

Responsible AI as Hegemony: Epistemic Exclusion, Technosolutionist Conformity, and Performative Compliance Among AI Practitioners in India

Atmadeep Ghoshal¹, Arshia Dutta², Ruba Abu-Salma¹, Sanjay Modgil¹, Martim Brandão¹

¹King's College London

²Royal Holloway, University of London

atmadeep.ghoshal@kcl.ac.uk arshia.dutta.2023@live.rhul.ac.uk ruba.abu-salma@kcl.ac.uk sanjay.modgil@kcl.ac.uk
martim.brandao@kcl.ac.uk

Abstract

This paper investigates how AI practitioners in India perceive and experience Responsible AI (RAI) through mixed-methods qualitative research with participants from industry and academia. We argue that RAI practice in India operates as a doubly hegemonic system in which global Western frameworks that privilege technical and procedural conceptions of responsibility converge with local caste-gender hierarchies to determine what counts as legitimate ethical knowledge. Drawing on interviews with 18 practitioners, we show how this double hegemony operates through three interlocking mechanisms. Counter-hegemonic perspectives, particularly corrective justice framings articulated by caste-marginalised practitioners, are shadowed through epistemic silencing. Surviving critique is channelled into technosolutionist forms that leave organisational power intact. The system is sustained through performative compliance, where organisations display RAI commitments while protecting institutional priorities from ethical interruption. These mechanisms produced conditions in which practitioners regularly recognised harm but lacked the authority, protection, or collective support to act on that recognition. We propose that countering RAI hegemony requires collective and enforceable infrastructures, including Corporate Social Responsibility (CSR)-embedded ethics residencies, anonymous reporting mechanisms, and vernacular harm taxonomies grounded in community knowledge rather than imported risk categories.

Introduction

Artificial intelligence (AI) has rapidly transformed from lab-researched technologies into foundational infrastructure shaping decisions in healthcare (Panch, Mattie, and Celi 2019), finance (Cao and Zhai 2022), education (Holmes and Tuomi 2022), and public services worldwide (Chen et al. 2021; Sheikh, Prins, and Schrijvers 2023). As ethical, social, and political ramifications grow more pronounced (UNESCO 2021), Responsible AI (RAI) has emerged to guide the ethical, transparent, and socially aligned development of AI systems. However, RAI discourse remains contested, with disagreements over which responsibilities to prioritise, how to operationalise them, and whose values should define standards (Jobin, Ienca, and Vayena 2019; Morley et al. 2021). Much RAI literature reflects Euro-American

perspectives and institutional priorities (Sambasivan et al. 2021), articulated through voluntary codes, fairness metrics, and auditing protocols (Winfield and Jirotko 2018; Vakkuri et al. 2019). Critics argue that such approaches risk reducing ethics to technical checklists, enabling ethics washing rather than fostering genuine accountability (Floridi 2019; Wagner 2019; Hao 2019), while many frameworks assume Western principles are universally applicable (Ghoshal et al. 2025), a stance criticised as epistemic imposition (Bhatt et al. 2022; Khanuja, Ruder, and Talukdar 2023). AI practitioners who design, develop, and deploy AI systems receive limited scholarly attention (Holstein et al. 2019; Madaio et al. 2020). Recent studies reveal that practitioners often articulate ethical concerns but lack the authority to enforce change (Widder et al. 2023), face organisational barriers that constrain implementation (Madaio et al. 2020), and bear individual responsibility for systemic issues (Ali et al. 2023). However, this practitioner-focused research has predominantly examined Euro-American contexts, leaving unclear how RAI unfolds in non-Western settings.

India provides a compelling case for examining RAI beyond Western contexts. As a global technology hub (Jalote and Natarajan 2019) and AI back office (Sarkar 2023), India hosts major software operations for multinational corporations while positioning itself as a development centre through initiatives such as AI for All (NITI Aayog 2018; Matthan and Javed 2025; India 2025). Indian practitioners contribute significantly to global AI systems while working within distinct socio-technical conditions marked by digital divides, linguistic diversity, entrenched hierarchies of caste, gender, and religion, and weak regulatory enforcement (Khanuja, Ruder, and Talukdar 2023). Despite India's central role in global AI development, practitioner perspectives remain largely unexplored. We report on a two-stage empirical investigation addressing this gap. We surveyed 16 AI practitioners to assess RAI awareness and guide interview design, then conducted in-depth interviews with 18 practitioners from industry and academia who have lived and worked in India. Our research questions are as follows:

- **RQ1:** How do AI practitioners in India perceive RAI?
- **RQ2:** What challenges do practitioners face in adopting RAI practices?
- **RQ3:** What structural conditions produce RAI hegemony

in India, and what collective infrastructures are needed to counter them?

