How Different Groups Prioritize Ethical Values for Responsible AI

Maurice Jakesch
Cornell University
New York, NY, USA
mpj32@cornell.edu

Zana Buçinca
Harvard University
Boston, MA, USA

Saleema Amershi
Microsoft Research
Redmond, WA, USA

Alexandra Olteanu
Microsoft Research
Montreal, QC, Canada

ABSTRACT

Private companies, public sector organizations, and academic groups have outlined ethical values they consider important for responsible artificial intelligence technologies. While their recommendations converge on a set of central values, little is known about the values a more representative public would find important for the AI technologies they interact with and might be affected by. We conducted a survey examining how individuals perceive and prioritize responsible AI values across three groups: a representative sample of the US population (N=743), a sample of crowdworkers (N=755), and a sample of AI practitioners (N=175). Our results empirically confirm a common concern: AI practitioners' value priorities differ from those of the general public. Compared to the US-representative sample, AI practitioners appear to consider responsible AI values as less important and emphasize a different set of values. In contrast, self-identified women and black respondents found responsible AI values more important than other groups. Surprisingly, more liberal-leaning participants, rather than participants reporting experiences with discrimination, were more likely to prioritize fairness than other groups. Our findings highlight the importance of paying attention to who gets to define "responsible AI."

CCS CONCEPTS

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

KEYWORDS

Responsible AI, value-sensitive design, empirical ethics

ACM Reference Format:
Maurice Jakesch, Zana Buçinca, Saleema Amershi, and Alexandra Olteanu. 2022. How Different Groups Prioritize Ethical Values for Responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3531146.3533097

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533097

1 INTRODUCTION

Advances in artificial intelligence (AI) have the potential to benefit people and society, but they also raise ethical challenges and concerns about possible adverse impacts [52]. Being prone to errors and biases, AI systems may harm people [2] for instance by reinforcing stereotypes [7] or by increasing social inequality [21]. While the larger consequences of AI can be difficult to anticipate [8], systems developed with broader human and societal values in mind stand a better chance of preserving these values [1, 23, 57]. To support the development of socially beneficial AI technologies, several private companies, public sector organizations, and academic groups have published ethics guidelines with values they consider important for responsible AI [37].

These AI ethics guidelines have been found to largely converge on five central values [37]: transparency, fairness, safety, accountability, and privacy. But these values may differ from what a broader and more representative population would consider important for the AI technologies they interact with. While prior work has shown that value preferences depend on peoples' backgrounds and personal experiences [14, 36], AI technologies are often developed by relatively homogeneous and demographically skewed subsets of the population [13, 32, 42]. Given the lack of reliable data on other groups' priorities for responsible AI, practitioners may unknowingly encode their own biases and assumptions into their concept and operationalization of responsible AI [48, 57].

In this work, we present the results of a survey we developed, validated, and fielded to elicit peoples' value priorities for responsible AI. Drawing on the traditions of empirical ethics [19, 53] and value elicitation research [22, 63], our survey asks participants about the perceived importance of a set of 12 responsible AI values both in general and in specific deployment scenarios. To increase robustness, respondents assessed values from three perspectives: value selection, contextual assessment of values, and comparative prioritization of values (detailed in §3.2).

We administered this survey to three different populations. We analyzed how value priorities of a US census-representative sample (N=743), a crowdworker sample (N=755), and an AI practitioner sample (N=175) vary by deployment scenario and individuals' backgrounds and experiences. We surveyed the value priorities of AI practitioners as they are often the ones making decisions about the AI technologies that are being developed, and compared their preferences to those of a more representative sample. We also consulted crowdworkers as they are already involved in producing data that

AI systems are evaluated on to explore the feasibility of involving them in the ethical assessment of AI systems as well.

Our results provide evidence that responsible AI values are perceived and prioritized differently by different groups. AI practitioners, on average, rated responsible AI values less important than other groups. At the same time, AI practitioners prioritized fairness more often than participants from the US-census representative sample who emphasized safety, privacy, and performance. We also find differences in value priorities along demographic lines. For example, women and black respondents evaluated responsible AI values as more important than other groups. We observed the most disagreement in how people traded-off fairness with performance. Surprisingly, participants reporting past experiences of discrimination did not prioritize fairness more than others, but liberal-leaning participants prioritized fairness more than conservative-leaning participants.

Our results highlight the need for AI practitioners to contextualize and probe their ethical intuitions and assumptions. The empirical approach to AI ethics explored in this study can help to increase the context sensitivity of the responsible AI development process. However, as we elaborate in the discussion, opinion research can inform ethical decision-making, but cannot replace sound ethical reasoning.

2 BACKGROUND

Our study draws on prior work on responsible AI, value sensitive design [23], empirical ethics [53], value elicitation [22, 63], and standpoint theory [36].

2.1 AI ethics guidelines and value-sensitive design

Science and technology studies theorize that computing technologies incorporate a tacit understanding of human nature [70]. Algorithms are described as value-laden artifacts [48] that encode developer assumptions, including ethical and political values [57]. From this perspective, a product team that decides to maximize the chance that a disease detection system will recognize a disease at the cost of increasing false alarms prioritizes certain values over others. Past work has shown that machine learning development and research often narrowly focus on technical values such as accuracy, efficiency, and generalization [6, 54]. In contrast, proponents of value-sensitive design [23, 24], reflective design [64], and critical technical practice [1] advocate that AI systems should be designed with broader human and societal values in mind.

What values developers of responsible AI systems should emphasize remains a key question. Some argue these values should be naturally embedded in an organization's culture [57]. Several organizations have also published guidelines describing what values they believe AI systems should embody. Jobin et al. [37] found these guidelines to converge around central values, but differ in how they construe these values and concepts. Critics note that reliable methods to translate values into practice are often missing [51, 57]. Some also argue that statements of high-level values and principles are too ambiguous and may gain consensus simply by masking the complexity and contending interpretations of ethical concepts [69]. For example, people may agree on the importance of fairness, but "fairness" in and by itself has little to say about what is fair and why [5].

Our study validates and contextualizes value priorities outlined in AI ethics guidelines. To date, there is little empirical data on values a broader and more representative public finds important for the AI technologies they interact with. Our empirical approach to AI ethics probes for possible blind spots in AI practitioners' and researchers' assumptions.

2.2 Empirical studies of human values and AI ethics

Eliciting people's values is a central pursuit in the social sciences [22]. Economists explain choices in the marketplace based on value theory, sociologists seek to understand which values are held by a community and how they change. Psychologists use value elicitation for therapy and counsel, and empirical ethicists enhance the context-sensitivity of their arguments by combining social scientific methods with ethical reasoning [53]. While drawing normative conclusions from empirical results is difficult, empirical data on ethical preferences can inform decision making [53].

Several studies have examined people's ethical intuitions concerning AI technologies. In the "moral machine" experiment, Awad et al. [2] generated a variety of moral dilemmas a self-driving car might find itself in and ask participants which course of action they recommend. They report significant cross-cultural differences in ethical preferences correlated with modern institutions and cultural traits. Hidalgo et al. [31] explored how people judge humans and machines differently when they make mistakes. They found that people tend to forgive machines more in scenarios with high intentionality. Similarly, Malle et al. [47] compared how people apply moral norms to humans versus robots. Most related to the empirical study of responsible AI values, Saxena et al. [60] have compared public perceptions of different fairness paradigms. Similarly, Grgic-Hlaca et al. [28] and Pierson [55] have studied which features people find fair to include in a prediction algorithm. They found substantial disagreement among participants [28], with e.g., women being less likely to include gender as a feature in a course recommendation algorithm if this might result in female students seeing fewer recommendations for science courses [55].

Going beyond previous work, we develop a responsible AI value survey to explore what values people find most important for responsible AI. Where previous studies have elicited preferences concerning specific technical implementations with convenience samples, we provide a first high-level perspective on a representative public's priorities for the AI system they interact with and might be affected by.

2.3 The impact of individual background & of context on how values are prioritized

Feminist empiricists and standpoint theorists argue that knowledge is achieved from a particular standpoint [71] and that social location systematically influences our experiences and decisions [36]. They hold that homogeneous communities are prone to false consensus effects [59] where individuals believe that the collective opinion of their own group matches that of the larger population. In homogeneous communities, inaccurate assumptions or biases can be

hard to recognize and correct [8, 36]. In communities comprised of individuals with diverse values and experiences, however, how assumptions influence reasoning becomes more visible [36, 45, 58]. Including historically underrepresented groups, in particular, may lead to rigorous critical reflection as their experiences may facilitate the identification of problematic background assumptions [36].

