Studying People to Study AI: Expert Perspectives on the Epistemic Fit and Barriers of Human Research in AI Safety & Ethics

Anonymous submission

Abstract

Safety risks of AI are becoming increasingly evident in human interactions with AI technologies. The prominent approaches to evaluate these risks favor more technical methods, such as model benchmarks and LLM simulations, overlooking empirical research methods, despite the latter's established role in grounding claims about real-world harm. To examine this trend, we conduct an expert survey ($n = 93$) and expert interviews ($n = 17$) with AI Safety & Ethics (AISE) researchers from *Technical*, *Sociotechnical*, *Governance*, and *Normative* backgrounds. Our findings suggest that although there is a consensus that human research is valuable for generating evidence for AISE, its adoption and acceptance are constrained by perceived validity issues, tangible resource barriers, epistemic and personal preferences in methods, and infrastructural constraints from the broader research community. In particular, *Technical* researchers tend to *value* human research less and *collaborate* across disciplines less. We propose recommendations for establishing the epistemic fit of human research within AISE and bridging the prohibitive limitations that researchers face.

Introduction

Research on the safe and ethical design and deployment of AI systems is a multi-faceted, fast-evolving, and critical area of AI research, encompassing a spectrum of disciplines, both technical and sociotechnical in nature (Leslie 2019; Hendrycks 2025). As the boundaries of the research community are continuously evolving, two dominant yet divisive camps within the broader field emerge as AI Safety (AIS) and AI Ethics (AIE) (Gyevnár and Kasirzadeh 2026). In this paper, we group them together with other adjacent disciplines, referring to the broader community as *AISE* for their shared concern about the impacts of AI on humans and society. And yet an interesting paradox remains unaddressed: despite AISE's mission to anticipate and prevent harm to humans, empirical human-centred research methodologies are underutilized in the field.

Specifically, we highlight that AISE research involving any human subjects is rare within published literature (Weidinger et al. 2023), being outnumbered by technical or normative approaches like LLMs simulations of human reactions to AI outputs (Li et al. 2026; Ni et al. 2026), static benchmarks with binary labels for harm (Reuel et al. 2024b; Yu et al. 2026), and AI constitutions guided by assumptions about normative value alignments (Bai et al. 2022). These approaches sacrifice construct validity for cost and time efficiency, risking the failure to capture the desired safety construct (Bean et al. 2026). Therefore, the evidence of harm for AISE is not well-substantiated if researchers rely on LLM simulations, normative assumptions, or incidental case studies. Research performed with human participants can fill the evidence gap (Weidinger et al. 2023). While there is a recent uptick in human subject research around interactional and experiential harms of AI — like large-scale experiments on emotional reliance (Kirk et al. 2025; Fang et al. 2025) and sycophancy (Cheng et al. 2026; Rathje et al. 2025) — the extent to which that human methods fit into the evolving epistemology of AISE is still unclear. The International AI Safety Report (Bengio et al. 2025), a review of AI risks for governance decision-making, calls for better evidence via observational records of the scale of harm from deepfakes and experimental studies on the impact of AI manipulation.

While past works (Gyevnár and Kasirzadeh 2026; Roytburg and Miller 2025) illuminated differences between research communities within AISE via literature reviews, we focus on the researchers themselves to understand: (**RQ1**) *across AISE disciplines, what do experts see as the epistemic fit of human research?*; and (**RQ2**) *What are the barriers that AISE researchers face in performing human research?* Importantly, while literature surveys analyse published research, we directly capture experts' motivations and deterrents behind *how* the research is pursued, or *why* it's not (Agapie, Haldar, and Poblete 2022; Paskov et al. 2026). **To understand these lived experiences, we surveyed ($n = 93$) and interviewed ($n = 17$) experts from *Technical*, *Sociotechnical*, *Governance*, and *Normative* disciplines who are actively contributing to the field.**

We find that while experts across disciplines agree that more evidence of AI's impact on humans is needed, and that technical-only research is often plagued with construct validity issues, they hold varied opinions on which human methods generate valid evidence. These differences are largely driven by epistemic incompatibility and low familiarity with human subject methodologies. In particular, *Technical* researchers are more likely to undervalue the contributions of human methods, and to collaborate across disciplines less frequently. Qualitative and mixed-methods approaches stand out as an underutilized yet highly valued

means of *explaining* mechanisms of harm, which challenges the quantitative leaning of AISE. In terms of barriers, participants are foremost impacted by a lack of tangible research resources like time, funding, and participant access. Many also report tensions arising from their mentors, institutions, and sectors. In the following sections, we present relevant literature, describe our data collection and analysis methods, detail our results, and discuss the implications for cross-disciplinary and cross-sector collaboration — including that AISE community should critically, not performatively, incorporate rigorous human research to generate evidence for harms from AI.

Related Works

We position our contributions in relation to a body of computing research that concerns epistemic divides in multi-disciplinary areas, the evolution of AISE and its sub-communities, and the problems where human research has been applied (or has the opportunity to be) in AISE.

Epistemic Disciplinary Divides

AISE is far from the first research community to grapple with the tensions inherent in uniting practitioners from disparate disciplinary traditions. Parallel challenges have unfolded in computational healthcare (Agapie, Haldar, and Poblete 2022), environmental science (Miller et al. 2008), and other complex, problem-driven fields that present tensions in the choice of methods to use, frameworks to ground in, and problems to work on. Kuhn and Hacking describes the nature of scientific progress as waves of research that fundamentally reframes the accepted paradigm of the scientific community. As fields with multi-disciplinary compositions mature and shift in their epistemologies, it raises questions about what forms of evidence count as legitimate and whose methods should be used to generate knowledge (Jacobs and Frickel 2009; Talbi and Van Woerden 2025). Such epistemic boundaries reflect the values embodied by the field, which may be dominated by the majority voices, such as machine learning favouring technical and quantitative optimization over societal impacts (Birhane et al. 2022a).

Human-computer interaction (HCI) also offers a parallel precedent, where the field originated from the positivist and quantitative paradigms of computer science and engineering (Duarte and Baranauskas 2016; Harrison, Tatar, and Sengers 2007). Later waves integrated cross-disciplinary theories from cognitive science and phenomenology, but due to the origins of the field being from technical disciplines, interpretivist and qualitative methods are often sidelined or misinterpreted (Soden, Toombs, and Thomas 2024). Deliberate advocacy for the inclusion of such methods was needed to push the field towards acceptance, offering a roadmap for how AISE may navigate incorporating human research. This aligns with Miller et al.'s concept of *epistemological pluralism*, which argues that different 'ways of knowing' can be integrated to gain a fuller description of a field. In this study, we examine how AISE experts currently navigate the clashing paradigmatic tensions of different disciplines, and how human research is positioned to contribute value to the field.

Evolution of AI Safety & Ethics

While the history of AI Safety and AI Ethics can be traced back to technology ethics (Vallor 2016), science and technology studies (Bijker, Hughes, and Pinch 1987), and safety engineering (Perrow 1984; Rasmussen 1997; Swuste et al. 2021), the two fields were substantially formalized from the mid-2010s as AI systems were being adopted in practice. AIS's initial focus was on AI alignment (Russell 2019; Amodei et al. 2016)—ensuring AI systems behave in accordance with human intentions, often expressed mathematically—with intellectual roots in Bostrom and Yudowsky's existential risk framing of Artificial General Intelligence (Bostrom 2014; Yudkowsky 2008). Scholars criticized AIS's ties to effective altruism, longtermism, and rationalism for fostering ideological homogeneity at the expense of excluding non 'in-group' voices (Ahmed et al. 2023; Gebru and Torres 2024). Furthermore, AIS' technology-centric perspective led to the sidelining of complex societal impacts of AI (Dahlgren Lindström et al. 2025; Walker et al. 2024). In response to these critiques, AIS has expanded to include broader perspectives such as technical AI governance (Reuel et al. 2024a) and system safety (Rismani et al. 2023; Dobbe 2022).

In contrast, AIE emerged from longer traditions of studying the societal impacts of algorithmic systems and gained significant momentum when the real-world deployment of AI systems began producing visible, documented harms (Birhane et al. 2022b; Shelby et al. 2023). This prompted research and policy communities to shape high-level ethical principles (Jobin, Ienca, and Vayena 2019) and venues like AIES (AAAI Conference on AI, Ethics, and Society) and FAccT (ACM Conference on Fairness, Accountability, and Transparency) gained prominence. AIE coalesced around a cluster of ethical principles, including fairness, transparency, privacy, sustainability, and accountability, distinctively centering the experiences of impacted communities (Costanza-Chock 2020) with various levels of rigour and success.

Some scholars working across both communities have increasingly tried to map and close the gap between them (Kasirzadeh 2025; Shen et al. 2024; Lazar and Nelson 2023; Roytburg and Miller 2025). Gyevnár and Kasirzadeh identify four common forms of engagement in AI ethics and safety researchers, including radical confrontation, disengagement, compartmentalized coexistence, and critical bridging. They argue for creating spaces that allow for productive exchanges between the two fields.

Human Research in AI Safety & Ethics

As AI systems become more deeply integrated into consequential domains, taxonomies of AI risk have increasingly attended to harms that are experienced by people. Sociotechnical frameworks of AI risk (Weidinger et al. 2022; Shelby et al. 2023) tackle the interactional and experiential dimensions of harm where the user experiences the outputs of the AI via direct interactions, or by existing in an AI-driven society. These include cognitive offloading and skill erosion (Chalkidis and Søgaard 2026), emotional over-reliance on AI systems (Saracini, Cornejo-Plaza, and Cippitani 2025), gradual disempowerment through AI-mediated

decision environments (Kulveit et al. 2025), susceptibility to AI-enabled manipulation (Matz et al. 2024), and much more. These harms are emergent, but not hypothetical. Yet dominant risk frameworks continue to foreground technical problems in alignment, robustness, and security (Hendrycks et al. 2021), leaving many of the most immediate human-centered harms underserved (Weidinger et al. 2023).

Where human methods do appear in AISE, they tend to be large-scale, quantitative, and narrowly evaluative – representing the perspectives and behaviours of an *average person* rather than understanding lived experience of harm. Reinforcement learning from human feedback (RLHF) represents the field's most institutionalized form of human engagement for value elicitation (Ouyang et al. 2022), but the human is treated as a training signal instead of the subject of research. Red-teaming (Perez et al. 2022) and scalable oversight research (Bowman et al. 2022) involve human participants but treat them as parts of the evaluation pipeline. Randomized controlled trials (RCTs) testing the impact of AI design on perceptions, beliefs, and performance are becoming more common (Paskov et al. 2026), but remain rare due to the high costs involved. In contrast, qualitative, participatory, and mixed-methods approaches have a more established presence in AI Ethics research, in investigating empirical accounts of how AI impacts affected communities (Birhane et al. 2022b). However, this methodological tradition has developed largely in parallel to AIS (Gyevnár and Kasirzadeh 2026; Roytburg and Miller 2025), keeping the two disciplines effectively separated.