Our findings reveal that practitioners experience RAI as an individual ethical burden without institutional support, framing it as either an ideology or a toolkit, while marginalised practitioners uniquely conceptualise it as corrective justice. We document how organisations prioritise performative compliance while enabling systematic data misuse, and how cultural narratives position RAI as feminine and a career liability. Our work makes four contributions: (1) empirical insights into how RAI is framed and contested in Indian AI practice; (2) identification of structural barriers at individual, organisational, and societal levels; (3) a theorisation of these challenges as a doubly hegemonic system in which global Western RAI frameworks and local caste-gender hierarchies converge to determine legitimate ethical knowledge, operating through epistemic silencing, technosolutionism, and performative compliance as interlocking mechanisms; and (4) a proposal of collective and enforceable infrastructures for countering the challenges.

Related Work

RAI: Concepts and Worldviews

Responsible AI addresses ethical concerns arising across AI development and deployment, building on value-sensitive design (Friedman 1996) and early research on computational bias (Friedman and Nissenbaum 1996). Its central commitments include fairness (Madaio et al. 2022; Deng et al. 2023b), transparency (Toy 2024; Corbett and Denton 2023), accountability (Novelli, Taddeo, and Floridi 2024), and privacy and security (Chowdhary, Ingber, and Andalibi 2025). These commitments appear across industry frameworks (Bellamy et al. 2018) and scholarship on algorithmic justice (Marjanovic, Cecez-Kecmanovic, and Vidgen 2022) and participatory design (Birhane et al. 2022). Critical scholarship, however, shows that RAI frameworks may extend Western norms while excluding alternative knowledge systems, thereby reproducing algorithmic colonisation (Adams 2021). Research on epistemic injustice similarly demonstrates how AI can produce testimonial and hermeneutical harms by marginalising non-Western epistemologies (Kay, Kasirzadeh, and Mohamed 2024). RAI therefore remains contested across different worldviews (Jakesch et al. 2022), yet limited attention has been paid to how these tensions are shaped by professional practice and organisational culture (Ruster and Davis 2025), particularly beyond the Global North (Okolo, Dell, and Vashistha 2022). We address this gap by examining how AI practitioners in India understand, negotiate, and identify with RAI in their everyday work.

Practising RAI

Translating RAI principles into practice remains challenging across roles including data scientists, engineers, product managers, and ethicists (Holstein et al. 2019). Practitioners may recognise ethical problems yet lack the authority to intervene (Madaio et al. 2020), operate within organisational power asymmetries (Widder et al. 2023), and encounter cross-functional communication barriers and techni-

cal decisions that constrain which models are considered acceptable (Lu et al. 2022; Muralikumar and McDonald 2024; Ashktorab et al. 2023). These difficulties reflect structural rather than solely individual failures. Organisations often concentrate ethical responsibility on practitioners through ethics washing (Ali et al. 2023), while institutional dynamics restrict their capacity to act (Widder et al. 2023; Deng et al. 2023a). Although RAI guidelines also fail to translate into practice in Global North settings (Sadek et al. 2024), India's combination of weak enforcement, caste and gender hierarchies, and epistemic asymmetries produces a distinct form of this failure, reducing complex inequalities to algorithmic problems (Birhane 2025) and exposing the limits of AI auditing as a sufficient response (Birhane et al. 2024).

RAI in South Asia

While RAI scholarship has developed theoretical frameworks and documented practitioner challenges, it remains largely grounded in Euro-American contexts. Research in India demonstrates why these frameworks cannot be transferred without attention to local sociotechnical conditions. Algorithmic fairness approaches often fail to reflect Indian realities (Sambasivan et al. 2020), while AI systems may reproduce Western norms that homogenise expression (Agarwal, Naaman, and Vashistha 2025). Cultural models of self also shape patterns of interaction (Ge et al. 2024), and Indian users relate differently to AI authority and trust (Kapania et al. 2022). Healthcare studies make these tensions visible through explainable AI tools for community health workers (Solano-Kamaiko et al. 2024), chatbots deployed through Accredited Social Health Activists (Ramjee et al. 2025), and ethnographic research on maternal health programmes (Ismail et al. 2023a). Together, these studies show how technical interventions encounter culturally specific constraints involving usability, literacy, and data availability (Anuyah et al. 2024; Ismail et al. 2023b). These challenges are situated within a policy landscape that balances AI innovation against risk mitigation (Biju and Gayathri 2023), alongside linguistic diversity that constrains deployment while also creating opportunities to bridge social divisions (Kalyanakrishnan et al. 2018).