Demographics and experiences not only affect background assumptions [18], but also shape people's values and ethical preferences [25, 27]. Rather than stemming from overarching belief systems, values often arise through particular social practices in a specific context [46]. As such, ethical intuition is contextual and socially situated [14]. For instance, what's fair to some people may seem unfair to others [43], and some people value privacy and autonomy more than others [69]. The population of AI practitioners is demographically skewed [13, 32, 42] with e.g., women and black people being underrepresented [16]. With their specific demographics and experiences, AI practitioners may bring their own preferences to what it means for AI to be "responsible" or "ethical", such as a bias towards deployment [40]. Responsible AI technologies developed within homogeneous communities may fail to account for the experiences and needs of various groups, so it remains crucial to scrutinize who gets to define AI ethics [38].

By surveying representative population samples about their priorities for responsible AI, we seek to validate the value prioritization in AI ethics frameworks. We explore the social relativity of responsible AI values to provide grounds for more critical reflection about possibly inaccurate assumptions and false consensus effects.

3 METHODS

To study how people perceive and prioritize responsible AI values, we combine instruments from value elicitation research [22] with the concepts and principles found in AI ethics guidelines [37]. We fielded an iteratively developed online survey with 743 census-representative participants, 755 crowd workers, and 175 AI practitioners.

3.1 Survey development

We adapted the Schwartz Value Survey [61, 62] to apply it to responsible AI values. The Schwartz Value Survey has been used to study individual and intercultural differences in general human values in over 60 countries [63]. Based on an inventory of human values, the Schwartz Value Survey asks respondents to self-report which values are most important to them. Respondents rate the importance of each value on a Likert scale while explanations for each value are shown.

Selecting and explaining responsible AI values. To adapt the Schwartz Value Survey to the study of AI ethics, we constructed an inventory of responsible AI values. The responsible AI values we chose for our survey are based on a review of published AI ethics guidelines. We drew on work by Jobin et al. [37] finding that AI ethics guidelines commonly refer to transparency, justice & fairness, non-maleficence, accountability, privacy, beneficence, freedom & autonomy, trust, and dignity. To this list, we added system performance, as it is a central value in AI research and development [6] that is often used to compare AI models and to make deployment decisions.

As responsible AI values are abstract and participants may not easily understand how they apply in the context of AI technologies [11], we provided additional explanations. To formulate explanations for each value, we drew again on existing AI ethics guidelines, including Microsoft's responsible AI principles [50], Google's AI Principles [26], the Montreal Declaration for the Responsible Development of Artificial Intelligence [52], the Deloitte AI ethics guide [15], IBM's Principles for Trust and Transparency [35], and the EU's Ethics guidelines for trustworthy AI [67].

We tested and iterated on different explanations of responsible AI values in four crowdsourcing pilot studies (N_1=40, N_2=80, N_3=40, N_4=160). Each pilot asked participants whether they understood an explanation through both Likert scales and open-ended responses. Based on the pilot results, we substituted "non-maleficence" with "safety" and "beneficence" with "social good," as the former were not well-understood by participants. We also explicitly referred to "*human* autonomy" to avoid confusion with autonomous cars and robots. Finally, we did not include "trust" as it appeared overly general and overlapped with other values such as transparency and accountability.

We phrased the explanations in simple, non-technical language, all following the same structure. Each explanation starts with a sentence describing what a system embodying the value would do, followed by an example of steps developers might take to realize a value, e.g.: "*An AI system that respects people's autonomy avoids reducing their agency. Developers of autonomy-preserving AI systems ensure, as far as possible, that the system provides choices to people and preserves or increases their control over their lives.*" By complementing a general definition with specific operationalizations of a value, the framing provides a tangible understanding of the value while maintaining a degree of generality. See Appendix A.2 for a complete list of the explanations we used in our survey.

Identifying pairs of possibly conflicting responsible AI values. In addition to assessments of values themselves, we asked participants about their preferences in cases of conflicting values [4]. For example, ensuring fairness might require collecting additional sensitive data, potentially diminishing privacy. To identify value conflicts, we searched for mentions of conflicts in the literature for each pair of values in the responsible AI value inventory. We found prior discussions of trade-offs between privacy & performance [3, 66], fairness & privacy [3, 20], fairness & performance [12, 56], safety & transparency [10, 33, 49], and autonomy & safety [44]. We combined the value explanations developed above to introduce the conflicts to participants, e.g. "*The developers realize that minimizing the collection of sensitive data (ensuring privacy) may make the system's predictions less accurate (reducing performance). Should they prioritize privacy or performance?*"

Constructing hypothetical AI deployment scenarios. We used hypothetical scenarios to make value assessments more tangible and to elicit judgments in specific contexts. We produced four hypothetical deployment settings validated through two pilot studies (N_1=180, N_2=160). To design these scenarios, we selected 25 AI systems people may have encountered in everyday settings starting with a list of general AI use cases [17]. We developed short explanations of these use cases and asked pilot participants whether they found them understandable and relatable. Based on the pilot results,

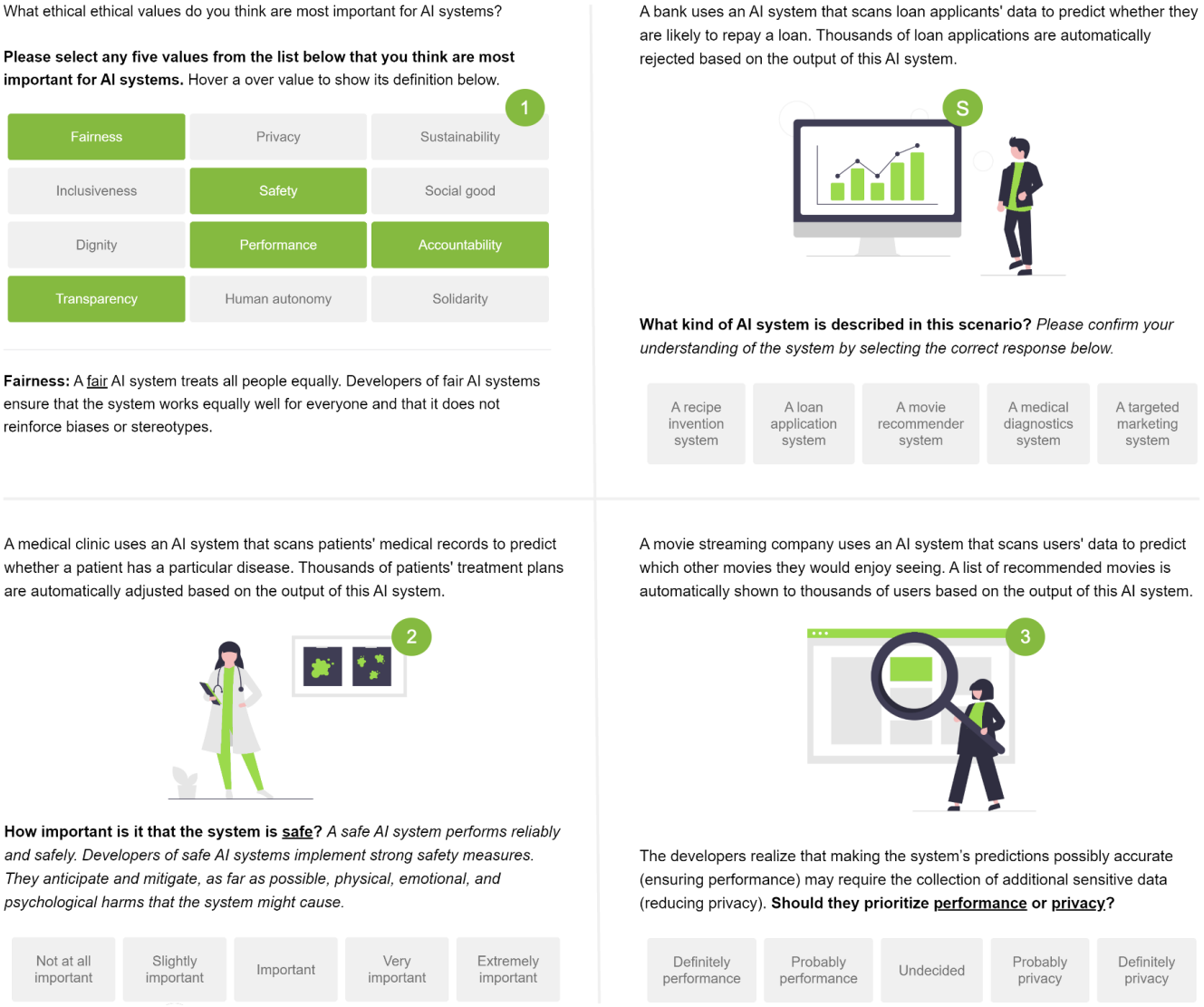

Figure 1: Overview of the main survey components. Participants first completed a value selection task (1). After confirming the understanding of the respective deployment scenario (S), they evaluated how the importance of values in context (2). Finally, participants indicated how they would prioritize values when they are in conflict (3).