This methodological narrowness draws parallels to the field of explainable AI (XAI), a case study in which a field built technical methods premised on the assumption that model transparency improves human-AI collaboration, only for empirical studies to reveal that explanations can increase over-reliance and worsen decision outcomes (Kaur et al. 2024; Bansal et al. 2021). AISE risks repeating this sequence at a larger scale, investing in technical mitigations for harms whose human dimensions have not yet been adequately characterized. We therefore approach this paper as a way to understand *why the deeply human problems of AISE are not always approached with human research*.

Methods

We gather AI Safety & Ethics expert insights through two complementary methods: a survey ($n = 93$), and in-depth interviews ($n = 17$). This mixed-methods design allows us to identify trends across disciplines and enrich them with nuanced accounts from individual researchers (Creswell and Clark 2017; Tashakkori and Teddlie 2021).

Expert Survey

The survey covers both participants' personal practices and broader perspectives on their discipline and AISE as a field. Inclusion criteria were intentionally broad, requiring only that participants self-identify as working on AISE-relevant research. This encompasses independent researchers without institutional affiliation, governance practitioners who may not publish, and those who may feel excluded by narrower definitions of AI Safety. Professional identities were verified through self-reported areas of focus and, where applicable, professional email addresses. Two attention check questions were included; responses failing both were excluded. The survey comprised four sections (see Appendix A for full question text):

1. **Background and Experience:** Professional background in AISE, including primary research area (*Technical*, *Sociotechnical*, *Governance*, or *Normative*). We use more granular categories than AIS and AIE to be inclusive of disciplines that may fall outside those epistemic boundaries, and further clustering by shared similarities in methods and ontologies.
2. **Practice and Aspirations:** Ratings of how often participants currently perform, and how useful they perceive, various human research methods.
3. **Scenario-Based Perceptions:** Ratings of the usefulness of human vs. non-human methods for addressing four AISE risk scenarios, validated to vary on imminence and risk level. This section captures field-level views rather than discipline-specific practices.
4. **Barriers to Human Research:** Degree to which resource, knowledge, and methodological barriers have impacted participants' engagement in human research.

Expert Interview

Semi-structured interviews were conducted to gain richer, descriptive accounts to contextualize the survey findings. Each interview covered the participant's research area and its relationship (or lack thereof) to human research, cross-disciplinary collaboration practices, perceived AI harms, and views on the value and fit of human methods in AISE. For those with prior human research experience, we additionally probed barriers they had encountered.

Analysis followed a mixed-methods approach in which survey results provided the organizing structure for qualitative coding (Creswell and Clark 2017). We developed a codebook deductively from the research questions, refined inductively using themes emerging from the interviews (Braun and Clarke 2006). The initial codebook was developed by the first author and then refined through independent coding by two additional authors, with final themes agreed upon collaboratively by the full team. Author positionality, including expertise in human research and engagement with AISE, is reported at the end of the paper.

Recruitment and Participants

We recruited primarily through the annual conference of the International Association for Safe & Ethical AI[1] (IASEAI 2026), which convenes multidisciplinary AISE researchers across academia, industry, non-profits, and government. Unlike AI Ethics-focused venues such as FAccT and AIES, IASEAI includes a substantial technical AI Safety community, making it well-suited to our cross-disciplinary aims. We contacted attendees from an internal database who met a minimum threshold of research engagement (e.g., at least

[1] https://www.iaseai.org

PID	Yrs. of Exp.	Sector & Role	Region
$P1_S$	2	Academia (Student)	N. America
$P2_T$	1	Academia (Student)	N. America
$P3_G$	6	Academia (Researcher)	Europe
$P4_S$	5	Academia (Researcher)	N. America
$P5_T$	2	Academia (Student)	N. America
$P6_S$	5	Academia (Researcher)	N. America
$P7_G$	6	Academia (Researcher)	N. America
$P8_G$	9	Non-Profit (Manager)	Europe
$P9_S$	4	Academia (Student)	Europe
$P10_T$	1	Academia (Researcher)	N. America
$P11_T$	9	Industry (Researcher)	N. America
$P12_T$	2	Academia (Student)	N. America
$P13_N$	7	Industry (Researcher)	Europe
$P14_T$	3	Non-Profit (Researcher)	Asia
$P15_G$	10+	Industry (Researcher)	N. America
$P16_N$	6	Academia (Student)	N. America
$P17_S$	2	Academia (Student)	N. America

Table 1: Summary of interview participants. They are referred to in the text by their participant number $PID_{Primary\ Area}$, where the subscript indicates the primary area they reported in the survey (e.g. *S* for *Sociotechnical*). For the role, 'Researcher' can refer to faculty, research leads, postdocs, or research engineers.

one published paper or a senior position in a relevant organization). We supplemented this with open recruitment via team social media and targeted Google Scholar searches for authors of high-impact papers in AI safety, ethics, alignment, and harm evaluation — prioritizing underrepresented areas when needed.[2] Participation in was not compensated as the experts were motivated to contribute to field-advancing research of their own volition.

The survey sample ($n = 93$) comprised primary research areas of *Technical* ($n_T = 32$), *Sociotechnical* ($n_S = 26$), *Governance* ($n_G = 25$), and *Normative* ($n_N = 10$). The sample skewed toward WEIRD contexts (Septiandri et al. 2023), with participants located primarily in North America (47.3%) and Europe (44.1%). Most hold or are completing a PhD (67.8%), work in academia (70.9%), and are substantially or fully engaged in research (88.2%). Years of AISE research experience ranged from 1 to 10+ (M=4.44, SD=2.70). Recruiting through IASEAI allowed us to capture non-academic participants while retaining a diverse sample active across sectors. Full demographic details are in Tables 2–7 and Figure 4 in Appendix B.

For interviews, we sought diversity across disciplines and sectors, though participants who opted into a follow-up interview were likely already engaged with human research, introducing selection bias. The final interview sample ($n = 17$) included participants across all four research areas (6 *Technical*, 5 *Sociotechnical*, 4 *Governance*, and 2 *Normative*), across experience levels (M=4.7 years), and across academia, non-profits, and industry. See Table 1 for details.

[2]For example, to target *Normative*, we contacted recent authors from the *Minds and Machines* journal.

Results

We organize our results in two sections: RQ1) how experts perceive the epistemic fit of human research within AISE, and RQ2) the resource, infrastructural, and sector limitations that impede it. Each section is supported by quantitative survey analyses and thematic interview findings. Of the 17 interviewees, the majority had direct or adjacent experience with human research—six had conducted it themselves, two engaged domain expertsin participatory ways, and two had prior human research experience in other fields—providing a range of perspectives on both practice and aspiration.

RQ1: Epistemic Value and Disciplinary Divides

We explore the epistemic fit of human research by examining the degree to which human research methods are perceived as generating legitimate, useful evidence within AISE. Specifically, what can human studies offer that other approaches cannot? How do disciplinary training and epistemic priors shape researchers' valuations of that evidence? And what does the interplay of disciplines within AISE mean for the acceptance of human methods?

The Human Evidence Gap. Participants agreed that AISE research objectives are becoming increasingly human-centered and value-driven, shifting from theoretical and hypothetical problems toward the interactions between AI systems and people ($P6_S$, $P9_S$, $P12_T$, $P15_G$, $P17_S$) (Dotan and Milli 2019; Birhane et al. 2022a). High-impact sociotechnical harms demand urgency in collaborative action ($P3_G$) and play into the moral conscience of AI researchers ($P17_S$). However, dominant computational approaches are frequently limited by *construct validity* problems, as they struggle to capture subjective constructs like what harm to humans actually means in practice ($P2_T$, $P6_S$, $P9_S$, $P11_T$, $P12_T$, $P15_G$). Many technically oriented participants themselves critiqued a positivism-rooted ideology that reduces nuanced phenomena to single optimizable metrics, flattening the very constructs AISE aims to evaluate. Gathering evidence from domain-specific stakeholders can challenge these assumptions and surface unexpected findings ($P1_S$, $P4_S$), though there is still a lack of consensus on which AISE issues warrant evaluation ($P6_S$) or which human methods are appropriate ($P14_T$).

> *People have this obsession with a single metric. Like, "We can just make a scale from 0 to 100", which is totally arbitrary. There is no construct validity. It's literally a number we decided, and now we're going to declare this as the metric for goodness, and everyone will align around those numbers.* — **P11**$_T$, on the construct validity of benchmarks.

Human research is valued for constructing evidence of AI's impact, in terms of both benefits and harms. It can counter or validate normative assumptions embedded in AI design and evaluation ($P4_S$, $P6_S$, $P12_T$), and ground theory on socially complex topics like fairness, cognition, and human-AI collaboration ($P5_T$, $P8_G$, $P10_T$, $P16_N$). As $P12_T$ notes, *"the really hard part is actually defining what is a bad output, and my work was assuming you have this defined"* — a definition that cannot be assumed without evidence.

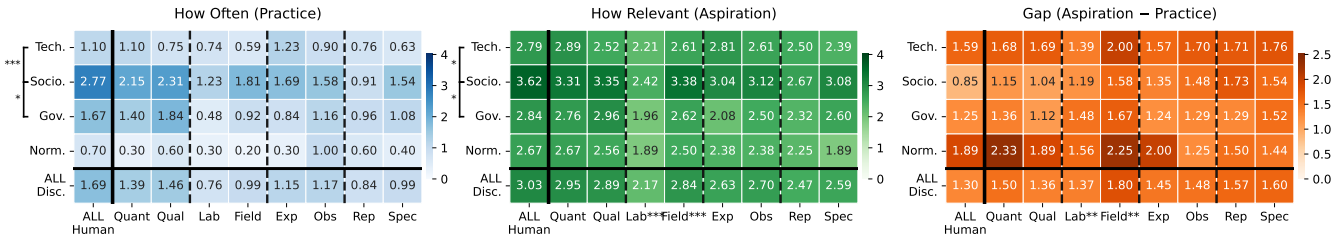


Figure 1: Within their disciplines, how often do participants *practice* human research, and how much they would *aspire* to perform methods relevant to their research. Both are measured on a Likert scale mapping to values between 0 - 4. Disciplines or research dimensions that have a significant difference are indicated (*** for $p < .001$, ** for $p < .01$, * for $p < .05$).