Placing India within broader Global South debates reveals related structural patterns. African practitioners face weak enforcement and infrastructural barriers (Okolo 2023), which Birhane characterises as forms of algorithmic colonisation that reproduce dependency (Birhane 2023). In India, these dynamics are shaped by resource constraints and entrenched inequalities (von Braun and Baumüller 2021), even as the state advances claims around digital sovereignty (Png 2022). Research from Bangladesh similarly documents weak enforcement, voluntaristic data practices, and the individualisation of ethical labour (Mayeesha, Islam, and Ahmed 2025; Elahi, Srinivasa, and Emruli 2025). India and Bangladesh therefore converge in the gap between stated principles and practice, but differ through India's sovereign AI framing and the role of caste in producing distinctive forms of epistemic exclusion. This literature establishes the need for situated, practitioner-informed RAI approaches that address structural inequality and power rather than import-

ing universalist frameworks from the Global North.

Theoretical Framework

We adopt Gramsci's theory of hegemony as our primary analytical framework (Gramsci 1998). Gramsci argues that dominant classes maintain power not through coercion alone, but through the production of consent, whereby particular ideologies become naturalised as common sense, making alternative ways of thinking difficult to articulate within existing institutional arrangements. Central to this framework is the concept of transformism, through which counter-hegemonic critique is absorbed and redirected into forms that ultimately leave dominant structures intact (Gramsci 1998). We supplement this with two related analytical frameworks. Dotson's 2011 account of epistemic violence provides tools for examining how organisational structures silence practitioners through testimonial quieting and smothering, processes through which speakers are denied knower status within the very settings where their knowledge is most consequential. Morozov's 2013 concept of technosolutionism extends transformism into contemporary technological contexts, identifying how political and structural problems are systematically reframed as technical optimisation problems. Together, these frameworks orient our analysis towards questions of institutional power, knowledge legitimacy, and the conditions under which ethical critique can and cannot be heard. We apply them to the Indian context where, as Brown et al. (2025) have begun to show, cultural hegemony shapes which concerns within AI research are treated as legitimate. The intersection of global Western RAI discourse and local caste-gender hierarchies creates a doubly hegemonic configuration that our data allows us to examine empirically.

Methodology

Our study used a two-phase sequential design, beginning with a survey of Indian AI practitioners that assessed general awareness of RAI and informed subsequent semi-structured interviews. We use AI practitioner" to include those professionally involved in the design, development, deployment, evaluation, or governance of AI systems, spanning technical roles such as machine learning engineers and data scientists and non-technical roles such as product managers, policy researchers, and ethics reviewers (Rismani and Moon 2023). We also include academic researchers because research and deployment in the Indian AI ecosystem frequently overlap, with academics contributing to applied projects and industry practitioners participating in research-adjacent governance work (Madaio et al. 2022). We distinguish responsibility" as practitioners' perceived normative obligations for the ethical conduct of their work from "accountability" as the organisational or structural mechanisms through which such obligations are attributed, enforced, and redressed (Novelli, Taddeo, and Floridi 2024). This distinction allows us to examine cases in which practitioners experience responsibility without corresponding institutional accountability. Participants had lived and worked in India and had experience in AI development, deployment, or governance, including

practitioners currently based in India and those who had relocated while maintaining ties to the Indian AI ecosystem.

Survey

Instrument Development and Pilot Testing The survey assessed awareness of RAI principles and practices and informed the interview guide. It covered demographics and professional background, engagement with RAI frameworks, organisational structures and resources, familiarity with Indian AI policy, and self reported implementation challenges. We first conducted cognitive interviews with five convenience sampled participants from our professional and personal networks, comprising three industry practitioners and two academics, to identify culturally appropriate terminology and refine questions on organisational structures, Indian AI policy, and cultural considerations. We then piloted the survey with three additional practitioners, two from industry and one from academia. Feedback on clarity, length, and functionality led to minor wording changes and clarifying examples for policy questions. Cognitive and pilot interview participants were excluded from the main dataset to prevent potential bias. Following piloting, we administered the survey through Google Forms to 16 participants, evenly divided between industry and academia, whose roles included machine learning engineers, product managers, postdoctoral researchers, and assistant professors. Recruitment drew on professional networks and alumni associations from Indian technical institutions, LinkedIn, WhatsApp, specialised forums, partnerships with AI labs and technology companies, conference networks, and snowball sampling. Data collection ran for six weeks between April and May 2025, and completion took approximately 11 minutes. Participants were offered a 500 INR, or 5 USD, Amazon gift card, although only three accepted, with most declining because they wished to contribute to academic research. All participants were of Indian ethnicity. Thirteen were based in India and three in the UK, EU, and US with lived experience in India. Among those based in India, two identified as caste minorities, two as a religious minority, and one as an ethnic minority. Demographics appear in Table 1. We analysed open ended responses qualitatively. Two researchers independently coded responses inductively, met regularly to discuss decisions, and resolved discrepancies through consensus. Codes were iteratively refined into three broader patterns concerning gaps in institutional accountability, divergent framings across roles, and limited policy awareness, which directly informed the interview protocol by identifying tensions for deeper qualitative exploration.