we further refined the scenarios and kept only the 10 scenarios that were most easily understood by pilot participants. The second pilot then asked participants which scenarios they understood best and whether the AI system's decisions were highly consequential. Based on the responses, we selected two well-understood high-stake and low-stake scenarios for the study:

(a) Medical: An AI system used by a medical clinic to predict whether a patient has a disease (high-stake)
(b) Banking: An AI system used by a bank to predict whether an applicant will repay a loan (high-stake)
(c) Marketing: An AI system used by a marketing company to match ads to viewers (low-stake)
(d) Streaming: An AI system used by a streaming company to recommend movies to users (low-stake)

Each scenario states the entity controlling the AI system and the type of data the system is using. It then elaborates what predictions are being made and what actions are being taken based on the prediction, e.g.: "A medical clinic uses an AI system that scans patients' medical records to predict whether a patient has a particular disease. Thousands of patients' treatment plans are automatically adjusted based on the output of this AI system." The full list of scenarios is included in Appendix A.4.

3.2 Survey procedure

After providing informed consent, participants received a high-level introduction both covering the general goals of AI and noting the complex decision-making involved in the AI system development beyond technical challenges (see Appendix A.1). Figure 1 illustrates the subsequent survey steps which combined three value elicitation tasks: (1) *value selection*—select five responsible AI values (out of the 12) that are deemed most important in general, (2) *contextual assessment*—evaluate the perceived importance of seven central responsible AI values (transparency, fairness, safety, accountability, privacy, autonomy, and performance) in a specific deployment setting, and (3) *comparative assessment*—recommend what product teams should do when values are in conflict.

Participants selected the five most important values for AI systems in general, with explanations displayed when a value was hovered over. They then read the first scenario and confirmed their understanding of the deployment setting. Overall, participants encountered four scenarios. In scenarios 1 and 2, participants indicated how important they thought three responsible AI values were in the given situation on a 5-point Likert scale. In scenario 3, participants evaluated one more value and then two value conflicts by indicating which value they thought should be prioritized in the given situation. Finally, they evaluated three value conflicts in the fourth and last scenario. For every rating, participants were given the option to explain their choices.

After completing the rating tasks, participants indicated their familiarity with machine learning, user research, and their personal experiences with discrimination. We selected these experiential correlates based on the hypothesis that personal experience might inform ethical preferences [14]. For example, user researchers may have learned to empathize with users, whereas respondents trained in ML may have better insight into the technical constraints of responsible AI. We also asked participants to report their gender identity, age, ethnicity, political views, sector of work, and highest level of education. Again, these demographic correlates were selected to explore to what extent social location influences the perceived importance of responsible AI values [36]. For all experiential and demographic questions, participants could choose not answer.

3.3 Participant recruitment

To examine how different groups assess responsible AI values, we surveyed three populations:

A US census-representative sample ($N=743_1$) was recruited by Qualtrics to gain insights into how the general population assesses the importance of responsible AI values. The recruitment process combined a variety of methods to minimize biases and performed stratified random sampling to match the US census along gender, age, race, region, and household income. Participant compensation was handled by Qualtrics.

A convenience US-based crowdworker sample (N_2=755) was recruited via the Clickworker crowdsourcing platform. Participants were US-based and likely previously contributed to the training of AI models by e.g., providing data labels. Each participant received USD 2.8 for a median participation time of 8 minutes. While crowdworkers are not directly involved in the AI development process, their judgments are often a key ingredient to machine learning systems. We explored whether their assessments could serve as proxies for the ethical intuition of a more representative population.

A sample of AI practitioners (N_3=175) was recruited through an open call on Twitter (N=156) and internal mailing lists (N=19) at a large tech company. Our call for participation targeted US-based participants whose work is related to AI/ML. We confirmed their background in the survey, but ultimately rely on self-reported expertise. For the internal mailing lists, we specifically targeted teams doing AI/ML related work. Participants could choose to enter a raffle to win one of five $50 gift vouchers after study completion. AI practitioners are a relevant population that makes key decisions throughout the AI development process. We explore whether their value judgments differ from those of the more general population.

We had to work with different types of compensation due to differences in respondent type and recruitment method across samples. However, we aimed to provide roughly commensurate compensation across recruitment methods. The study was IRB approved, and we obtained informed consent from all our participants.

3.4 Data quality control

To counterbalance ordering effects, the arrangement of scenarios, values, and conflict questions was randomized. In addition, the order of response options was randomly flipped per participant. For the conflict questions, we also randomized the internal order of the conflict, e.g. fairness vs. performance was inverted to performance vs. fairness. A pop-up window asked participants to slow down whenever they attempted to submit responses in under 3 seconds per survey page to deter spammers and inattentive participants. The four scenario introductions throughout the survey served as attention and comprehension checks for our participants. We removed all participants that had failed more than one attention check from our analysis to increase response quality, reducing the relevant samples to N_1=516, N_2=607, N_3=140 respectively.

4 RESULTS

4.1 What values are deemed as most important in general?

In Task 1, participants selected five values they deemed most important for AI systems out of an inventory of 12 responsible AI values (Figure 2). 76% of respondents from the US-census representative sample selected safety among the top 5 responsible AI values. Over 60% of participants in this representative panel also selected performance, privacy, and accountability among the most important values. Respondents from the crowdworker sample selected accountability less often, but their preferences were largely consistent with those from the US-census representative sample. AI practitioners' preferences were less focused. Compared to the US-census representative sample, practitioners selected humanist values such as fairness, inclusiveness, dignity, and solidarity more often and were less likely to select safety and performance among the most important values.

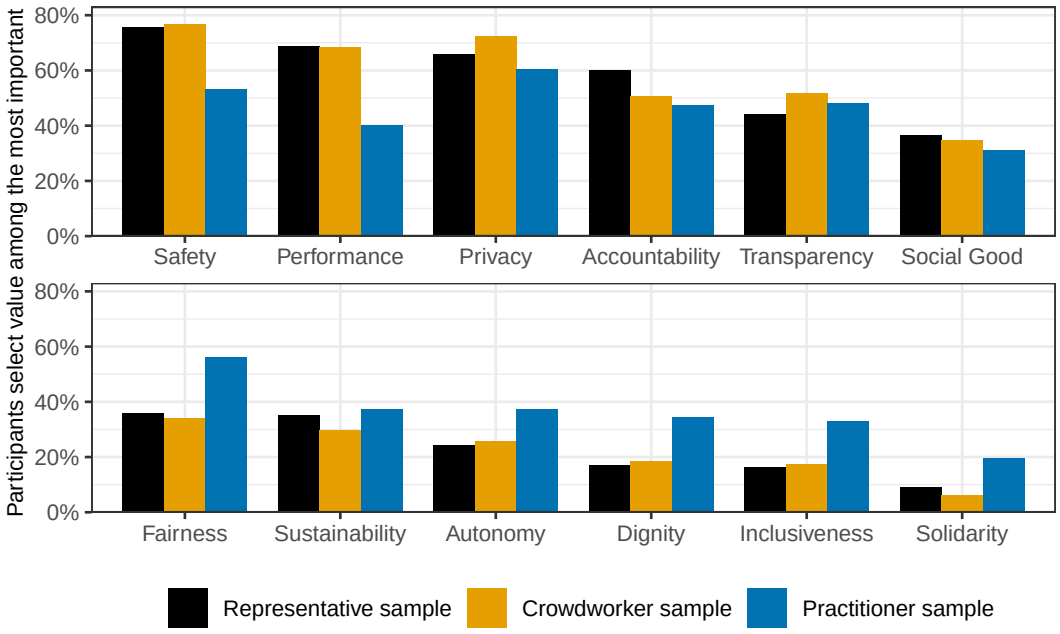

Figure 2: AI practitioners' value priorities differ from those of the general public. N_1=516, N_2=607, N_3=140. The x-axis shows the 12 responsible AI values respondents chose from, while the y-axis indicates how often respondents selected a value among the five most important. Participants from the US-census representative sample and the crowdworker sample selected safety, performance, and privacy most often among their five most important values, while practitioners selected fairness more often.