This need for human evidence is particularly acute for governance. P11$_T$ and P14$_T$ argue that policy should be supported by empirical evidence such as large-cohort A/B experiments, but this is not yet standard. P8$_G$ states that *"the studies we need for AI policy are non-existent yet"*, and P7$_G$ echoes that *"we have such little evidence on the effects on people."* In technical AI governance and alignment, interest centers on catastrophic and emergent harms. Incident case studies (e.g., the AI Incidents Database[3]) offer indicators of impending risk (P7$_G$), but human-validated evidence of harm at scale, or of effective mitigations, remains lacking.

> *AI safety people are trying to be proactive and front-run against hypothetical harms, or harms that seem very likely to spring up but haven't concretely materialized yet. [...] There's a very fine, unclear line about when technology amplifies human agency or reduces it. A really important question that would be nice to solve is not necessarily a monolithic benchmark for this, but a collection of indicators that correspond with human disempowerment.* — **P2**$_T$, on the need for human evidence for AIS problems.

The survey corroborates these interview findings. We measured both the prevalence of human research in participants' current practices and its perceived value if incorporated. Human empirical research was decomposed into four methodological dimensions: *quantitative* vs. *qualitative* paradigms, *lab* vs. *field* settings, *experimental* vs. *observational* designs, and *representative* vs. *specialized* sampling. For each, we computed a *gap score* (aspiration − practice) to quantify the discrepancy between what researchers do and what they would value doing. Results are shown in Figure 1; full analysis details are in Appendix C.

Consistent with disciplinary expectations, *Sociotechnical* researchers practice and aspire to human research at significantly higher rates than *Technical* (practice: $Z = 4.19, p_{adj} < .001$; aspiration: $Z = 2.69, p_{adj} = .007$) and *Governance* (practice: $Z = 2.67, p_{adj} = .007$; aspiration: $Z = 2.79, p_{adj} = .005$), by Holm-corrected Dunn post-hoc tests following Kruskal-Wallis comparisons.[4] Direct contrasts between *Technical* and *Governance* were largely non-significant, other than *Governance* reporting more qualitative practice and *Technical* showing stronger aspirational preference for experimental designs. Critically, across all disciplines and dimensions, gap scores were consistently positive, indicating that human research is valued beyond what is currently practiced.

At the research dimension level, Wilcoxon signed-rank tests revealed a single significant difference: researchers aspire to conduct *field* studies more than *lab* studies. This has two implications. First, no other dimension — including the paradigmatically contentious quantitative/qualitative divide — produced significant differences, suggesting that diverse forms of human research are broadly considered epistemically acceptable within AISE. Second, the field-over-lab preference signals a collective interest in *ecological validity* over experimental control. In summary, there is clear interest for human research across disciplinary lines, but an aspiration-practice gap means it is not being conducted at a rate commensurate with the field's needs.

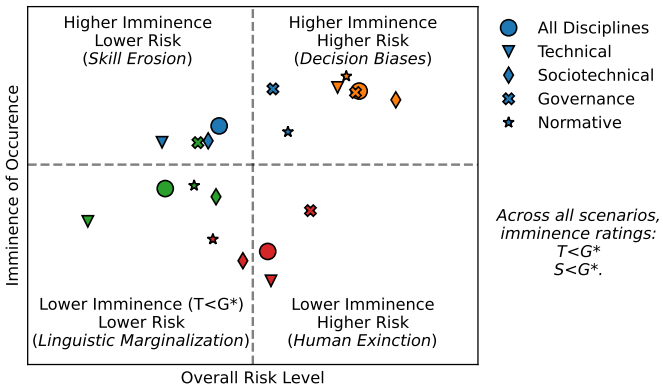
Disciplinary Differences in Valuing Human Research. We next examine whether and why disciplines differ in how they value human versus non-human methods. In the scenario-based survey section, participants rated the usefulness of both human and non-human methods for addressing four AISE risk scenarios varying on risk level (low vs. high) and imminence (future vs. current). Participants were not asked to represent their discipline, so these ratings reflect general beliefs about the field, albeit filtered through the lens of their epistemic training. Results are in Figure 2 and the detailed analysis is in Appendix C. The full description of scenarios and methods is provided in Appendix A.

Participants agreed broadly on the classification of scenarios: *Skill Erosion* and *Decision Biases* were rated as more imminent, while *Decision Biases* and *Human Extinction* were rated as higher risk. The one disciplinary divergence was in imminence ratings: *Governance* rated scenarios as more imminent than both *Sociotechnical* ($Z = 2.62, p = .04$) and *Technical* ($Z = 3.25, p = .007$), consistent with governance researchers' orientation toward anticipating risks before they fully materialize. Otherwise, disciplines were in close alignment on perceived severity.

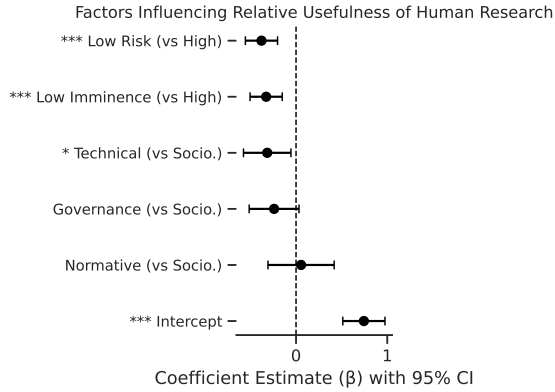
To test whether discipline predicted the relative value assigned to human (e.g. 'test intervention with users') versus non-human (e.g. 'develop an automated benchmark') methods, we fit a mixed-effects model with random intercepts by participant, modelling the 'human − non-human' usefulness gap as a function of risk level, imminence, and

[3] https://incidentdatabase.ai/

[4] *Normative* is excluded due to small sample size.



(a) Rating of the 1) imminence, and 2) risk level (combined *likelihood* × *severity*) of four various AISE risks. The average ratings (large circle marker) agree with the a priori classification of the scenarios. Inter-disciplinary differences only manifest as *Governance* rating scenarios as more imminent than other disciplines.



(b) Mixed-effects estimates of factors influencing the relative perceived usefulness of human research methods. **Human methods are valued less when the scenarios have low risk and imminence, and by the *Technical* discipline in comparison to *Sociotechnical*.**

Figure 2: Survey participants' ratings of AISE risk scenarios (left) and their preferences for human vs non-human research approaches for the scenarios (right). These ratings capture the general underlying beliefs, not discipline-specific practices.

discipline (reference: *Sociotechnical*). The positive intercept ($\beta = 0.74, p < .001$) confirms that *Sociotechnical* researchers favoured human methods overall. This preference was attenuated for low-risk ($\beta = -0.38, p < .001$) and future scenarios ($\beta = -0.33, p < .001$), indicating that the perceived advantage of human methods diminishes when stakes are lower or harms more distal. ***Technical* researchers showed a significantly smaller preference for human methods than *Sociotechnical*** ($\beta = -0.32, p = .017$); *Governance* showed a similar trend that did not reach significance ($\beta = -0.24, p = .084$).

The interviews offer explanations for this observed undervaluation of human methods, stemming from a combination of legitimate methodological concerns and epistemic bias. On the legitimate side, construct and measurement validity were the foremost concerns, particularly for RCTs (Paskov et al. 2026). P8$_G$ critiques existing human-AI interaction RCTs as *"really methodologically poorly designed and not controlled"*, and P6$_S$ echoes that some human studies *"aren't credible because their methodology doesn't answer the [research question]."* P13$_N$ finds issues in how constructs like *"trust in AI is measured and operationalized in empirical studies"*. Scalability is also a concern: frequently updated models cannot always be evaluated with human participants (P11$_T$), and large-population effects may be more tractable via simulation (P16$_N$).

Beyond legitimate concerns, discipline-specific biases also play a role. P6$_S$ and P8$_G$ both emphasized the importance of incorporating qualitative analysis to *"explain idiosyncrasies and unique perspectives of individuals"* (P6$_S$) and *"investigate how people use evidence and outputs of AI"* in complex, strategic tasks (P8$_G$). However, technically oriented researchers reported less familiarity with qualitative methods and sometimes held partial views of them — perceiving qualitative evidence as perhaps *"easier to manipulate than quantitative ones"* (P2$_T$). P12$_T$ notes that even in the relatively qualitative domain of red-teaming, reporting of *how* attacks succeeded is often too thin to generalize or reproduce, reflecting underinvestment in the rigour of qualitative works. Human methods that map onto quantitative optimization are preferred. For example, 'AI uplift' studies (RCTs of how AI scales human capabilities) are reported as the paradigm of choice for many (P2$_T$, P7$_G$).

> *I think they actually need to explain why qualitative studies are helpful. Yes, it's only 10 people, but you can still learn a lot from 10 people, as long as you ask the right questions and you do careful work. But because there's an obsession with scaling, there's an obsession with numbers, and this is a big barrier that is hard to overcome.* — **P6**$_S$, on the lack of literacy of qualitative methods in AIS.

The survey's methodological barriers section (see Figure 3, which is combined with RQ2's analysis on tangible barriers) corroborates this: *Sociotechnical* researchers reported significantly less impact from the categories of *'uncertainty about methods'* and *'preference for using existing datasets'* than other disciplines. Aesthetic and professional factors also play a role: some researchers stated to be deterred by the 'messiness' of human data (P6$_S$, P12$_T$), while others cite a lack of prestige associated with empirical methods in their communities (P5$_T$). P2$_T$ suggests that HCI research may be more appealing to AIS researchers through *"aesthetically changing"* the semantics and motivations used. Nevertheless, even technically oriented experts acknowledge the limits of purely computational approaches, as P9$_S$ reflects, *"we try very hard to avoid humans, but then in some settings, we find that actually involving humans is the only way to answer the research questions."*

Challenges in Cross-Disciplinary Bridging. The preceding findings find agreement that human evidence matters, but persistent disciplinary siloes constrain its adoption.