Interviews

Protocol Development and Participant Recruitment We iteratively developed our interview guide based on the survey findings and RAI implementation literature, translating the three survey patterns into targeted probes. These probes examined organisational power dynamics and compliance environments in the absence of institutional support; explored role-specific framings and the gap between organisational rhetoric and practitioners' lived experiences; addressed how India's social structures shape approaches to

Interview participants

ID	Domain	AI Exp.	Age/Gen.	Location
P1	Acad.	6	30–35/W	India
P2	Ind.	4	25–30/M	EU
P3 ^C	Acad.	10	35–40/M	India
P4 ^R	Ind.	4	25–30/W	India
P5	Acad.	12	40–45/M	India
P6	Ind.	7	30–35/NB	India
P7	Acad.	5	30–35/W	EU
P8	Acad.	4	25–30/M	India
P9	Ind.	9	35–40/W	India
P10	Acad.	8	30–35/W	India
P11	Ind.	4	30–35/M	India
P12	Ind.	8	35–40/W	US
P13	Acad.	4	25–30/W	India
P14	Acad.	11	35–40/M	India
P15	Ind.	6	30–35/W	India
P16	Ind.	4	25–30/NB	India
P17	Acad.	4	25–30/M	India
P18 ^C	Ind.	5	30–35/W	UK

Survey respondents

ID	Domain	Role	Age/Gen.	Location
S1	Acad.	Assistant Professor	30–35/W	India
S2 ^C	Acad.	Lecturer	35–40/M	India
S3 ^R	Acad.	Postdoctoral Fellow	30–35/W	India
S4	Acad.	Associate Professor	40–45/NB	India
S5	Acad.	Assistant Professor	30–35/W	EU
S6 ^C	Acad.	Lecturer	35–40/NB	India
S7	Acad.	Senior Lecturer	40–45/W	India
S8	Acad.	Postdoctoral Fellow	30–35/W	UK
S9 ^E	Ind.	ML Engineer	25–30/M	India
S10	Ind.	Data Scientist	30–35/W	India
S11 ^R	Ind.	Product Manager	30–35/M	India
S12	Ind.	AI Researcher	35–40/W	US
S13	Ind.	Software Engineer	25–30/M	India
S14	Ind.	Data Engineer	25–30/W	India
S15	Ind.	Policy Specialist	35–40/W	India
S16	Ind.	UX/AI Practitioner	30–35/NB	India

Abbreviations: Acad. = Academia; Ind. = Industry; W = Woman; M = Man; NB = Non-binary; AI Exp. = years of experience with AI. Superscript denotes self-reported minority status: ^C = caste minority; ^R = religious minority; ^E = ethnic minority.

Table 1: Participant demographics.

RAI; and elicited practitioners’ perspectives on institutional reforms and public engagement. Interviews followed a semi-structured format and included 18 participants. Participants were recruited through targeted convenience sampling, including outreach to women-in-tech networks and technology unions composed exclusively of caste and religious minority technology professionals in India, to ensure representation of underrepresented practitioners. Professionals from smaller organisations and those with international experience were recruited separately through personal networks.

Piloting and Data Collection We piloted our interview guide with four practitioners (two from academia, two from industry) to test the flow, timing, and effectiveness of the questions. Pilot interviews revealed the need for additional questions on power dynamics and cultural considerations, which we incorporated into the final protocol. Pilot interviewees were excluded from the main analysis to prevent contamination bias, as exposure to preliminary protocol versions could influence response patterns and compromise data independence (van Teijlingen and Hundley 2002). Following pilot testing, we conducted 18 semi-structured interviews with practitioners representing diverse demographic and professional backgrounds. Interviews were conducted in English or Hindi based on participant preference. Pilot interviewees, survey participants, and main interviewees were three distinct groups with no overlap. All interviews were audio-recorded and transcribed with consent, lasting 45–70 minutes (mean: 52 minutes), generating approximately 15 hours of data. All participants received INR 1,000 (approximately £10) compensation. The sample included 18 participants evenly split between academia and industry, aged from their mid-20s to mid-40s, with 4–12 years of AI experience.