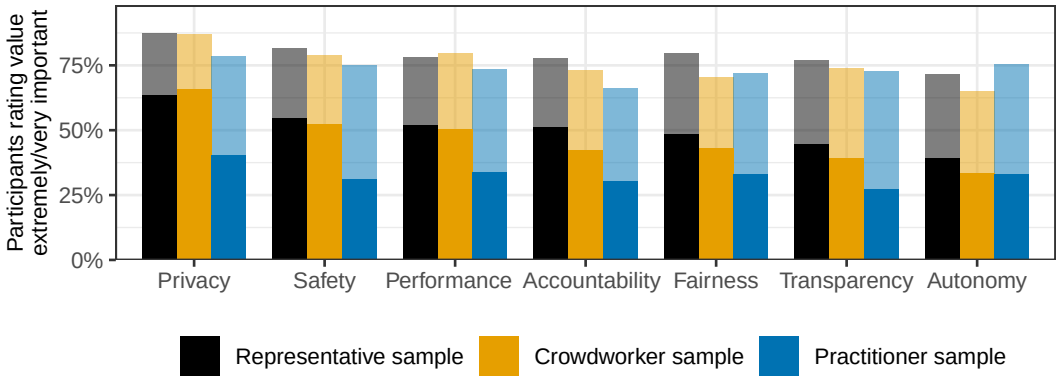

Figure 3: Representative participants rated responsible AI values as more important than AI practitioners did. N=140 to 607 ratings per bar. The x-axis shows the assessed responsible AI values and the y-axis indicates how often respondents evaluated the responsible AI value as very important (light) or extremely important (dark).

4.2 How important are values in specific deployment scenarios?

In Task 2 participants evaluated how important they considered a value in the context of a specific deployment scenario (Figures 3 and 4). The perceived importance of performance, accountability, fairness, and transparency varied significantly across deployment settings. In general responsible AI values were rated as very or extremely important. Compared to both the US-census representative and the crowdworker samples, on average, AI practitioners evaluated responsible AI values, and privacy, safety, and performance, in particular, as less important. We also observed significant variation of perceived importance across deployment settings, with responsible AI values being considered most important in the medical context and least important in the streaming context. A more

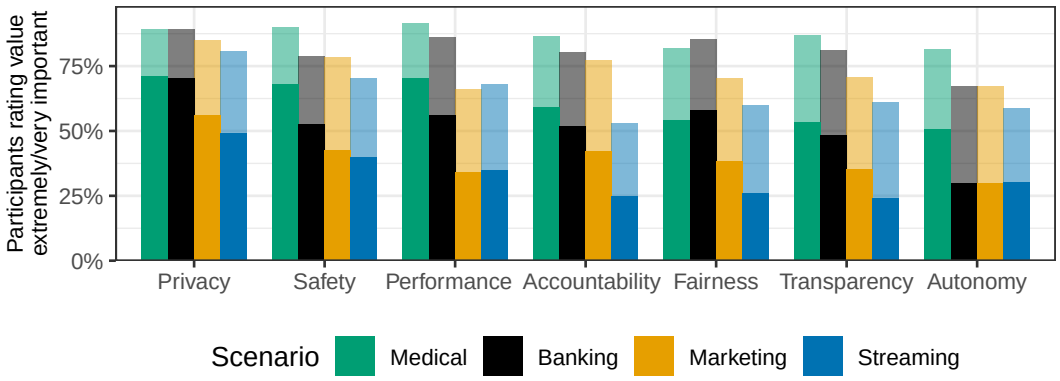

Figure 4: Responsible AI values were rated as most important in the medical and banking scenarios. N=287 to 344 ratings per bar aggregated across samples. The y-axis shows how often respondents evaluated the responsible AI value as very important (light) or extremely important (dark). The perceived importance of other values is dependent on the application context.

detailed graph showing responses by both sample and scenario is included in the Appendix A.6.

4.3 How values are prioritized when in conflict

In Task 3 participants suggested how values should be prioritized when in conflict (Figures 5 and 6). Respondents from all participant samples agreed on prioritizing safety over autonomy and transparency. Across scenarios, a majority of respondents agreed on prioritizing privacy over performance or fairness. Most disagreement was observed when performance and fairness conflicted: Participants from the US representative sample were almost equally split in their preferences for fairness versus performance. Crowdworkers were less likely to prioritize performance and AI practitioners were more likely to prioritize fairness than the US-census representative participants.

Across scenarios, respondents prioritized privacy over performance and fairness, and safety over autonomy and transparency. Again, the performance-fairness trade-off produced most variation: Participants prioritized performance in the medical and streaming scenario, and fairness in the banking and marketing scenario.

4.4 Demographics and experiential correlates of responsible AI value priority

To explore how demographic and experiential factors correlate with participants' assessments, we mapped their responses to a 5-Likert scale that preserves the direction of the original scale. Treating ordinal scales as interval scales is controversial, but the scales in our study have a unit of measurement with comparable-size intervals and a zero point, so a continuous analysis is meaningful and justifiable [41]. To examine how various demographic, experiential, or contextual factors may explain the variance in respondents' assessments, we used linear regression to build simple baseline models that predict their assessments.

Table 1 shows parameter estimates of linear regression models fitted to predict how important respondents consider a value in a specific scenario. The model constant corresponds to a white man from the US-census representative sample evaluating a responsible AI value in the banking scenario. The parameter estimates confirm that the perceived importance of values varies significantly across deployment settings. They also confirm that, compared to the US-census representative sample, AI practitioners evaluated most values as less important. Women and black respondents, on average, evaluated most responsible AI values as more important than other groups. Among the experiential correlates, a self-reported liberal political leaning was associated with a higher valuation of privacy. Self-reported experiences with discrimination predicted lower perceived importance of performance but were not statistically significantly correlated with other responsible AI values. While familiarity with ML did not predict different value priorities, respondents reporting to be familiar with UX research evaluated most responsible AI values as more important.

Table 2 shows parameter estimates predicting participants' preference in the case of conflicting responsible AI values. Positive coefficients correspond to a preference for the top value. Responses vary significantly by deployment context, but only the response to the fairness-performance trade-off varies by sample. Women respondents were more likely to prioritize safety over transparency than other groups, and black respondents were more likely to prioritize safety over autonomy. While participants reporting experiences of discrimination were more likely to prioritize fairness over privacy, they were not more likely to prioritize fairness over performance than other groups. Instead, participants with liberal political learning were more likely to prioritize fairness over performance and privacy than other groups. Familiarity with ML neither predicted a preference for performance over privacy nor fairness.

Some variables were correlated with each other. For example, the practitioner sample contains fewer women respondents ($r=-0.14$, $p<0.01$) and black respondents ($r=-0.11$, $p<0.01$), but more educated ($r=0.33$, $p<0.01$) and liberal-leaning ($r=0.2$, $p<0.01$) respondents. Similarly, liberal-leaning respondents were younger ($r=-0.13$, $p<0.01$) and more likely to report experiences with ML ($r=0.1$, $p<0.01$) and discrimination ($r=0.09$, $p<0.01$). However, a correlation analysis (included in the Appendix A.6) suggests that no covariates were

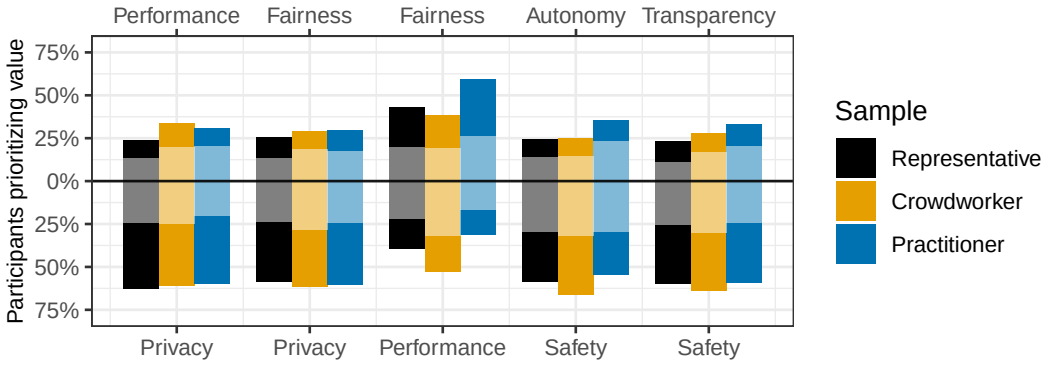

Figure 5: Participants across groups prioritized privacy and safety over fairness, but disagreed on the fairness vs. performance tradeoff. N = 104 to 607 ratings per bar. The conflicting value pairs are shown on the top and bottom, e.g., performance vs. privacy on the left. The proportion of respondents prioritizing the top value is shown to the top and the proportion of respondents prioritizing the bottom value to the bottom. Respondents expressing a strong preferences are shaded in dark, whereas weak preferences are lightly shaded. Undecided respondents are omitted.