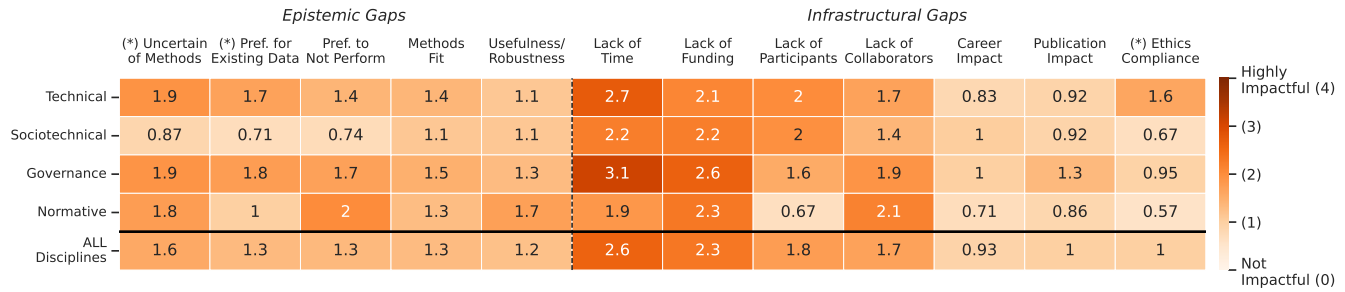


Figure 3: Ratings for the impacts of epistemic (RQ1) and resource and infrastructural (RQ2) barriers to human research. Participants could also select N/A; such responses were excluded from analysis. Categories with a significant Kruskal-Wallis test of group differences are indicated with * ($p < .05$).

Here, we examine how those epistemic gaps manifest in collaboration and what conditions support bridging.

AISE (as well as the broader community of AI in general) is an evolving field. This is reflected in the expanding scope of technical AI conferences ($P9_S$), shifts in community composition ($P6_S$), and the fact that 77 of 93 survey participants identify with more than one research area. Yet in both cross-disciplinary engagement and collaboration frequency shown in Figure 4, *Technical* researchers report the lowest, suggesting a stronger tendency toward disciplinary siloing in AIS. ***Technical* collaborates significantly less than all other disciplines**[5](*Sociotechnical*: $Z = -3.57, p_{adj} = .001$; *Governance*: $Z = -3.45, p_{adj} = .001$) by Holm-corrected Dunn post-hoc tests following Kruskal-Wallis. As $P13_N$ emphasizes, *"[AISE] problems are definitely too complex to have only one person in the room"*, but the field remains dominated by technical safety perspectives.

Tracking with prior research, the AIS/AIE divide is particularly salient (Gyevnár and Kasirzadeh 2026; Roytburg and Miller 2025). Participants noted that while methods differ across these communities, their goals are aligned ($P9_S$) and *"the methods are different, but the mental structure is the same"* ($P13_N$). The barrier is not disagreement about what matters, but access to knowledge and people. $P13_N$ described difficulty finding entry points into human-centred AI expertise, saying *"connecting the experts in human-centered AI with people who are interested in human-centered AI is not easy... are there textbooks on human-centered AI anywhere? That would be great for people like me."* At the same time, bridging must be bidirectional. HCI and social science researchers should also actively engage with AIS communities rather than waiting to be discovered ($P6_S$). Institutional infrastructure that supports interdisciplinary exchange facilitates these collaborations ($P1_S$, $P6_S$, $P16_N$), but is not universally available for everyone ($P12_T$, $P14_T$, $P17_S$).

> *It's important to acknowledge others' perspective... but we shouldn't expect fields to then just flatten their differences completely, right? Like, there is value in differences. It's just, can we approach those differences in a structured way? Rather than immediately going into this reflex of, "I hate this", or "I love this"?* — **P6**$_S$, on constructively bridging between different epistemologies.

[5]*Normative* not included due to small sample size.

Experts appreciated cross-disciplinary venues like IASEAI, but felt that the community still reflected AIS over AIE, as *"the safety groups have money and presence"* ($P15_G$), and the opportunities for bridging were superficial and *"didn't quite work this year"* ($P6_S$). Top-down unification efforts carry risks of epistemic collapse and academic gatekeeping, and $P6_S$ observes that *"AI safety has not yet matured into a field where human studies are accepted as a necessary component of evaluation."* This creates a dual risk. If human research is excluded from AISE's epistemic boundaries, the evidence base for making AI systems safe will be fundamentally incomplete. But if human research adopted as a performative checklist item rather than a rigorous contribution, the result is what we term *human-washing*: the superficial inclusion of human subjects that creates the appearance of validity. Both outcomes diminish the field's capacity to generate credible evidence of harm and advance the objective of making AI safe and ethical.

RQ2: Barriers to Human Research: Resources, Infrastructure, and Sector Boundaries

Having examined how AISE experts value human research and why a gap persists between aspiration and practice, we now turn to the tangible barriers that prevent researchers from executing human studies. We identify three levels of constraint: resource-related, infrastructural, and sector-level dynamics, reflecting prior findings on barriers of human research (Agapie, Haldar, and Poblete 2022). Survey ratings on the degree to which each barrier was felt are shown in Figure 3 and interview accounts provide explanatory depth.

Resource Limitations. The most direct constraints concern the material resources available to researchers: time, funding, and high-quality participants. No significant differences across disciplines were found for any resource barrier in the survey, suggesting these constraints are felt broadly. Averaged across all respondents, *lack of time* was ranked most impactful, followed by *funding*, then *participant access*. These barriers often compound each other, such as long recruitment periods consuming time that delays other work, and waiting on funding decisions stalls momentum. Ethics approval adds further administrative delay, which $P4_S$ notes can disrupt the continuity of a research project.

I had a discussion with [a colleague] who's doing large-scale human subjects research, and he just wanted to go back and do a [computational] benchmark... you can spend a few months and get a good result, and then it's done, and then you can move on. — **P9**$_S$, on the comparatively substantial time required for human research.

While time is often weighed internally by the researcher, funding poses a hard external constraint. P10$_T$ was unable to conduct a neuroscience-based human-AI interaction study due to losing a major grant. The estimated cost of 300k USD was prohibitive despite a fully developed proposal. This sharply contrasts with the cost of computational work, which P9$_S$ estimates at *"less than a thousand dollars per paper."* However, funding conditions vary, as P1$_S$ and P4$_S$ reported adequate support from their labs. P6$_S$, who participates in grant reviews, observed promising movement towards including sociotechnical research in AISE funding.

Participant access introduces both logistical difficulty and quality trade-offs. P1$_S$ underwent multiple recruitment rounds after encountering scam respondents in a study requiring highly specific participant qualifications, reflecting the increasing issues in fraudulent participants observed in HCI (Panicker et al. 2024). P4$_S$, P11$_T$, and P13$_N$, who work in health-specific contexts, describe difficulty accessing medical specialists and clinical populations. Field studies present additional complexity, as P15$_G$ notes that in these contexts, *"you're really dependent on partnership building, partnership quality, and it might be 10 times as much time investment for a smaller scale."* Recruiting non-specialist populations such as crowdworkers or undergraduate students is more attainable (P6$_S$), but is perceived to reduce the relevance and validity of the findings (P2$_T$).

Infrastructural Limitations. Beyond individual resources, barriers are embedded in the *infrastructure* surrounding researchers — the organizational arrangements, mentorship structures, and professional incentives that shape what research is possible and viable to sustain careers (Lee, Dourish, and Mark 2006). These constraints fall heaviest on junior researchers, who depend most on their immediate environment for access and legitimacy.

Mentorship is a significant mechanism of both restriction and access. *Normative* participants rated *'lack of collaborators'* as one of their highest impact barriers, and the highest of any discipline. More broadly, interviews captured accounts of advisors actively discouraging human research: P9$_S$'s mentors were opposed to working with human subjects; P5$_T$ and P16$_N$'s institutions and leadership did not recognize the value of empirical methods; and P17$_S$'s sociotechnical interests were directly in tension with their advisor's focus on AI. In the absence of formal interdisciplinary structures, junior researchers are left to build networks informally, either through cold outreach (P17$_S$) or student communities (P2$_T$, P16$_N$, P17$_S$).

My advisor is a standard computer science, technical person who just wants to improve on the benchmarks and achieve AGI. I don't think I've been able to get through to him, so my goal with my PhD is to just shoehorn in the things that I think are important, while on the surface making it still look like an AI research PhD. — **P17**$_S$, on navigating tensions between their interests and their advisor's.

Career and publication pressures, while not highly rated in the survey overall, can be acutely constraining for specific individuals. P5$_T$ describes difficulty establishing credibility at intersection of HCI and theoretical AI research, compounded by the incompatible publication norms of those fields. They also recount gendered barriers around establishing legitimacy within a highly theoretical research group, which deters them from pursuing human research. This points to a broader pattern in AISE that privileges formal, quantitative, and technical contributions. These implied epistemic values reflect whose research is seen as legitimate.

I want to be accepted as a theoretical researcher. Because if I do human studies, people will think I am just here because I'm a woman, and I'm doing 'woman' research. — **P5**$_T$, on facing biases faced in theoretical research.

Sector-Level Constraints. Barriers to human research also vary by sector, and opportunities for mitigation lie partly in cross-sector collaboration. In non-profits and industry, the absence of formal ethics review infrastructure impedes engagement with human subjects, particularly in high-risk research contexts (P8$_G$, P11$_T$, P14$_T$) (Metcalf and Crawford 2016). P8$_G$, a non-profit research manager, addresses this by funding academic student fellowships, reflecting a broader view that human research is best conducted in academic institutions, which provide ethical oversight and research neutrality.

A lot of evidence from the labs is coming from companies that have business incentives to give impression that the models they are producing are extremely robust. Everyone thinks, "the AI labs have the best scientists, so they know what they're talking about, so this is evidence that we're going to take for the policymaking"... which I think is very problematic. — **P8**$_G$, on ethical issues of industry research.

The involvement of industry in AISE research is a source of both problematic implications and opportunity (Abdalla and Abdalla 2021). Participants raised concerns about the influence of unregulated, commercially incentivized actors on the research landscape (P2$_T$, P11$_T$), and about the rigour and neutrality of industry-produced research (P8$_G$, P13$_N$, P15$_G$). For example, Facebook's emotional contagion study (Kramer, Guillory, and Hancock 2014) is mentioned as a reference for ethical failure (P9$_S$, P15$_G$). At the same time, industry's access to large-scale deployment data and high-quality research talent cannot be overstated. P15$_G$ argues that, *"if [companies] doing risky research, or they want their research to be rigorous and not just internal A-B testing, I think they should be more academic."* Interesting, P2$_T$ adds that this credibility tension is not one-directional, saying *"that there's probably some prejudice from the AI safety community's perspective, where [AI Safety thinks] a lot of the good research hasn't come from traditional academia."*

The AI community is in its infancy with that, because they're all just still figuring out. It took 7 years to negotiate the IAEA at the height of the Cold War and it took 20 years to negotiate the ITER agreement. Normally, these things take a long time. — **P3**$_G$, on international collaboration.