14 participants were based in India, and 4 were based in the EU, UK, and US (of Indian ethnicity and with lived experience in India). Our sample included men, women, and non-binary participants. Three Indian participants reported minority status (caste, religious, or ethnic). Demographics are detailed in Table 1.

Data Analysis Interviews were transcribed verbatim. We employed inductive thematic analysis following Braun and Clarke’s six-phase framework (Braun and Clarke 2006). Two researchers collaboratively ensured rigour and reliability through repeated reading and note-taking to familiarise themselves with the data. Both researchers independently coded all transcripts, meeting regularly to refine the codebook and resolve discrepancies through consensus until agreement and code saturation were reached (i.e., no new codes were developed). Both researchers then coded all transcripts using the finalised codebook and resolved any remaining disagreements through discussion. The researchers then searched for themes by grouping related codes and identifying patterns. They systematically reviewed and refined preliminary themes to ensure they accurately represented the coded data and addressed the research questions. Finally, they defined and named the final themes with clear definitions and scope, maintaining detailed records of all analytic decisions and theme development throughout the process. This systematic approach enabled a comprehensive understanding while ensuring rigour through collaborative interpretation and consensus-building. Sample size in qualitative research is debated (Braun and Clarke 2006), and data saturation is inherently subjective (Braun and Clarke 2019). We assessed data saturation by iteratively refining our themes until we determined that we had fully captured the

meanings necessary to address our research questions. Our sample size of 18 participants aligns with qualitative guidelines that prioritise depth over generalisability (Vasileiou et al. 2018).

Survey Findings

The survey (n=16) scoped how Indian practitioners understood RAI and identified issues for deeper exploration in interviews. Three patterns informed the interview protocol. First, although 8/16 respondents reported internal RAI frameworks, many described limited authority, time, and resources for applying them. Common concerns included Global North bias (9/16), lack of localised examples (9/16), insufficient guidance (5/16), and language barriers (4/16). Participants also described responsibility as largely documentary, with expectations recorded in policies but little corresponding capacity to act. We therefore added probes on organisational power, resource constraints, and compliance without institutional support. Second, industry respondents emphasised auditing and compliance, whereas academics more often foregrounded justice and equity. RAI engagement was also predominantly industry-oriented (12/16), with participants noting that it appeared more often in presentations than in everyday workflows. Interviews consequently examined how organisational setting and professional role shaped RAI practice and the gap between formal commitments and operational practice. Third, 9/16 respondents were initially unaware of India's national RAI strategy, although several reviewed it when prompted by the survey. Their critiques focused on its derivative character and limited attention to caste, linguistic diversity, and bottom-up participation. At the same time, 15/16 considered policy important. We therefore added questions on caste, language, economic disparity, public participation, and institutional reform. Together, these survey findings shaped an interview protocol focused on how organisational structures, professional roles, and Indian social contexts condition the practice of RAI.

Interview Findings

Interviews enabled us to investigate the tensions identified in the survey and capture how practitioners negotiate RAI in everyday contexts. We organise our findings across individual, organisational, and societal levels, building on prior qualitative work on RAI practice that distinguishes individual ethical burdens from organisational constraints (Widder 2024; Ali et al. 2023). We extend this framework by introducing a societal level to capture cultural narratives and educational dynamics that precede workplace entry and are particularly salient in the Indian context. Addressing RQ1, we examine how practitioners conceptualise RAI at the individual level. Addressing RQ2, we document the structural and institutional challenges they encounter at the organisational and societal levels, including weak governance, systematic data misuse, and cultural barriers that shape RAI practice before practitioners enter professional settings.

Different Conceptions of Responsibility

Before practitioners could describe what constrained Responsible AI, they first had to articulate what they understood it to mean. These definitions varied widely, revealing different assumptions about the purpose of RAI and the contexts in which practitioners engaged with it.