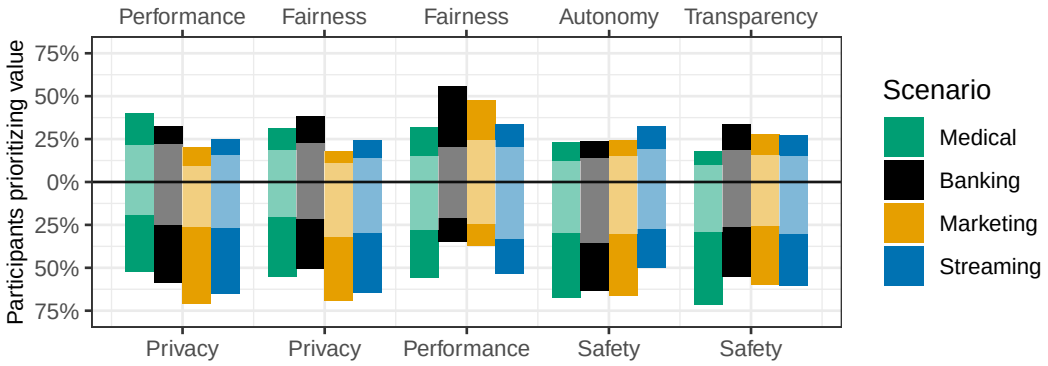

Figure 6: Value priorities vary by context, but most participants prioritized privacy and safety across most scenarios. N = 276 to 341 ratings per bar aggregated across samples. The proportion of respondents prioritizing the top value are shown to the top and the proportion of respondents prioritizing the bottom value to the bottom. Respondents expressing a strong preferences are shaded in dark, weak preferences are lightly shaded.

highly correlated (r>0.7). The variance inflation factor remained below 1.5 across all covariates, indicating little to no multicollinearity issues [29].

5 DISCUSSION

AI practitioners' value priorities for responsible AI differ from those of the general public. Our results empirically corroborate a commonly raised concern: AI practitioners' value preferences for responsible AI are not representative of the value priorities of the wider US population. Compared to a US-census representative public, AI practitioners evaluated responsible AI values as less important in general and emphasized a different set of values.

US-census representative and crowdworker respondents agreed on what values they found most important: safety, privacy, and performance. Practitioners, in comparison, were more likely to prioritize fairness, dignity, and inclusiveness.

These findings align with prior research finding that different groups have different normative expectations of how AI systems should behave in specific situations [2, 28, 31, 55]. Our findings extend prior work by demonstrating how AI practitioners' ethical preferences differ from other groups'. We also show that groups not only differ in their judgment of specific behaviors and technical details, but may disagree on the importance of the very values at the core of responsible AI. *The disagreement in value priorities highlights the importance of paying attention to who gets to define what constitutes "ethical" or "responsible" AI.* Responsible AI guidelines [37] may emphasize a different set of values depending on who writes them and who is consulted. We hypothesize that consulting populations outside the Western world about their priorities for responsible AI would surface even starker disagreement about the values underlying responsible AI [39, 63].

Table 1: Regression analysis with simple baseline models predicting the value importance ratings based on scenario, sample, and demographic correlates. The constant corresponds to a white male respondent from the US-census representative sample assessing a value in the banking scenario. Bold text indicates statistical significance.

	Dependent variable:						
	Value importance rating						
	Privacy	Safety	Perform.	Account.	Fairness	Transp.	Autonomy
Marketing system	**-0.051****	-0.024	**-0.14*****	**-0.046***	**-0.10*****	**-0.08*****	-0.004
Medical system	0.005	**0.06*****	**0.044***	0.030	-0.031	0.029	**0.10*****
Streaming system	**-0.08*****	**-0.09*****	**-0.12*****	**-0.18*****	**-0.18*****	**-0.16*****	-0.036
Crowdworker sample	0.005	-0.020	0.002	**-0.05*****	**-0.06*****	**-0.032***	**-0.041***
Practitioner sample	**-0.08*****	**-0.057***	-0.052	**-0.10*****	**-0.057***	**-0.09*****	0.005
Women respondents	**0.04*****	**0.039****	**0.05*****	**0.04***	**0.05*****	**0.028***	**0.07*****
Gender-diverse resp.	-0.042	-0.031	-0.046	-0.017	0.086	-0.010	0.041
Black respondents	**0.046***	**0.052***	**0.062****	0.006	**0.08*****	0.019	0.031
Hispanic respondents	0.024	0.042	0.044	0.034	0.020	0.021	0.023
Asian respondents	0.001	-0.025	-0.017	-0.018	-0.031	**-0.057***	-0.007
Age	0.001	-0.0001	-0.001	-0.0004	-0.001	-0.0003	0.001
Education	-0.020	-0.047	-0.026	0.009	0.007	0.012	0.013
Political leaning	**0.060***	0.011	0.027	0.023	**0.056***	0.049	0.006
Exp. with discrimination	-0.027	0.011	**-0.057***	0.002	0.004	-0.008	0.005
Familiarity with ML	-0.037	-0.007	0.005	-0.028	-0.016	-0.011	0.009
Familiarity with UX	**0.046***	0.043	**0.053***	0.034	**0.055***	**0.057***	0.043
Constant	**0.82*****	**0.80*****	**0.83*****	**0.82*****	**0.81*****	**0.77*****	**0.62*****
Observations	1,246	1,246	1,246	1,246	1,246	1,246	1,246
R^2	0.082	0.084	0.150	0.130	0.119	0.111	0.070
Adjusted R^2	0.070	0.072	0.139	0.118	0.107	0.099	0.058
Residual Std. Error	0.212	0.233	0.229	0.238	0.244	0.242	0.254
F Statistic	6.84***	7.05***	13.51***	11.42***	10.33***	9.58***	5.81***
Note:							*p<0.05; **p<0.01; ***p<0.001

What might explain the differences in value priorities between AI practitioners' and other groups? Our results provide limited insight into plausible drivers of differences in values. First, women and black respondents assessed responsible AI as more important than other demographic groups. Their relatively low representation in the AI practitioner sample compared to the representative sample (only 40% and 2.2% compared to 52% and 15% respectively) explains about 15% of the lower importance ratings AI practitioners assigned to values in general. Increasing the representation of e.g., women and black researchers in AI [13, 32, 42] may thus result in responsible AI values receiving more attention.

Another demographic variable that robustly predicted differences in value preferences was respondents' political leaning. Liberal-leaning respondents were 10% more likely to select fairness amongst the most important values than conservatives, and were 15.5% more likely to prioritize fairness in the fairness-performance trade-off. Compared to the representative sample, AI practitioner respondents were substantially more likely to self-identify as liberal-leaning (52% compared to 26%), explaining about 27% of practitioners' different evaluation of fairness. This result is in line with the broader research on value differences along ideological lines [9, 68]. It highlights that guidelines for responsible AI need to navigate a polarized value landscape.

Other demographic and experiential variables, however, were less predictive of how our participants assessed responsible AI values. Respondents reporting experience with discrimination were more likely to prioritize fairness over privacy, but did not evaluate fairness as more important than other groups. When asked whether developers should prioritize fairness over performance, participants from minoritized groups and participants reporting experience with discrimination were as undecided as other groups. While previous work identified performance as the central value in machine learning research [6], our results do not suggest that AI practitioners or respondents familiar with machine learning were more likely to value performance. Participants trained in user experience research, however, evaluated responsible AI values more important in general.

Can AI practitioners use crowdsourcing to complement their ethical intuitions in the development process? Our findings emphasize the need for bringing in a diversity of perspectives when decisions are made about the development and operationalization of responsible

Table 2: Regression analysis with simple baseline models predicting the value preference ratings based on scenario, sample, and demographic correlates. The constant corresponds to a white male respondents from the US-census representative sample recommending a value prioritization in the banking scenario. Bold text indicates statistical significance.