Strengthening cross-sector collaboration is broadly viewed as necessary given the pace of AI development ($P15_G$). Governance and policy decisions are increasingly reliant on evidence that does not yet exist at the required scale, but in turn, these decisions will play a substantial role in controlling the speed and veracity of AI development. $P3_G$ further focuses on international politics, drawing comparisons to prior large-scale collaborations between countries in times of heightened tension. Ultimately, the timeline for multi-disciplinary, cross-sector, and international collaboration to prevent both existential and cumulative risks of AI is an urgent, epistemic priority.

Discussion

Our results show that experts across disciplines agree that empirical research with human participants offers meaningful evidence to AISE — and yet, despite epistemic fit, it remains underused. Although experts demonstrated interest for various methods of human research, many highlighted a predominant positivist bias that favours more technical methodologies and metrics, limiting the field's ability to answer research questions that require interpretivist and qualitative empirical approaches. Additionally, practical barriers and individual researcher biases further reinforce dominant research priorities. As the AISE community continues to iteratively define its identity, top-down efforts in constructing the epistemic boundaries of what methods and whose voices get to be highlighted must be approached critically. We synthesize recommendations and implications for bridging the epistemic gaps (RQ1) and tangible barriers (RQ2).

Epistemic Implications for AISE Researchers

Field-Level Epistemic Boundaries. Multidisciplinary fields such as human-computer interaction, public health and environmental sciences have long struggled to bridge epistemic divides, as integrating different values, methods and ways of work takes intentional efforts throughout the entire life-cycle of a research project (Talbi and Van Woerden 2025). The framework of *epistemological pluralism* (Miller et al. 2008) proposes reconciling disciplinary differences by recognizing them as complementary instead of competing perspectives. AISE faces a similar imperative. Our results demonstrate *Technical* researchers systematically undervalue human methods relative to their *Sociotechnical* counterparts, suggesting that the epistemic divide in AISE is actively shaping what evidence gets legitimized. Left unaddressed, this tension may compound the risk of ideological homogeneity in AISE community-building (Ahmed et al. 2023; Dahlgren Lindström et al. 2025). Thus, we pose that AISE should welcome the challenge of incorporating different 'ways of knowing' to expand potential avenues to identify, measure and mitigate AI harm (Gyevnár and Kasirzadeh 2025).

Methodological Literacy for Researchers. Beyond field-level recommendations, improving the presence of human research in AISE also requires change at the individual researcher level, particularly among *Technical* researchers, who report the lowest cross-disciplinary engagement. We do not argue that all researchers should conduct human studies – after all, disciplinary divides can provide constructive value (Gyevnár and Kasirzadeh 2026) – but that practitioners should understand the established methodological literature of adjacent fields (Talbi and Van Woerden 2025). The case of AI uplift studies illustrates both the promise and the limits of current practice. Despite involving human participants, uplift studies largely reproduce the positivist paradigm dominant in machine learning: controlled lab conditions, aggregate effect sizes, and quantitative outcome measures. Practitioners have themselves acknowledged core challenges around task realism in human uplift studies leveraging RCTs (Paskov et al. 2026), yet the body of HCI and social science research that has theorized and addressed exactly these ecological validity problems is rarely engaged. This insularity reflects a broader pattern: AISE's dominant positivist orientation privileges measuring the likelihood of harm while overlooking the interpretivist methods needed to understand the mechanisms through which harm unfolds and how it is experienced in practice.

Prevent "Human-Washing". While encouraging broader adoption of empirical human research methods in AISE, we must equally caution against superficial or performative participant engagement, as we termed *human-washing*. Past work has flagged similar patterns in machine learning research, like *diversity washing* (Whitney and Norman 2024) and *participation washing* (Sloane et al. 2022) that result in harmful misrepresentation and exploitation instead of inclusion. *Human-washing* poses the same risk, where the inclusion of human participants is enacted without the necessary methodological rigour, creating the appearance of empirical grounding while undermining it. Our interview participants note the concern that some existing human studies in AISE are insufficiently designed to answer the research questions they claim to address. As the field's epistemic boundaries evolve, and as venues like IASEAI expand their scope, there is a potential risk that human research becomes a checklist requirement. Mandating human research without cultivating the literacy and infrastructure to do it well would create downstream issues with validating evidence legitimacy.

Implications for Resources and Infrastructures

Overcoming Resource Challenges The most direct lever available to institutions is funding, and the cost asymmetry between technical and human research is stark: a computational paper can cost under a thousand dollars, while a rigorous human-AI interaction study can run into the hundreds of thousands. Funding bodies, like government agencies, philanthropic foundations, and industry research programs, should explicitly account for the higher cost structures of human research in their grant mechanisms to acknowledge the imbalance and incentivize appropriately. Equally important is the framing and terminology used to describe funding opportunities, as opting for technical topics implicitly signals what methods are valued, discouraging human-centered proposals before they are written. This pattern can be observed in past grants that primarily targeted technical AI safety, with a focus on AI alignment: OpenAI's Superalignment

Fast Grant[6], the AI Security Institute's Alignment Project[7], and so forth. Beyond funding opportunities, technical researchers without prior human subjects training would benefit from accessible resources for navigating ethics compliance so that Institutional Review Board (IRB) processes do not pose a significant barrier to entry.

Mentorship and Training. Mentorship functions as a significant mechanism of methodological gatekeeping in AISE, with junior researchers who pursue human studies frequently facing discouragement from supervisors, limited access to relevant collaborators, and in some cases the need to obscure their methodological interests entirely. Departments and graduate programs should formalize pathways for interdisciplinary mentorship so that students are not wholly dependent on a single advisor's agenda and skill set (Jacobs and Frickel 2009). Mitigation strategies such as co-supervision arrangements, visiting researcher programs, and funded rotations are examples to facilitate cross-pollination and to build stronger collaboration networks. Fellowship programs represent a parallel opportunity: existing AIS fellowships at non-profits and industry organizations have demonstrated an effective model for training junior researchers, but currently reflect the field's methodological insularity. Expanding these programs to include human-research tracks, or funding parallel structures in sociotechnical and governance areas, would both develop empirically literate researchers and signal that human methods are valued in high-prestige settings.

Cross-Sector Collaboration. Our participants suggested a collaboration model structured around integrated academic-industry partnerships, rather than isolated parallel efforts. Additionally, non-profit organizations are well-positioned to act as neutral intermediaries, insulating research directives from commercial influence while preserving access to industry resources and facilitating data-sharing arrangements that could give academic researchers access to deployment-scale human evidence under independent ethics oversight. At the policy level, governance bodies should treat the production of human evidence as an active priority, funding the infrastructure that makes it possible rather than waiting for it to emerge organically. International coordination will be necessary for evidence about cross-jurisdictional harms, and while precedents from large-scale scientific collaborations suggest this is achievable, the pace of AI development means the window for establishing this infrastructure is already compressed.

Influence and Power. As custodians of knowledge production, institutions in AISE exert substantial influence over which research agendas are legitimized and prioritized, with consequences that extend beyond the field into public discourse and policymaking. Industry research from the tech giants Anthropic, Google and OpenAI have, until very recently, concentrated efforts on the evaluation of AI abilities and behaviours (Achiam et al. 2023; Kalai et al. 2025) over how real users engage with AI (Shen and Tamkin 2026; Phang et al. 2025). Notably, industry papers tend to receive more citations than those authored by researchers in academia (Strauss et al. 2025), further demonstrating their contributions receive more attention and traction. This trend shapes what gaps are considered worth filling, and what methods are considered rigorous. Potential mitigations include conflict-of-interest disclosure requirements at journals and conferences, and citation audits in systematic reviews to surface whose work is or is not being built upon.

Furthermore, the concentration of prominent voices in AISE compounds this dynamic. Researchers from elite universities and major laboratories often share epistemological assumptions favoring technical and quantitative approaches that collectively marginalize the human-centered, qualitative and participatory methods. When the figureheads of AISE do not practice or value human-centered research methods, it is difficult for such methods to gain traction regardless of their epistemic merit. As a remedy expert panels, advisory boards, and parliamentary testimony should mandate disciplinary and methodological diversity, explicitly recognizing that lived experience of AI harm is a form of evidence that technically-oriented research will not reveal.

Limitations

Our findings should be interpreted with the following limitations. Recruitment was targetted to IASEAI attendees and the researchers' existing professional networks rather than a systematic sampling of the entire AISE community. The survey was also skewed toward WEIRD contexts, and industry researchers were underrepresented relative to their influence. Voluntary interview participants were likely already predisposed toward human research; this means those who are most skeptical of human methods might be underrepresented in our qualitative findings. While our four discipline categories provided a useful organizing structure, they were not intended to capture the full complexity of the researchers' identities. Finally, because the interviews were intentionally open-ended, the themes that emerged did not always map onto the survey constructs, which limits direct comparisons between data sources.

Conclusion

Navigating epistemic tensions is a natural step in the development of an emergent scientific field, and AISE is no exception. Nonetheless, past work has identified the predominance of technical AI Safety research paradigms in AISE literature. We contribute to this discussion by examining the lived experiences of the research practitioners in AISE via expert survey and interviews. Our findings reveal overall agreement that empirical human research adds value but is underutilized. Critically, experts from *Technical* backgrounds rate human methods as *less useful*, and tend to *collaborate less* across disciplines. This imbalance risks reducing the AISE research agenda to the point of epistemic hegemony, implicitly delegitimizing other approaches such as human-centred methods. We pose that explicitly inclusive, yet not performative, epistemic boundaries are paramount to the recognition of human evidence from human research paradigms to holistically address AI-related harms.

[6]https://openai.com/index/superalignment-fast-grants/

[7]https://alignmentproject.aisi.gov.uk/research-agenda

Positionality

The academic and research backgrounds of the research team heavily informed the research questions and analytical methods used in this paper. All authors are academics from Computer Science and Information Sciences, with a focus on human-computer interactions, responsible AI, and systems safety. Everyone has experience with conducting human empirical research, ranging from high-powered controlled experiments to ethnographic studies. The first and last authors come from quantitative-leaning backgrounds in HCI and computational social science, and other authors have expertise in qualitative and critical methods. We therefore chose a mixed-methods approach that combines surveys and interviews. Three of the authors, including the first author, also engage with technical AI safety and have direct experience with synthesizing from multidisciplinary research.