Ideology, Toolkit, or Justice Practitioners' definitions of Responsible AI reflected fundamentally different understandings of what the concept was intended to achieve and whom it should serve. For the majority of academic participants (n=8) and some industry participants (n=2), RAI was inseparable from a broader normative vision of what India should become as a society. P1 articulated this directly, stating that *"Responsible AI for a nation like ours has to be about uplifting the poor, helping the honest and safeguarding the vulnerable. It cannot be just about bias and fairness, we have to go past that."* Others in this group emphasised national integrity, linguistic diversity, and obligations towards communities historically excluded from technological development, positioning RAI as a matter of social commitment rather than technical compliance. For some industry practitioners (n=7) navigating delivery timelines and client expectations, the same question produced a categorically different response. P11 described RAI in terms resembling technical debt management: something real and eventually necessary, operationalised through audits and benchmarks rather than normative commitment. While considered important in principle, RAI was also viewed as potentially restrictive given the pace at which products were expected to move. The most substantively distinct framings came from practitioners whose relationships with Indian institutions had been shaped by direct experiences of exclusion. P3 and P18, both from caste minority backgrounds, described RAI as a form of corrective justice that emerged not through scholarly engagement but through accumulated experiences of discretionary institutional power operating without accountability. P3 recalled watching a relative spend hours at a ration store arguing with an official who refused to provide the designated quantity of wheat and rice after the relative caught him manipulating the weights of his beam balance. For P3, the prospect of AI reducing such discretionary power represented a felt necessity rather than a technocratic ambition. P18 extended this framing when asked what Responsible AI should accomplish, stating that *"I actually want AI to take over the role of people who abuse the system and take advantage of those who are weak. And when that happens, Responsible AI will have the most important role to play."* Both participants emphasised that this vision depended on responsible design, noting that irresponsible systems could simply encode the same prejudices at scale, removing the human face without removing the harm. P14 situated all three framings within a shared structural problem when discussing governance, arguing that the concentration of authority among political leaders, executives, and academic gatekeepers was precisely what limited RAI effectiveness in practice. They argued that *"Responsible AI has to be decentralized if you actually want it to work. Otherwise it will become like the other laws in the country. It would look*

good on paper, but essentially will have no impact at all."

Personal Ethics and Professional Compromise Practitioners in industry and academia consistently described responsibility as central to their professional identity, but their accounts became more complex when they reflected on what this meant for their day-to-day work. The ethical weight practitioners attached to their roles extended beyond formal job descriptions. P9, who worked at a legal AI company specialising in analysing data for case summaries, explained that decisions about data privacy and algorithmic fairness carried significance beyond the workplace because she understood how the systems she built could affect people's access to credit, employment, and surveillance. She stated, *"My job at my company is to automate report creation for certain specific legal cases. I know if something goes wrong there, my work can affect a person's life and family adversely"*. P15, who worked in healthcare AI, described a similar level of awareness, noting that systems developed by her team to recommend patient treatments to doctors had real-world consequences and that ignoring these implications felt like a moral failure, regardless of her formal responsibilities. She explained, *"I am aware if my team's product recommends something to a doctor which does not consider the patient's state of affairs or financial capacity, then the patient will be burdened with an unsound decision. Now the company has not asked me to ensure this, but I cannot not do it."* Both accounts reflected not only ethical commitment but also an acute recognition of the gap between practitioners' sense of responsibility and the organisational structures available to support it. P11 captured this tension, stating that *"at the end of the day I need this job and I need to be able to survive. I need to be able to put food on my plate. But I also need to be someone who I am."* This tension highlighted the difficulty of sustaining personal ethics alongside professional demands.

Gendered Silencing Women practitioners in industry and academia described workplace dynamics that extended beyond a lack of support, reflecting patterned forms of epistemic marginalisation that constrained how their contributions were received and recognised. P1 recounted an instance in which she raised concerns about algorithmic discrimination against women in India during a staff room discussion, only for a senior colleague to reframe her argument as personal sentiment rather than technical critique. This reframing undermined both the content and authority of her contribution. Similarly, P10 described being routinely spoken over and ignored in syllabus team discussions for gender and AI ethics courses, leading her to document concerns in writing because written records were harder to dismiss or overwrite. These dynamics also appeared in institutional and policy contexts. P13, while serving on a state government advisory committee on AI policy, raised concerns about how proposed welfare and policing systems might reproduce existing patterns of gender-based exclusion. A senior male professor dismissed her intervention, arguing that focusing on women's issues was less important than priorities likely to attract state investment and recognition. This framing positioned enfranchisement as secondary to capital-oriented

considerations and was not challenged by others present. Industry accounts further illustrated how marginalisation was normalised through informal practices. P15 described identifying gendered patterns in a user-profiling system that assigned labels such as "emotional", "frisky", and "volatile" disproportionately to accounts presenting as women. When she raised this concern, her manager dismissed it as not legally actionable and trivialised the observation through a personal remark that elicited laughter, effectively ending further discussion. Here, humour functioned as a mechanism of deflection that diffused critique while reinforcing the legitimacy of the system. Participants further highlighted how these experiences were intensified through intersections of caste, religion, and gender. P4, an AI researcher from a religious minority community, described a "double impossibility" in which she had to establish the legitimacy of her knowledge before advancing substantive arguments. She explained, *"This happened sometime during a company skills seminar and workshop where I was trying to point out to a speaker that we need to build benchmark datasets from the ground up by understanding people's experiences and focus more on those historically oppressed. The speaker immediately retorted back saying that it is easy for women to say all this because they get all the advantages in a company compared to men and being a religious minority like you means no one can argue due to fear of law."* P18 similarly reflected on these intersecting constraints, noting that *"as a Dalit woman AI practitioner, it was very difficult for me in India to be in the field. When you are a woman and a caste minority and you have to fight these spontaneous challenges where people will irritate you and write stuff that would seem like minorities want to be the centre of attention and you cannot go with the flow."* Together, these accounts illustrate how structural and interactional dynamics constrain not only participation but also the conditions under which knowledge claims are recognised and legitimised within AI practice.