	Dependent variable:				
	Value preference rating				
	Privacy. vs. performance	Privacy vs. fairness	Performance vs. fairness	Safety vs. autonomy	Safety. vs. transparency
Marketing system	**0.164****	**0.301*****	**0.113***	0.067	0.080
Medical system	**-0.112***	**0.113***	**0.378*****	0.078	**0.259*****
Streaming system	0.091	**0.209*****	**0.342*****	**-0.150****	0.086
Crowdworker sample	-0.079	0.048	**0.152*****	0.058	0.015
Practitioner sample	-0.060	0.114	-0.091	-0.078	0.062
Women respondents	0.055	0.038	0.057	0.076	**0.113****
Gender-diverse resp.	0.219	-0.128	-0.182	0.060	-0.214
Black respondents	-0.006	-0.098	-0.109	**0.168****	-0.029
Hispanic respondents	-0.050	**-0.160***	0.082	-0.017	-0.021
Asian respondents	-0.033	0.024	-0.113	0.033	0.020
Age	0.001	**0.003***	-0.002	0.0002	0.0002
Education	0.104	-0.126	-0.067	0.045	-0.063
Political leaning	0.010	**-0.191***	**-0.226****	0.091	-0.017
Exp. with discrimination	-0.113	**-0.205****	0.023	**-0.160***	0.083
Familiarity with ML	0.008	0.073	-0.008	0.090	-0.078
Familiarity with UX	-0.093	0.124	0.026	0.048	0.016
Constant	**0.262***	0.100	-0.057	0.102	0.159
Observations	1,246	1,246	1,246	1,246	1,246
R^2	0.032	0.056	0.080	0.038	0.031
Adjusted R^2	0.019	0.043	0.068	0.025	0.018
Residual Std. Error	0.712	0.679	0.701	0.665	0.688
F Statistic	2.535***	4.526***	6.719***	3.009***	2.459**
Note:					*p<0.05; **p<0.01; ***p<0.001

AI. Crowdworkers are often the go-to convenience sample, but to what extent could they provide a reliable lens into the values that a broader population expect AI systems to adhere to?

As in prior research [34], we find that the value priorities of crowdworkers largely align with those of the US-census representative sample. Our results also show that often a majority of participants agreed on value trade-offs. For example, respondents from all samples prioritized privacy over performance across all deployment scenarios. The agreement raises the question of whether and when product teams could use such results to e.g., justify prioritizing privacy over performance.

Here, consensus alone may not justify practical requirements within specific contexts of use. Rather than providing definite answers, the approach developed in this paper provides "values levers" [65]: organizational processes that take the implicit work of value judgments in technology development and transform it into an explicit matter of debate and documentation. Empirical data on different groups' preferences can both inform the development process of responsible AI and provide opportunities for critical reflection. Rather than prescribing value priorities, responsible AI guidelines could ask practitioners to justify their choices whenever they go against commonly held value preference.

5.1 Limitations

The quantitative approach to value elicitation explored above has its benefits: It allows consulting large and representative samples of stakeholders and integrates well with existing crowdwork infrastructures. At the same time, it needs to be complemented by qualitative, small-n investigations like interviews or focus groups for a comprehensive understanding of value differences across social groups. For example, the current study did not explore how groups understand or interpret values differently, what other values some groups might have wanted to include, or why it is that e.g. women, on average, rated responsible AI values as more important.

The results of this survey also should be interpreted with care. No normative "ought" can be derived from a descriptive "is" [53]. We cannot conclude that safety ought to be prioritized over autonomy from the observation that the respondents in our samples suggested so. Our results aim to increase the context sensitivity of

responsible AI decisions, not to prescribe a specific course of action. Empirical ethical research does not replace ethical reasoning but offers perspectives and critical reflections.

Finally, knowledge-dependent tensions arise when contrasting the perspectives of experts and laypeople. One may argue that non-expert perspectives lack the technical and organizational insight required to evaluate AI systems. However, as we are focusing on ethical rather than technical questions, non-experts have their own valid and legitimate forms of knowledge [30] that experts might not be aware of.

6 CONCLUSION

Recently published guidelines for responsible AI seem to converge on a set of central values. However, little is known about the values a more representative public would find important for responsible AI. We conducted a survey comparing how US-representative respondents, crowdworkers, and AI practitioners perceive and prioritize responsible AI values. Our findings show that, compared to the general public, AI practitioners find responsible AI values less important and are likely to focus on a different set of values. Our findings underline the need for more diverse ethical judgement to be incorporated into the AI development process. Crowdworkers, who are already involved in the AI development process, resemble the general public in their value priorities and might provide valuable input.

ACKNOWLEDGMENTS

We thank our colleagues from across Microsoft who provided insight and expertise that greatly assisted the research. We are particularly grateful to Su Lin Blodgett, Stephanie Ballard, Michael Madaio, Emery Fine and Kate Crawford for their comments on the research framing and survey design.

REFERENCES

[1] Philip Agre and Philip E Agre. 1997. *Computation and human experience*. Cambridge University Press.
[2] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature* 563, 7729 (2018), 59–64.
[3] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. 2019. Differential privacy has disparate impact on model accuracy. *Advances in Neural Information Processing Systems* 32 (2019), 15479–15488.
[4] Solon Barocas and Danah Boyd. 2017. Engaging the ethics of data science in practice. *Commun. ACM* 60, 11 (2017), 23–25.
[5] Reuben Binns. 2018. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*. PMLR, 149–159.
[6] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2021. The Values Encoded in Machine Learning Research. *arXiv preprint arXiv:2106.15590* (2021).
[7] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of Bias in NLP. *arXiv preprint arXiv:2005.14050* (2020).
[8] Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. 2020. Overcoming Failures of Imagination in AI Infused System Development and Deployment. *arXiv preprint arXiv:2011.13416* (2020).
[9] Valerie Braithwaite. 1998. The value orientations underlying liberalism-conservatism. *Personality and individual differences* 25, 3 (1998), 575–589.
[10] Claudia Cappelli, Herbet Cunha, Bruno Gonzalez-Baixauli, and Julio Cesar Sampaio do Prado Leite. 2010. Transparency versus security: early analysis of antagonistic requirements. In *Proceedings of the 2010 ACM symposium on applied computing*. 298–305.
[11] Stephen Cave, Claire Craig, Kanta Dihal, Sarah Dillon, Jessica Montgomery, Beth Singler, and Lindsay Taylor. 2018. Portrayals and perceptions of AI and why they matter. (2018).
[12] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*. 797–806.
[13] Kate Crawford. 2016. Artificial intelligence's white guy problem. *The New York Times* 25, 06 (2016).
[14] Zachary Davis and Anthony Steinbock. 2021. Max Scheler. In *The Stanford Encyclopedia of Philosophy* (Fall 2021 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.
[15] Deloitte. 2020. *Bringing Transparency and Ethics into AI?* https://perma.cc/8LPD-JN74
[16] Edward C Dillon Jr, Juan E Gilbert, Jerlando FL Jackson, and LJ Charleston. 2015. The state of African Americans in computer science-the need to increase representation. *Computing Research News* 21, 8 (2015), 2–6.
[17] Cem Dilmegani. 2018. *100 AI Use Cases and Applications*. https://perma.cc/6A78-PTRJ
[18] Roel Dobbe, Sarah Dean, Thomas Gilbert, and Nitin Kohli. 2018. A broader view on bias in automated decision-making: Reflecting on epistemology and dynamics. *arXiv preprint arXiv:1807.00553* (2018).
[19] Michael Dunn, Mark Sheehan, Tony Hope, and Michael Parker. 2012. Toward methodological innovation in empirical ethics research. *Cambridge Quarterly of Healthcare Ethics* 21, 4 (2012), 466–480.
[20] Michael D Ekstrand, Rezvan Joshaghani, and Hoda Mehrpouyan. 2018. Privacy for all: Ensuring fair and equitable privacy protections. In *Conference on Fairness, Accountability and Transparency*. PMLR, 35–47.
[21] Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.
[22] Baruch Fischhoff. 1991. Value elicitation: Is there anything in there? *American psychologist* 46, 8 (1991), 835.
[23] Batya Friedman. 1996. Value-sensitive design. *interactions* 3, 6 (1996), 16–23.
[24] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)* 14, 3 (1996), 330–347.
[25] Manuela Fumagalli, Roberta Ferrucci, Francesca Mameli, Sara Marceglia, Simona Mrakic-Sposta, Stefano Zago, Claudio Lucchiari, D Consonni, F Nordio, Gabriella Pravettoni, et al. 2010. Gender-related differences in moral judgments. *Cognitive processing* 11, 3 (2010), 219–226.
[26] Google. 2020. *Our Principles*. https://perma.cc/VHZ5-DJJJ
[27] Jesse Graham, Peter Meindl, Erica Beall, Kate M Johnson, and Li Zhang. 2016. Cultural differences in moral judgment and behavior, across and within societies. *Current Opinion in Psychology* 8 (2016), 125–130.
[28] Nina Grgic-Hlaca, Elissa M Redmiles, Krishna P Gummadi, and Adrian Weller. 2018. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 world wide web conference*. 903–912.
[29] Joseph F Hair. 2009. Multivariate data analysis. (2009).
[30] Sandra Harding. 1992. Rethinking standpoint epistemology: What is" strong objectivity?". *The Centennial Review* 36, 3 (1992), 437–470.
[31] César A Hidalgo, Diana Orghian, Jordi Albo Canals, Filipa De Almeida, and Natalia Martin. 2021. *How humans judge machines*. MIT Press.
[32] White House. 2016. Preparing for the future of artificial intelligence. Executive Office of the President National Science and Technology Council. Committee on Technology.
[33] Yiqing Hua, Armin Namavari, Kaishuo Cheng, Mor Naaman, Thomas Ristenpart, and Cornell Tech. 2021. Increasing Adversarial Uncertainty to Scale Private Similarity Testing. *arXiv preprint arXiv:2109.01727* (2021).
[34] Connor Huff and Dustin Tingley. 2015. "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research & Politics* 2, 3 (2015), 2053168015604648.
[35] IBM. 2020. *IBM's Principles for Trust and Transparency*. https://perma.cc/8LPD-JN74
[36] Kristen Intemann. 2010. 25 years of feminist empiricism and standpoint theory: Where are we now? *Hypatia* 25, 4 (2010), 778–796.
[37] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.
[38] Anna Jobin, Kingson Man, Antonio Damasio, Georgios Kaissis, Rickmer Braren, Julia Stoyanovich, Jay J Van Bavel, Tessa V West, Brent Mittelstadt, Jason Eshraghian, et al. 2021. AI reflections in 2020. *Nature Machine Intelligence* 3, 1 (2021), 2–8.
[39] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. "Because AI is 100% right and safe": User Attitudes and Sources of AI Authority in India. In *CHI Conference on Human Factors in Computing Systems*. 1–18.
[40] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
[41] Thomas R Knapp. 1990. Treating ordinal scales as interval scales: an attempt to resolve the controversy. *Nursing research* 39, 2 (1990), 121–123.