References

Abdalla, M.; and Abdalla, M. 2021. The grey hoodie project: Big tobacco, big tech, and the threat on academic integrity. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 287–297.

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Agapie, E.; Haldar, S.; and Poblete, S. G. 2022. Using HCI in cross-disciplinary teams: a case study of academic collaboration in HCI-health teams in the US using a team science perspective. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–35.

Ahmed, S.; Jaźwińska, K.; Ahlawat, A.; Winecoff, A.; and Wang, M. 2023. Building the epistemic community of AI safety. *SSRN Electron. J.*

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete Problems in AI Safety. *arXiv [cs.AI]*.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Bansal, G.; Wu, T.; Zhou, J.; Fok, R.; Nushi, B.; Kamar, E.; Ribeiro, M. T.; and Weld, D. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–16.

Bean, A. M.; Kearns, R. O.; Romanou, A.; Hafner, F. S.; Mayne, H.; Batzner, J.; Foroutan Eghlidi, N.; Schmitz, C.; Korgul, K.; Batra, H.; et al. 2026. Measuring what matters: Construct validity in large language model benchmarks. *Advances in Neural Information Processing Systems*, 38.

Bengio, Y.; Mindermann, S.; Privitera, D.; Besiroglu, T.; Bommasani, R.; Casper, S.; Choi, Y.; Fox, P.; Garfinkel, B.; Goldfarb, D.; et al. 2025. International ai safety report. *arXiv preprint arXiv:2501.17805*.

Bijker, W. E.; Hughes, T. P.; and Pinch, T., eds. 1987. *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. Cambridge, MA: MIT Press.

Birhane, A.; Kalluri, P.; Card, D.; Agnew, W.; Dotan, R.; and Bao, M. 2022a. The values encoded in machine learning research. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 173–184.

Birhane, A.; Ruane, E.; Laurent, T.; Brown, M. S.; Flowers, J.; Ventresque, A.; and Dancy, C. L. 2022b. The forgotten margins of AI Ethics. *arXiv [cs.CY]*.

Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bowman, S. R.; Hyun, J.; Perez, E.; Chen, E.; Pettit, C.; Heiner, S.; Lukošiūtė, K.; Askell, A.; Jones, A.; Chen, A.; et al. 2022. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*.

Braun, V.; and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2): 77–101.

Chalkidis, I.; and Søgaard, A. 2026. Brainrot: Deskilling and Addiction are Overlooked AI Risks. *arXiv preprint arXiv:2605.03512*.

Cheng, M.; Lee, C.; Khadpe, P.; Yu, S.; Han, D.; and Jurafsky, D. 2026. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391(6792): eaec8352.

Costanza-Chock, S. 2020. *Design Justice: Community-Led Practices to Build the Worlds We Need*. Cambridge, MA: MIT Press.

Creswell, J. W.; and Clark, V. L. P. 2017. *Designing and conducting mixed methods research*. Sage publications.

Dahlgren Lindström, A.; Methnani, L.; Krause, L.; Ericson, P.; de Rituerto de Troya, I. M.; Coelho Mollo, D.; and Dobbe, R. 2025. Helpful, harmless, honest? Sociotechnical limits of AI alignment and safety through Reinforcement Learning from Human Feedback. *Ethics Inf. Technol.*, 27(2): 28.

Dobbe, R. 2022. System safety and artificial intelligence. In *2022 ACM Conference on Fairness Accountability and Transparency*, FAccT '22, 1584–1584. New York, NY, USA: ACM.

Dotan, R.; and Milli, S. 2019. Value-laden disciplinary shifts in machine learning. *arXiv preprint arXiv:1912.01172*.

Duarte, E. F.; and Baranauskas, M. C. C. 2016. Revisiting the three HCI waves: A preliminary discussion on philosophy of science and research paradigms. In *Proceedings of the 15th Brazilian Symposium on Human Factors in Computing Systems*, 1–4.

Fang, C. M.; Liu, A. R.; Danry, V.; Lee, E.; Chan, S. W.; Pataranutaporn, P.; Maes, P.; Phang, J.; Lampe, M.; Ahmad, L.; et al. 2025. How ai and human behaviors shape psychosocial effects of extended chatbot use: A longitudinal randomized controlled study. *arXiv preprint arXiv:2503.17473*.

Gebru, T.; and Torres, É. P. 2024. The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. *First Monday*.

Gyevnár, B.; and Kasirzadeh, A. 2025. AI safety for everyone. *Nat. Mach. Intell.*, 7(4): 531–542.

Gyevnár, B.; and Kasirzadeh, A. 2026. Bridging the gap in the responsible AI divides. *arXiv [cs.CY]*.

Harrison, S.; Tatar, D.; and Sengers, P. 2007. The three paradigms of HCI. In *Alt. Chi. Session at the SIGCHI Conference on human factors in computing systems San Jose, California, USA*, 1–18.

Hendrycks, D. 2025. *Introduction to AI safety, ethics, and society*. Taylor & Francis.

Hendrycks, D.; Carlini, N.; Schulman, J.; and Steinhardt, J. 2021. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*.

Jacobs, J. A.; and Frickel, S. 2009. Interdisciplinarity: A critical assessment. *Annual review of Sociology*, 35(1): 43–65.

Jobin, A.; Ienca, M.; and Vayena, E. 2019. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.*, 1(9): 389–399.

Kalai, A. T.; Nachum, O.; Vempala, S. S.; and Zhang, E. 2025. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*.

Kasirzadeh, A. 2025. Two types of AI existential risk: decisive and accumulative: A. Kasirzadeh. *Philosophical Studies*, 182(7): 1975–2003.

Kaur, H.; Conrad, M. R.; Rule, D.; Lampe, C.; and Gilbert, E. 2024. Interpretability gone bad: The role of bounded rationality in how practitioners understand machine learning. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1): 1–34.

Kirk, H. R.; Davidson, H.; Saunders, E.; Luettgau, L.; Vidgen, B.; Hale, S. A.; and Summerfield, C. 2025. Neural steering vectors reveal dose and exposure-dependent impacts of human-AI relationships. *arXiv preprint arXiv:2512.01991*.

Kramer, A. D.; Guillory, J. E.; and Hancock, J. T. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24): 8788–8790.

Kuhn, T. S.; and Hacking, I. 1970. *The structure of scientific revolutions*, volume 2. University of Chicago press Chicago.

Kulveit, J.; Douglas, R.; Ammann, N.; Turan, D.; Krueger, D.; and Duvenaud, D. 2025. Gradual disempowerment: Systemic existential risks from incremental AI development. *arXiv preprint arXiv:2501.16946*.

Lazar, S.; and Nelson, A. 2023. AI safety on whose terms? *Science*, 381(6654): 138.

Lee, C. P.; Dourish, P.; and Mark, G. 2006. The human infrastructure of cyberinfrastructure. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, 483–492.

Leslie, D. 2019. Understanding artificial intelligence ethics and safety. *arXiv preprint arXiv:1906.05684*.

Li, L.; Chen, H.; Namkoong, H.; and Peng, T. 2026. Llm generated persona is a promise with a catch. *Advances in Neural Information Processing Systems*, 38.

Matz, S. C.; Teeny, J. D.; Vaid, S. S.; Peters, H.; Harari, G. M.; and Cerf, M. 2024. The potential of generative AI for personalized persuasion at scale. *Scientific Reports*, 14(1): 4692.

Metcalf, J.; and Crawford, K. 2016. Where are human subjects in big data research? The emerging ethics divide. *Big Data & Society*, 3(1): 2053951716650211.

Miller, T. R.; Baird, T. D.; Littlefield, C. M.; Kofinas, G.; Chapin III, F. S.; and Redman, C. L. 2008. Epistemological pluralism: reorganizing interdisciplinary research. *Ecology and society*, 13(2).

Ni, B.; Wang, Y.; Wang, L.; Kveton, B.; Dernoncourt, F.; Xia, Y.; Chen, H.; Luera, R.; Basu, S.; Mukherjee, S.; et al. 2026. A survey on llm-based conversational user simulation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4266–4301.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Panicker, A.; Nurain, N.; Ibrahim, Z.; Wang, C.-H.; Ha, S. W.; Wu, Y.; Connelly, K.; Siek, K. A.; and Chung, C.-F. 2024. Understanding fraudulence in online qualitative studies: from the researcher's perspective. In *Proceedings of the 2024 CHI conference on human factors in computing systems*, 1–17.

Paskov, P.; Wei, K.; Hong, S. Z.; Bateyko, D.; Roberts-Gaal, X.; Ezell, C.; Praninskas, G.; Chen, V.; Bhatt, U.; and Guest, E. 2026. RCTs & Human Uplift Studies: Methodological Challenges and Practical Solutions for Frontier AI Evaluation. *arXiv preprint arXiv:2603.11001*.

Perez, E.; Huang, S.; Song, F.; Cai, T.; Ring, R.; Aslanides, J.; Glaese, A.; McAleese, N.; and Irving, G. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*.

Perrow, C. 1984. *Normal Accidents: Living with High-Risk Technologies*. New York: Basic Books.

Phang, J.; Lampe, M.; Ahmad, L.; Agarwal, S.; Fang, C. M.; Liu, A. R.; Danry, V.; Lee, E.; Chan, S. W.; Pataranutaporn, P.; et al. 2025. Investigating affective use and emotional well-being on ChatGPT. *arXiv preprint arXiv:2504.03888*.

Rasmussen, J. 1997. Risk Management in a Dynamic Society: A Modelling Problem. *Safety Science*, 27(2–3): 183–213.

Rathje, S.; Ye, M.; Globig, L.; Pillai, R.; de Mello, V.; and Van Bavel, J. 2025. Sycophantic AI increases attitude extremity and overconfidence.

Reuel, A.; Bucknall, B.; Casper, S.; Fist, T.; Soder, L.; Aarne, O.; Hammond, L.; Ibrahim, L.; Chan, A.; Wills, P.; Anderljung, M.; Garfinkel, B.; Heim, L.; Trask, A.; Mukobi,

G.; Schaeffer, R.; Baker, M.; Hooker, S.; Solaiman, I.; Luccioni, A. S.; Rajkumar, N.; Moës, N.; Ladish, J.; Bau, D.; Bricman, P.; Guha, N.; Newman, J.; Bengio, Y.; South, T.; Pentland, A.; Koyejo, S.; Kochenderfer, M. J.; and Trager, R. 2024a. Open problems in technical AI governance. *arXiv [cs.CY]*.

