms of global access. We attempted to recruit an equal number of male and female participants from over 30 countries, aiming for 50 participants from each country, but were unsuccessful in many countries. Consequently, our final participant pool comprised individuals from Australia, Chile, Israel, the United Kingdom, the United States, and South Africa. Notably, there was an unexpected lack of participation from regions such as China, India, South America, and Africa, some of the most populous areas globally. Additionally, Prolific's enforced binary gender option further limited our reach to populations that do not fall into the binary classification of gender.

Additionally, the use of online platforms excludes certain groups, such as those without Internet access, children, individuals in countries with different payment methods, and people not connected to platform users, resulting in a sampling bias. Prolific primarily recruits participants through word of mouth and social media.[7] While convenient, results may be skewed toward those with stronger opinions or experiences with AI, as people more interested in or affected by AI are more likely to participate, creating a self-selection bias. This bias is particularly evident in crowdsourcing platforms like Prolific or MTurk, where individuals can choose which tasks to take, with interest in the survey topic being a significant motivator for participation (Kaufmann, Schulze, and Veit 2011). However, prior work indicates that this is less of a prominent reason for *professional* MTurk users, who primarily select surveys based on compensation rather than interest in the topic (Lovett et al. 2018).

How Much Should We Pay Participants? In addition to access considerations, we decided to compensate all participants based on the payment rates of our home institution. However, this decision prompts a critical question: Should compensation be adjusted based on the participant's country or kept consistent across all countries? This dilemma underscores the challenges of fair participant treatment across diverse regions and economies, emphasizing the need for a nuanced compensation approach that considers both institutional standards and the participants' economic situations.

Examining the data from our literature review, it is evident that researchers employ diverse methods for compensating survey participants, ranging from unspecified (not included in the paper) to free or paid, with varying amounts reported (refer to the paper for detailed findings). Given the well-documented ethical concerns surrounding data annotation practices in AI research, particularly in terms of labor and payment (Wang, Prabhat, and Sambasivan 2022; Tahaei et al. 2023; Tan and Cabato 2023), the question of what constitutes an ethical way of compensating survey participants remains open for discussion.

How to Analyze Data? Finding a suitable framework for a top-down analysis of data, especially regarding the benefits and risks of AI, was difficult. This challenge arises mainly because most existing frameworks come from Western countries, and there is a lack of comprehensive frameworks or taxonomies for a complex and rapidly changing field like AI. For example, the difficulties in applying AI safety taxonomies are well-recognized due to the ever-evolving nature of AI (Rismani et al. 2023).

Given these constraints, our approach predominantly involved bottom-up analysis for most of the questions. An exception was made for the "trustworthiness" question, in which we applied a combination of top-down and bottom-up methods. For the top-down component, we used the NIST AI Risk Management Framework (NIST 2023) as a starting point. However, this choice introduced a US-centric bias into our analysis, potentially excluding or marginalizing other cultural views. The use of a framework developed within a specific cultural and institutional context raises important questions about the universality and applicability of our analysis across different contexts. Our positionality statement provides an initial reflection on our analytical perspective. Yet, a more in-depth consideration of the frameworks we use and their impact on our results highlights the broader issue of institutional and methodological positionalities. This presents a question for our research community: How can we effectively address these positionalities?

C Additional Materials for the Systematic Literature Review

Query

- *Who? Public.* "public" OR "representative" OR "population" OR "citizen" OR "citizens" OR "civic" OR "community" OR "non-expert" OR "non-experts".
- *What? AI.* "artificial intelligence" OR "machine learning" OR "deep learning" OR "AI".
- *What? Perceptions.* "thoughts" OR "feel" OR "feels" OR "feeling" OR "experience" OR "experiences" OR "feelings" OR "perception" OR "perceptions" OR "perceive" OR "attitude" OR "attitudes" OR "opinion" OR "opinions" OR "view" OR "views".
- *How? Surveys.* "survey" OR "surveys" OR "poll" OR "polls" OR "questionnaire" OR "questionnaires".

[7] "Participants are primarily recruited to Prolific via word of mouth, including word of mouth via social media. When Prolific was founded in 2014, our participants were recruited via three channels: 1) Social media (e.g., Facebook, Twitter, Reddit, and various other online forums). 2) Flyer distribution on university campuses. 3) The Prolific referral scheme (ceased March 2019). This allowed participants to invite their social network to join Prolific, in return for small cash incentives for the referrer" (Prolific 2023).

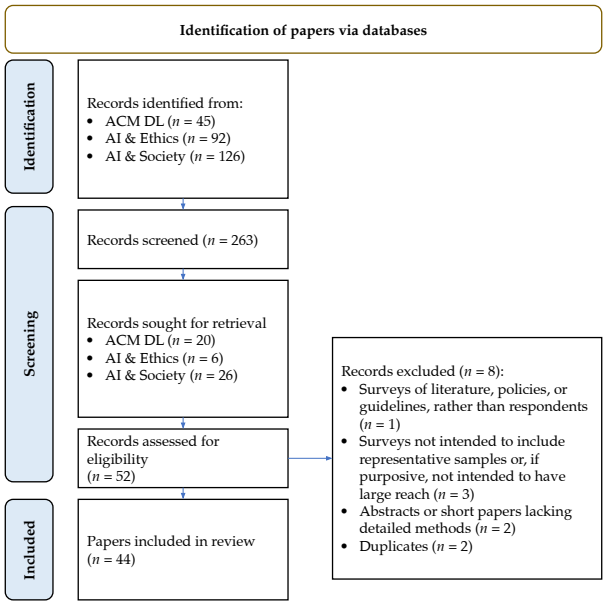

Figure 1: Prisma diagram for the systematic literature review. See the paper for details.

Prisma Diagram

- ACM DL ($n = 45$): We searched the entire ACM database, applying the aforementioned query to titles or abstracts from the most recent two years. This yielded nine records, with two overlapping with FAccT and AIES searches, and one exclusion as it was not a survey of people. For FAccT ($n = 18$) and AIES ($n = 17$), we executed the query on FAccT and AIES papers searching by title or abstract without a date restriction to broaden our reach. We also excluded the AI clause, as these conferences are inherently centered on AI and ethics. After manually screening the results, we included five papers from FAccT and four from AIES in our final collection.
- AI & Ethics ($n = 92$) and AI & Society ($n = 126$): We carried out a full-text search in Springer's database for the past two years using our query, noting that the database does not offer options for searching by title or abstract. All titles and abstracts were manually reviewed, applying our exclusion criteria, resulting in 26 papers from AI & Society and six from AI & Ethics being selected. Following a secondary review, three papers from AI & Society were excluded, while all papers from AI & Ethics were retained.