Organisational Structures Enabling Irresponsibility

Practitioners' individual struggles were embedded within organisational conditions that not only failed to support ethical practice but actively constrained it. Across accounts from industry and academia, participants described structural arrangements that limited responsible AI practice while maintaining its rhetorical appearance.

Economic Incentives and Performative Compliance Culture Participants with experience across Indian and European organisational contexts drew consistent distinctions in how responsible practice was resourced and incentivised. P2 described Indian startups where venture capital cycles and scaling targets positioned responsible practices as costs rather than standards, with ethical risks considered only when directly raised by clients. This produced what multiple participants characterised as a performative compliance culture, where ethical commitments appeared in documentation and slide decks but were addressed reactively, and where raising concerns carried professional risk rather than institu-

tional support. These conditions were particularly visible in dataset preparation. P12 described overseeing post-training work in India-origin AI labs where annotators operated under throughput targets that made careful deliberation structurally impossible, with managers instructing annotators to mark ambiguous outputs as acceptable rather than declining or refining them. In at least two instances she was aware of, preference pairs were fabricated when pipelines fell behind schedule, with managers explaining that *“the client doesn’t care what’s in the pairs as long as we have the numbers right.”* The same conditions shaped safety evaluation practices. Findings that implicated client design decisions were removed from red-team reports before delivery, as P9 described, with final reports stating that models had passed without significant findings. When P6 raised concerns that a model was producing derogatory terms such as “hijra” and “kinnar” for queer communities, the response invoked both regulatory absence and cultural normativity to dismiss the concern. The manager stated that *“these communities are a menace to society, why do you want to give them so much value, our culture is not supportive of it anyway, there is no point in being extra moral. If we care about them, our clients will not care extra for us.”* This framing positioned the concern as not only unnecessary but also culturally inappropriate, effectively removing it from the domain of legitimate technical critique. P2 situated these dynamics within a broader economic logic, noting that *“when an economy is driven only by short-term profits, implementing Responsible AI becomes very difficult. That is what is happening in India. [...] Responsible AI is expensive and stifles innovation.”*

Weak Governance and Institutional Complicity Participants described governance structures that existed formally but, in practice, served to evade accountability. Compliance roles were often held by personnel without technical expertise or organisational authority to challenge senior leadership. P11 recounted how a concern he raised about model hallucination was evaluated not on ethical grounds but on whether the resulting harm would be attributable to the company. He described the issue saying, *“The company where I am working right now is building an AI-powered advisory tool for financial planning that is deployed directly to retail customers. During internal testing, we consistently observed the model confidently generating fabricated fund names, inaccurate return figures, and nonexistent regulatory provisions. When I flagged that customers could make real financial decisions based on this false information, my supervisor directly told me as long as we have a disclaimer in the terms of service we are fine and at the end of day you are responsible for compliance, not me.”* The same transactional logic operated at senior levels of organisational decision-making. P12 described a case where a post-deployment audit revealed that a model was hallucinating citations to medical guidelines in a clinical decision-support product, with leadership comparing the projected cost of retraining against the likelihood that any misdiagnosis could be traced back to the system, and proceeding with the existing model while framing the calculation as *“just how risk is managed here, it’s not personal.”* Participants also described politi-

cal relationships as an informal layer of governance protection specific to the Indian context. When companies anticipated scrutiny for harmful model outputs, including content promoting political polarisation or limiting minority rights, leadership sometimes managed this risk through proximity to politicians who could influence enforcement bodies rather than through remediation. For startups seeking government grants, such proximity was sometimes a condition of access. P9 described third-party evaluation firms being warned not to criticise sovereign AI models too harshly. P16 described a mental health chatbot evaluated for domestic violence support that was found to centre spirituality and cultural deference to the spouse rather than clinical referral or rights-based framing, with the evaluation team declining to formally flag this because *“we don’t want to malign their company, if we make this a big issue we probably won’t get another contract from them.”* P11 situated these dynamics within a broader structural condition, observing that *“I don’t think we can ever get rid of malpractices unless we remove the unnecessary bureaucratic issues in India. Imagine even if Responsible AI legislation is passed, implementing will always become a hurdle and some organisations will always go for the corruption route.”*