Reuel, A.; Hardy, A.; Smith, C.; Lamparth, M.; Hardy, M.; and Kochenderfer, M. J. 2024b. Betterbench: Assessing ai benchmarks, uncovering issues, and establishing best practices. *Advances in Neural Information Processing Systems*, 37: 21763–21813.

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*, volume 2 of *CHI '23*, 1–18. New York, NY, USA: ACM.

Roytburg, D.; and Miller, B. 2025. Mind the Gap! Pathways Towards Unifying AI Safety and Ethics Research. *arXiv preprint arXiv:2512.10058*.

Russell, S. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking.

Saracini, C.; Cornejo-Plaza, M. I.; and Cippitani, R. 2025. Techno-emotional projection in human–GenAI relationships: A psychological and ethical conceptual perspective. *Frontiers in Psychology*, 16: 1662206.

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*, 160–171.

Shelby, R.; Rismani, S.; Henne, K.; Moon, A.; Rostamzadeh, N.; Nicholas, P.; Yilla-Akbari, N.; Gallegos, J.; Smart, A.; Garcia, E.; and Virk, G. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, volume 24 of *AIES '23*, 723–741. New York, NY, USA: ACM.

Shen, H.; Knearem, T.; Ghosh, R.; Alkiek, K.; Krishna, K.; Liu, Y.; Ma, Z.; Petridis, S.; Peng, Y.-H.; Qiwei, L.; et al. 2024. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions. *arXiv preprint arXiv:2406.09264*, 2406: 1–56.

Shen, J. H.; and Tamkin, A. 2026. How AI Impacts Skill Formation. arXiv:2601.20245.

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*, 1–6.

Soden, R.; Toombs, A.; and Thomas, M. 2024. Evaluating interpretive research in HCI. *Interactions*, 31(1): 38–42.

Strauss, I.; Moure, I.; O'Reilly, T.; and Rosenblat, S. 2025. Real-world gaps in ai governance research. *arXiv preprint arXiv:2505.00174*.

Swuste, P.; Groeneweg, J.; Guldenmund, F. W.; van Gulijk, C.; Lemkowitz, S.; Oostendorp, Y.; and Zwaard, W. 2021. *From safety to safety science: The evolution of thinking and practice*. London: Routledge.

Talbi, M.; and Van Woerden, R. 2025. Reflections on interdisciplinary research in practice: Epistemological conflicts. *European Journal for Philosophy of Science*, 15(4): 66.

Tashakkori, A.; and Teddlie, C. 2021. *Sage handbook of mixed methods in social & behavioral research*. SAGE publications.

Vallor, S. 2016. *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. New York: Oxford University Press.

Walker, A. M.; DeVito, M. A.; Badillo-Urquiola, K.; Bellini, R.; Chancellor, S.; Feuston, J. L.; Henne, K.; Kelley, P. G.; Rismani, S.; Shelby, R.; and Zhang, R. 2024. "what is safety?": Building bridges across approaches to digital risks and harms. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, CSCW Companion '24, 736–739. New York, NY, USA: ACM.

Weidinger, L.; Rauh, M.; Marchal, N.; Manzini, A.; Hendricks, L. A.; Mateos-Garcia, J.; Bergman, S.; Kay, J.; Griffin, C.; Bariach, B.; et al. 2023. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986*.

Weidinger, L.; Uesato, J.; Rauh, M.; Griffin, C.; Huang, P.-S.; Mellor, J.; Glaese, A.; Cheng, M.; Balle, B.; Kasirzadeh, A.; et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 214–229.

Whitney, C. D.; and Norman, J. 2024. Real risks of fake data: Synthetic data, diversity-washing and consent circumvention. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*, 1733–1744.

Yu, C.; Engelmann, S.; Cao, R.; Ali, D.; and Papakyriakopoulos, O. 2026. How should AI Safety Benchmarks Benchmark Safety? *arXiv preprint arXiv:2601.23112*.

Yudkowsky, E. 2008. Artificial Intelligence as a Positive and Negative Factor in Global Risk. In Bostrom, N.; and Cirkovic, M. M., eds., *Global Catastrophic Risks*. Oxford: Oxford University Press.

A Survey Questions

A.1 Background and Experience

Primary Research Area. What is the broad area of AI research that your work primarily aligns with?

- Technical (model training/evaluation, AI safety/alignment, interpretability, proofs)
- Normative (philosophy, ethics theory, conceptual work)
- Sociotechnical (psychology, HCI, STS, empirical AI ethics/fairness, social science)
- Governance (policy, law, economics)

Secondary Research Area. If applicable, is there a secondary area of Safe & Ethical AI research that your work aligns with?

- Technical (model training/evaluation, AI safety/alignment, interpretability, proofs)
- Normative (philosophy, ethics theory, conceptual work)
- Sociotechnical (psychology, HCI, STS, empirical AI ethics/fairness, social science)
- Governance (policy, law, economics)
- N/A
- Other:

Collaboration. How much do you collaborate with each area of Safe & Ethical AI research?

Response scale: 1 (*Never*) — 2 — 3 — 4 — 5 (*Always*)

1. Technical (model training/evaluation, AI safety/alignment, interpretability, proofs)
2. Normative (philosophy, ethics theory, conceptual work)
3. Sociotechnical (psychology, HCI, STS, empirical AI ethics/fairness, social science)
4. Governance (policy, law, economics)

Specific Discipline. Please write your specific discipline (e.g., model evaluation, critical computing, policy…).

Education. What is your highest level of education?

- Bachelor's (current or completed)
- Professional Master's (current or completed)
- Research Master's (current or completed)
- PhD (current)
- PhD (completed)

Sector. What sector do you work in?

- Industry (Private Sector)
- Government (Public Sector)
- Non-Profit
- Academia (Student)
- Academia (Faculty/Researcher)

Role Type. What category does your role fall under?

- Researcher
- Engineer
- Data Scientist / Analysis
- Policy Analyst
- Manager
- Director
- Other:

Country. In which country do you primarily work?

Research Involvement. To what degree are you engaged in research[8] activities in your current role?

- Primarily research-focused (e.g. research is full-time)
- Substantial involvement (e.g. research is part-time)
- Moderate involvement (e.g. research is occasional)
- Little involvement (e.g. keeps up with literature but does not actively do research)
- Manager
- Very little involvement

Experience. How many years of experience do you have conducting research on AI or AI-related topics? *[0–10+]*

A.2 Practice and Aspirations

Practice. How often do you perform the following types of human empirical research in your Safe & Ethical AI work?

Response scale: 1 (*Never*) — 2 — 3 — 4 — 5 (*Always*)

- Human empirical research (overall)
- Quantitative research (e.g., behavioural and survey data)
- Qualitative research (e.g., interview data)
- Lab studies (controlled setting)
- Field studies (real-world setting)
- Experimental studies (comparing multiple conditions)
- Observational studies (examining existing phenomenon)
- Representative studies (generalizable findings)
- Targeted community studies (context-specific

Aspiration. To what extent would you incorporate the following human empirical research methods in your Safe & Ethical AI work? (e.g. how relevant/useful would they be, if incorporated?)

Response scale: 1 (*Definitely Not*) — 2 — 3 — 4 — 5 (*Definitely*)

Repeat the same block of items.

[8] Research is not restricted to traditional academic publishing. It encompasses alternative formats, like data collection for policy and research dissemination via in-depth preprints and blog posts.

A.3 Scenario-Based Perceptions

Skill Erosion (Low Risk, High Imminence). Widespread reliance on AI coding assistants poses the concern that users' programming skills will erode. Future software developers may not know code syntax or be able to debug.

Response scale: 1 (*Very Low*) — 2 — 3 — 4 — 5 (*Very High*)

- How would you characterize the overall risk level? (Risk level = likelihood x severity of harm to humans)
- What is the likelihood that this has already manifested?

In your opinion, how useful would the following methods be in understanding and addressing this risk?

Response scale: 1 (*Not Useful*) — 2 — 3 — 4 — 5 (*Very Useful*)

1. Draw on established theories of learning to write guidelines for AI that supports skill retention.
2. Consult education experts on what types of skills are important to preserve to write as guidelines to the AI.
3. Run large-scale computational benchmarks to measure to what degree the AI's responses either empowers the user, or takes away their decision agency in programming tasks.
4. Collect programming performance tests from company employees longitudinally to track performance changes.

Linguistic Marginalization (Low Risk, Low Imminence). As AIs perform much better in dominant languages like English, there is concern that global linguistic diversity will be diminished, with some languages becoming fully lost in the future.

Response scale: 1 (*Very Low*) — 2 — 3 — 4 — 5 (*Very High*)

- How would you characterize the overall risk level? (Risk level = likelihood x severity of harm to humans)
- What is the likelihood that this has already manifested?

In your opinion, how useful would the following methods be in understanding and addressing this risk?

Response scale: 1 (*Not Useful*) — 2 — 3 — 4 — 5 (*Very Useful*)

1. Train the AI to improve its multi-lingual representation to prevent mapping everything into the semantic space of the dominant languages.
2. Develop a feature to make AI output in the local language and cultural context, and evaluate how users respond to it in real usage scenarios.
3. Create a linguistic diversity benchmark that measures the AI's language output distribution across prompts.
4. Measure the change in linguistic diversity across social media posts over time as a proxy for this collapse.

Decision Biases (High Risk, High Imminence). Decision-making AI tools have the power to influence judicial, medical, and hiring outcomes. There is concern that such systems, trained on biased data, may perpetuate inequalities at a large scale.

Response scale: 1 (*Very Low*) — 2 — 3 — 4 — 5 (*Very High*)

- How would you characterize the overall risk level? (Risk level = likelihood x severity of harm to humans)
- What is the likelihood that this has already manifested?

In your opinion, how useful would the following methods be in understanding and addressing this risk?

Response scale: 1 (*Not Useful*) — 2 — 3 — 4 — 5 (*Very Useful*)

1. Implement mathematical fairness constraints that the AI must satisfy during training.
2. Evaluate how adding a transparency feature to the AI can improve the human judge's fairness.
3. Benchmark the performance of the AI on various social bias datasets.
4. Analyze large-scale dataset of how AI-assisted decisions impacted the individuals evaluated.

Human Extinction (High Risk, Low Imminence). AIs may developed goals misaligned with human values. If this happens, it could pursue those goals in ways humans cannot predict or prevent. This could lead to catastrophic outcomes where humanity loses control over its future: technological trajectories we cannot reverse, resource allocation we cannot influence, or even deliberate human extinction.

Response scale: 1 (*Very Low*) — 2 — 3 — 4 — 5 (*Very High*)

- How would you characterize the overall risk level? (Risk level = likelihood x severity of harm to humans)
- What is the likelihood that this has already manifested?