[42] Liana Christin Landivar. 2013. Disparities in STEM employment by sex, race, and Hispanic origin. *Education Review* 29, 6 (2013), 911–922.
[43] Min Kyung Lee and Su Baykal. 2017. Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing*. 1035–1048.
[44] Sonia Livingstone, Leslie Haddon, Anke Görzig, and Kjartan Ólafsson. 2011. Risks and safety on the internet. *The perspective of European children. Full findings and policy implications from the EU Kids Online survey of* (2011), 9–16.
[45] Helen E Longino. 2020. *Science as social knowledge*. Princeton university press.
[46] Alasdair MacIntyre. 1981. The nature of the virtues. *Hastings Center Report* (1981), 27–34.
[47] Bertram F Malle, Matthias Scheutz, Thomas Arnold, John Voiklis, and Corey Cusimano. 2015. Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 117–124.
[48] Kirsten Martin. 2019. Ethical implications and accountability of algorithms. *Journal of Business Ethics* 160, 4 (2019), 835–850.
[49] Ronald Meijer, Peter Conradie, and Sunil Choenni. 2014. Reconciling contradictions of open data regarding transparency, privacy, security and trust. *Journal of theoretical and applied electronic commerce research* 9, 3 (2014), 32–44.
[50] Microsoft. 2020. *Responsible AI*. https://perma.cc/7AKE-3GH3
[51] Brent Mittelstadt. 2019. AI Ethics–Too principled to fail. *arXiv preprint arXiv:1906.06668* (2019).
[52] University Montreal. 2017. *The Montreal Declaration for a Responsible Development of Artificial Intelligence*. https://perma.cc/8LPD-JN74
[53] Albert W Musschenga. 2005. Empirical ethics, context-sensitivity, and contextualism. *The Journal of medicine and philosophy* 30, 5 (2005), 467–490.
[54] Priyanka Nanayakkara, Jessica Hullman, and Nicholas Diakopoulos. 2021. Unpacking the Expressed Consequences of AI Research in Broader Impact Statements. *arXiv preprint arXiv:2105.04760* (2021).
[55] Emma Pierson. 2017. Demographics and discussion influence views on algorithmic fairness. *arXiv preprint arXiv:1712.09124* (2017).
[56] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. *arXiv preprint arXiv:1709.02012* (2017).
[57] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 33–44.
[58] Kristina Rolin. 2006. The bias paradox in feminist standpoint epistemology. *Episteme* 3, 1-2 (2006), 125–136.
[59] Lee Ross, David Greene, and Pamela House. 1977. The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of experimental social psychology* 13, 3 (1977), 279–301.
[60] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. 2019. How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 99–106.
[61] Shalom H Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*. Vol. 25. Elsevier, 1–65.
[62] Shalom H Schwartz. 1994. Are there universal aspects in the structure and contents of human values? *Journal of social issues* 50, 4 (1994), 19–45.
[63] Shalom H Schwartz. 2007. Basic human values: Theory, measurement, and applications. *Revue française de sociologie* 47, 4 (2007), 929.
[64] Phoebe Sengers, Kirsten Boehner, Shay David, and Joseph'Jofish' Kaye. 2005. Reflective design. In *Proceedings of the 4th decennial conference on Critical computing: between sense and sensibility*. 49–58.
[65] Katie Shilton. 2013. Values levers: Building ethics into design. *Science, Technology, & Human Values* 38, 3 (2013), 374–397.
[66] Reza Shokri and Vitaly Shmatikov. 2015. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*. 1310–1321.
[67] European Union. 2020. *Ethics guidelines for trustworthy AI*. https://perma.cc/8LPD-JN74
[68] Geoffrey A Wetherell, Mark J Brandt, and Christine Reyna. 2013. Discrimination across the ideological divide: The role of value violations and abstract values in discrimination by liberals and conservatives. *Social Psychological and Personality Science* 4, 6 (2013), 658–667.
[69] Jess Whittlestone, Rune Nyrup, Anna Alexandrova, and Stephen Cave. 2019. The role and limits of principles in AI ethics: towards a focus on tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 195–200.
[70] Terry Winograd, Fernando Flores, and Fernando F Flores. 1986. *Understanding computers and cognition: A new foundation for design*. Intellect Books.
[71] Alison Wylie, Robert Figueroa, and Sandra Harding. 2003. Why standpoint matters. *Science and other cultures: Issues in philosophies of science and technology* 26 (2003), 48.

A APPENDIX: SUPPLEMENTARY MATERIALS

A.1 Introduction and task

Artificial intelligence (AI) is a set of emerging technologies concerned with building smart systems or machines capable of performing tasks that typically require human intelligence. Besides technical challenges, building AI systems involves complex decision-making on what the system should or should not do. In this survey, we will ask you to assess the importance of ethical principles for four AI systems.

A.2 Value Description and Question Framing

RAI value	Description
Transparency	A transparent AI system produces decisions that people can understand. Developers of transparent AI systems ensure, as far as possible, that users can get insight into why and how a system made a decision or inference. How important is it that the system is transparent?
Fairness	A fair AI system treats all people equally. Developers of fair AI systems ensure, as far as possible, that the system does not reinforce biases or stereotypes. A fair system works equally well for everyone independent of their race, gender, sexual orientation, and ability. How important is it that the system is fair?
Safety	A safe AI system performs reliably and safely. Developers of safe AI systems implement strong safety measures. They anticipate and mitigate, as far as possible, physical, emotional, and psychological harms that the system might cause. How important is it that the system is safe?
Accountability	An accountable AI system has clear attributions of responsibilities and liability. Developers and operators of accountable AI systems are, as far as possible, held responsible for their impacts. An accountable system also implements mechanisms for appeal and recourse. How important is it that the system is accountable?
Privacy	An AI system that respects people's privacy implements strong privacy safeguards. Developers of privacy-preserving AI systems minimize, as far as possible, the collection of sensitive data and ensure that the AI system provides notice and asks for consent. How important is it that the system respects people's privacy?
Autonomy	An AI system that respects people's autonomy avoids reducing their agency. Developers of autonomy-preserving AI systems ensure, as far as possible, that the system provides choices to people and preserves or increases their control over their lives. How important is it that the system respects people's autonomy?
Performance	A high-performing AI system consistently produces good predictions, inferences or answers. Developers of high-performing AI systems ensure, as far as possible, that the system's results are useful, accurate and produced with minimal delay. How important is it that the system performs well?