In your opinion, how useful would the following methods be in understanding and addressing this risk?

Response scale: 1 (*Not Useful*) — 2 — 3 — 4 — 5 (*Very Useful*)

1. Use mechanistic interpretability to ensure the AI's decision-making processes are fully understood and aligned.
2. Engage with expert red teamers and evaluators to test out strategies against misalignment.
3. Construct and perform computation tests to identify signs of potential value misalignment from AI.
4. Interview teams that have deployed powerful AI systems to document instances where systems behaved in concerning ways, to understand early signs of misalignment.

A.4 Barriers to Human Research

Barriers. [If applicable] Please rate how severely these factors have impacted your ability to engage in human empirical research:

Response scale: 1 (*Not Impactful*) — 2 — 3 — 4 — 5 (*Highly Impactful*)

- Lack of access to participants
- Lack of funding
- Lack of time
- Lack of collaborators or advisors
- Difficulty with compliance to policy / ethics (IRB)
- Concerns about methodological fit with research goal
- Concern with the robustness or usefulness of human studies
- Concerns with acceptance to your typical publication venues
- Concerns about impact to career opportunities
- Uncertainty of how to implement human studies
- Uncertainty of how to implement human studies
- Preference to let other disciplines conduct human empirical research.

B Survey Participants Summary

Tables 2-7 summarize the survey participant statistics. Figure 4 shows the participants' report of their primary and secondary research areas and their collaboration frequencies across disciplines.

Region of Work	*n*	%
North America	44	47.3%
Europe	41	44.1%
Oceania	5	5.4%
Africa	1	1.1%
Africa	1	1.1%
Did not disclose	1	1.1%

Table 2: Participants' region of work.

Education Level	*n*	%
PhD (completed)	41	44.1%
PhD (ongoing)	22	23.7%
BSc (completed & ongoing)	13	14.0%
Research MSc (completed & ongoing)	10	10.8%
Professional MSc (completed & ongoing)	7	7.5%

Table 3: Participants' education level.

C Additional Survey Analysis

This section summarizes the statistical analyses for Figure 4 (collaboration), Figure 1 (practice and aspirations) and Figure 2 (AISE scenarios and human research usefulness). Main trends and significant findings are presented in the main text.

Sector of Work	*n*	%
Academia (Faculty/Researcher)	39	41.9%
Academia (Student)	27	29.0%
Non-Profit	18	19.3%
Industry	7	7.5%
Government	2	2.2%

Table 4: Participants' sector of work.

Role Type	*n*	%
Researcher	69	74.2%
Other	7	7.5%
Director	6	6.5%
Manager	4	4.3%
Policy Analyst	3	3.2%
Engineer	2	2.2%
Data Scientist	2	2.2%

Table 5: Participants' role type.

Research Involvement	*n*	%
Full-time	61	65.6%
Substantial	21	22.58%
Moderate	5	5.4%
Little	4	4.3%
Very little	2	2.2%

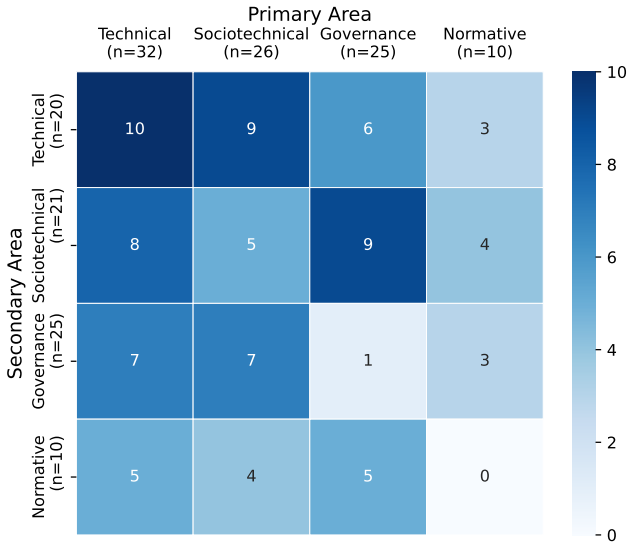
Table 6: Participants' involvement in research.

Years of Experience	*n*	%
1	14	15.1%
2	17	18.3%
3	7	7.5%
4	12	12.9%
5	13	14.0%
6	11	11.8%
7	3	3.2%
8	6	6.5%
9	4	4.3%
10+	6	6.5%

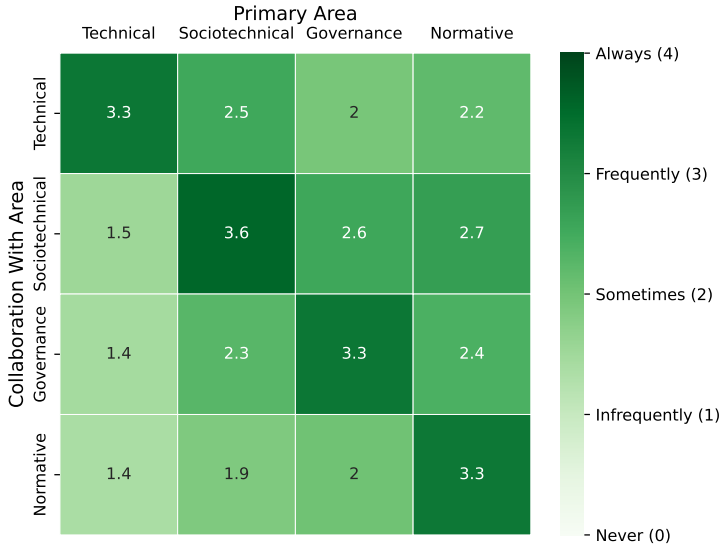
Table 7: Participants' self-reported years of experience in AISE (M=4.44, SD=2.70).

C.1 Statistical Tests for Figure 4 (Collaboration)

Disciplinary-Level Differences To characterize discipline differences for *collaboration frequency* (Likert ratings), we ran Kruskal-Wallis tests across the three main disciplinary groups (*Normative* is excluded from inference due to small *n*), with Holm-corrected Dunn post-hoc tests for significant items. The K-W test is significant ($\chi^2 = 17.12, p < .001$), with contrasts between *Technical* and *Sociotechnical* ($Z = -3.58, p_{adj} = .001$) and *Technical* and *Governance* ($Z = -3.45, p_{adj} = .001$) significant as well.



(a) Distribution of survey participants' primary and secondary (if selected) research areas.



(b) Participants' self-rated collaboration frequency between their primary research areas. *Technical* collaborates significantly less than all other disciplines.

Figure 4: Heatmap of survey participant's research engagement and collaborations across the four research areas.

There is no significance between *Sociotechnical* and *Governance*.

C.2 Statistical Tests for Figure 1 (Practice and Aspirations)

Disciplinary-Level Differences To characterize discipline differences for *practice* and *aspirations*, we ran Kruskal-Wallis tests across the three main disciplinary groups (*Normative* is excluded from inference due to small *n*), with Holm-corrected Dunn post-hoc tests for significant items and Benjamini-Hochberg correction applied across items. Due to the high number of tests (72 combined for K-W and Dunn for both sets of questions), the full set of results is not reported here. The dominant pattern is that *Sociotechnical* researchers rate human research more highly than both *Technical* and *Governance* researchers across nearly all items and both practice and aspiration. For the main category of *All Human Research*, *Sociotechnical* vs *Technical* is (practice: $Z = 4.19, p_{adj} < .001$; aspiration $Z = 2.69, p_{adj} = .01$) and *Sociotechnical* vs *Governance* is (practice: $Z = 2.67, p_{adj} = .01$; aspiration $Z = 2.78, p_{adj} = .01$). Direct contrasts between *Technical* and *Governance* researchers are largely non-significant, with the only exceptions being *Governance* conducting more qualitative research in practice ($Z = 3.25, p_{adj} = .002$) and *Technical* preferring experimental studies in aspiration ($Z = 2.27, p_{adj} = .04$).

Research Dimension Preferences To test for systematic preferences within each methodological dimension, we computed a balance score for each dimension (first pole minus second pole) and tested whether it differed from zero using Wilcoxon signed-rank tests, BH-corrected across the four dimensions. Significance is indicated on column labels in Figure 1. The only significant pole preference was for field over lab studies in researchers' aspiration ratings ($W = 207.0, p_{adj} < .001$), suggesting a collective desire for more ecologically valid work than is currently practiced.

C.3 Statistical Tests for Figure 2 (Scenarios)

Scenario Classification Validation To validate our scenario classifications, we first examined whether participants' ratings aligned with the intended structure of the four scenarios across two dimensions: risk level (low vs. high) and imminence (future vs. current). Across all participants, current scenarios were rated as significantly more imminent than future scenarios (means 2.86 vs. 1.60; Mann-Whitney $U = 6894, p < .001$), and high-risk scenarios were rated as significantly riskier than low-risk scenarios (means 2.97 vs. 2.27; $U = 10047, p < .001$). This pattern held when splitting by risk level: both low-risk (fut. mean 1.94 vs. curr. mean 2.66; $U = 2552, p < .001$) and high-risk (fut. mean 1.26 vs. curr. mean 3.07; $U = 1036, p < .001$) scenarios showed the expected imminence ordering. Together, these results suggest that participants broadly agreed with our a priori classification of the scenarios.

Discipline Differences in Scenario Classification We next examined whether ratings differed across disciplinary backgrounds. At the scenario level, a Kruskal-Wallis test revealed a significant between-discipline difference only for imminence ratings of the future low-risk scenario ($H = 7.88, p = .049$); no other scenario-level comparisons reached significance. Pairwise follow-up tests (Bonferroni-corrected $\alpha = .0083$) identified a single significant contrast:

Technical researchers rated this scenario as less imminent than *Governance* researchers (means 1.60 vs. 2.48; $U = 222, p = .008$). To increase statistical power, we repeated the analysis on ratings aggregated across all four scenarios. Here, a significant discipline effect emerged for imminence (H = 11.86, p = .008) but not risk ($H = 5.00, p = .172$). Pairwise tests showed that *Governance* researchers rated imminence systematically higher than both *Technical* (means 2.58 vs. 2.02; $U = 4554, p = .002$) and *Sociotechnical* (means 2.58 vs. 2.12; $U = 3722, p = .005$). The results suggest that while participants broadly agreed on the relative ordering of scenarios by risk and imminence, *Governance* researchers systematically perceived AI risk scenarios as more imminent than their *Technical* and *Sociotechnical* counterparts.