A.3 Value conflict framing

Value pair	Description
Fairness vs. performance	The developers realize that making the system treat all people equally (ensuring fairness) may make the system's predictions less accurate (reducing performance). Should they prioritize fairness or performance?
Fairness vs. performance (reverse)	The developers realize that making the system's predictions possibly accurate (ensuring performance) may mean that the system cannot treat all people equally (reducing fairness). Should they prioritize performance or fairness?
Fairness vs. privacy	The developers realize that making the system treat all people equally (ensuring fairness) may require the collection of additional sensitive data (reducing privacy). Should they prioritize fairness or privacy?
Fairness vs. privacy (reverse)	The developers realize that minimizing the collection of sensitive data (ensuring privacy) may mean that the system cannot treat all people equally (reducing fairness). Should they prioritize privacy or fairness?
Privacy vs. performance	The developers realize that minimizing the collection of sensitive data (ensuring privacy) may make the system's predictions less accurate (reducing performance). Should they prioritize privacy or performance?
Privacy vs. performance (reverse)	The developers realize that making the system's predictions possibly accurate (ensuring performance) may require the collection of additional sensitive data (reducing privacy). Should they prioritize performance or privacy?
Safety vs. autonomy	The developers realize that mitigating risks and potential harms (ensuring safety) may require limiting people's choices and control (reducing autonomy). Should they prioritize safety or people's autonomy?
Safety vs. autonomy (reverse)	The developers realize that giving people choices and control (ensuring autonomy) may introduce additional risks and potential harms (reducing safety). Should they prioritize people's autonomy or safety?
Safety vs. transparency	The developers realize that mitigating risks and potential harms (ensuring safety) may require to keep the system's decision process opaque (reducing transparency). Should they prioritize safety or transparency?
Safety vs. transparency (reverse)	The developers realize that revealing the system's decision process (ensuring transparency) may introduce additional risks and potential harms (reducing safety). Should they prioritize transparency or safety?

A.4 Application scenario framing

Scenario	Description
Banking	A bank uses an AI system that scans loan applicants' data to predict whether they are likely to repay a loan. Thousands of loan applications are automatically rejected based on the output of this AI system.
Medical	A medical clinic uses an AI system that scans patients' medical records to predict whether a patient has a particular disease. Thousands of patients' treatment plans are automatically adjusted based on the output of this AI system.
Marketing	A marketing company uses an AI system that scans the data of web users to predict which advertisements they will respond to. Thousands of advertisements are automatically shown to users based on the output of this AI system.
Streaming	A video streaming company uses an AI system that scans users' data to predict which other movies they would enjoy seeing. A list of recommended movies is automatically shown to thousands of users based on the output of this AI system.

A.5 Detailed result graphs

Please refer to Figures 7 and 8.

A.6 Covariate correlation analysis

	Crowd-workers	Practi-tioners	Women resp.	Diverse resp.	Black resp.	Hispanic resp.	Asian resp.	Age	Edu-cation	Pol. lean.	Dis-crimin.	Fam. ML
Women respondents	0.01	-0.14**										
Gender-diverse resp.	0.07*	-0.01	-0.15**									
Black respondents	0.01	-0.11**	0.05	-0.05								
Hispanic respondents	-0.01	-0.08**	-0.02	0.05	-0.02							
Asian respondents	0.09**	0.06*	-0.05	0.04	-0.09**	-0.07*						
Age	-0.24**	-0.14**	0.05	-0.10**	-0.04	-0.07*	-0.16**					
Education	0.06*	0.33**	-0.09**	0.00	-0.09**	-0.07*	0.12**	0.05				
Political leaning	0.03	0.20**	-0.05	0.15**	-0.01	0.04	0.05	-0.13**	0.17**			
Exp. w. discrimination	0.01	0.14**	0.04	0.15**	0.16**	0.05	0.07*	-0.14**	0.08**	0.09**		
Familiarity with ML	-0.02	0.23**	-0.08**	0.02	-0.01	-0.01	0.13**	-0.16**	0.22**	0.10**	0.18**	
Familiarity with UX	0.10**	0.05	-0.01	-0.01	0.06*	-0.02	0.09**	-0.13**	0.18**	0.01	0.19**	0.42**

Note: *p<0.05; **p<0.01

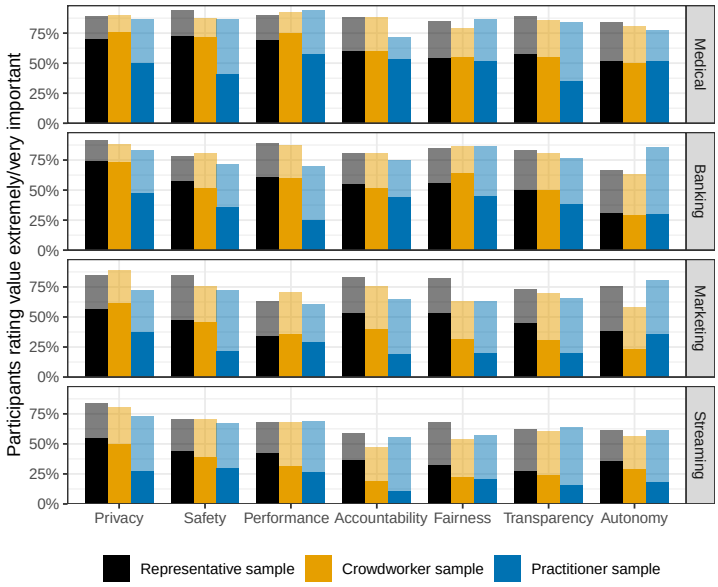

Figure 7: The perceived importance of values across deployment scenarios. N=28 to 171 ratings per bar. The x-axis shows the assessed responsible AI values and the y-axis indicates how often respondents evaluated the responsible AI value as very important (light) or extremely important (dark).

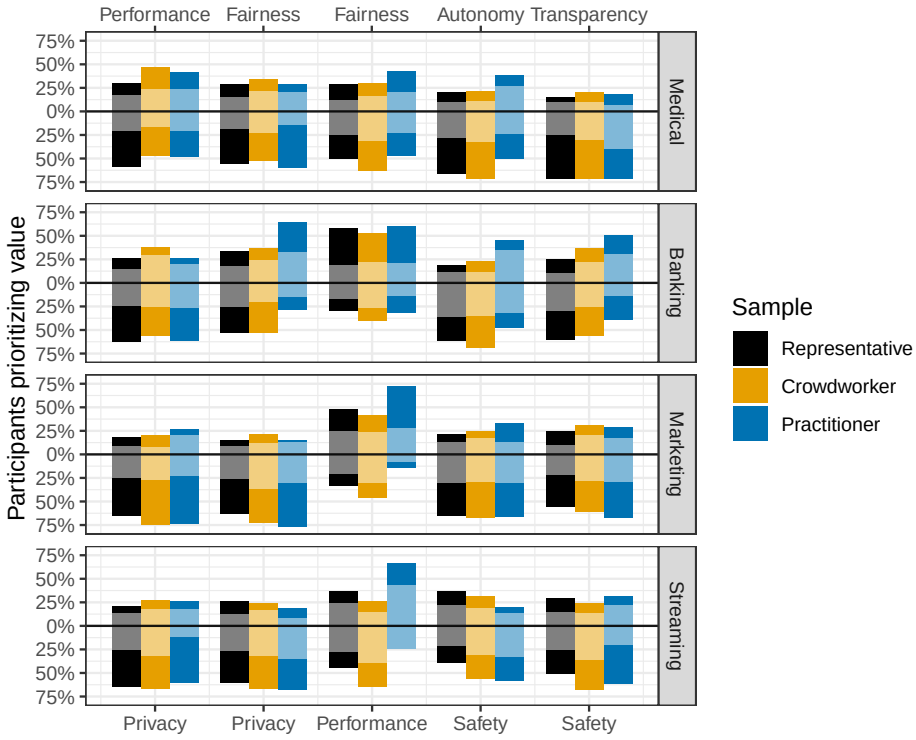

Figure 8: How values are prioritized in different deployment settings. N = 28 to 173 ratings per bar. The conflicting value pairs are shown on the top and bottom, e.g., privacy vs. performance on left. Respondents prioritizing the top value are shown to the top and responses prioritizing the bottom value to the bottom. Respondents expressing a strong preferences are shaded in dark, whereas weak preferences are lightly shaded. Undecided respondents are omitted.