Systematic Data Misuse Participants described data procurement and reuse practices that were not exceptional but constitutive of AI development in the environments where they had worked. P9 described annotation firms as the connective tissue of a supply chain in which mobile carriers, point-of-sale companies, marriage agencies, and political consulting firms supplied raw data consumed by AI development teams. The same datasets were copied, resold, and reused across multiple projects without the individuals concerned ever being informed. A participant working in an India-origin AI lab described the absence of protocols governing data handling, noting that *“I think no one in these small or mid size labs in our country follow any protocol. You will find data related to the most vulnerable populations without any kind of masking or redaction and you can easily make it out who they are and what they have said.”* Data collection practices in the field reproduced similar conditions. P15 described exercises in rural areas where residents provided voice recordings and household information under the framing that participation would improve government services, with consent forms presented only after recording sessions and handed to participants who in some cases could not read them. The same dataset was subsequently used for three separate model training tasks across different clients, none connected to the government services framing under which it had been collected. P6 described a parallel practice in which lower-income urban residents were recruited to record telephone conversations with third parties who were not informed, with the audio then sold as naturalistic conversational data. LGBTQ+ individuals were specifically targeted for voice data collection by some organisations, with those commissioning the work framing this to data collectors as *“someone has to do this, we need to uphold our culture.”* When P9 asked about data provenance, she was told it was proprietary or outside her need to know, describing the

broader condition this reflected by saying that *“the mentality of cutting costs on all accounts makes us use the same datasets everywhere and nobody cares about consent then.”* The regulatory environment provided no effective deterrent. P15 described a case in which a company found to have illegally collected biometric data was fined an amount smaller than what it had already earned from that dataset, creating conditions in which violations became calculated business decisions. Practitioners working within these conditions described a form of compelled complicity, with one participant noting that *“people have no idea their data is being traded, labelled, and fed into AI systems. They are not even part of the conversation, yet they are the raw material.”* Another described the professional logic sustaining silence among those who suspected data had been improperly sourced, saying that *“you know something is wrong with how the data came to you, but you also know if you raise it, the project gets delayed, the client gets angry, and you become the problem. So most people just look the other way.”*

Cultural and Societal Barriers

The organisational dynamics practitioners described were sustained and reproduced by educational norms and civic structures extending beyond individual workplaces. Participants’ accounts of these broader conditions revealed how the foundations for responsible AI practice were shaped before practitioners entered professional settings.

RAI as Feminine Work and Career Liability Academic participants described student resistance to RAI as shaped by overlapping cultural and professional pressures rather than simple disinterest. Students consistently framed engagement with ethics as a marker of technical weakness and as belonging more appropriately to the social sciences than to computer science and engineering. P1 noted that students often characterised serious engagement with ethical questions as making one better suited to other disciplines. This framing also affected enrolment, with male students significantly underrepresented in ethics courses and those who enrolled often not attending, as peer culture positioned ethics as a feminine endeavour. When P14 attempted to connect reinforcement learning concepts to the conditions of data workers, explaining how annotators were poorly paid and marginalised by rapid scaling, students responded directly that *“sir, this is not CS, let the people in social sciences look into all of this.”* When he introduced caste in the context of evaluation benchmarks, including work such as BharatBBQ that addresses these issues directly, students pushed back on the grounds that *“if we discuss the problem like this we will invite more problems and this will become a political discussion like gender.”* Resistance to caste as an evaluation category was entangled with anxiety about politicisation, positioning acknowledgement of structural harm as itself a professional risk. This risk had concrete professional consequences. P7 described a student who completed an AI ethics elective and was subsequently asked in a job interview for an e-commerce company whether studying ethics meant they might expose vulnerabilities or hold an activist mindset rather than being a real engineer. The student did not

get the job, and the account circulated among peers, with other students asking P7 whether they should omit ethics coursework from their CVs entirely. Several academic participants described responding to this pressure by rebranding courses from AI Ethics or Responsible AI to AI Safety, finding that the latter attracted more engagement while the underlying resistance to sociotechnical thinking persisted. The same dynamics shaped interdisciplinary encounters. P5 described humanities students who arrived enthusiastic about contributing to RAI but encountered sustained disrespect in joint courses with computer science peers who questioned their presence as a barrier to innovation, reflecting that *“sometimes when I hear my students it seems to me that they want Engineering and CS to be insulated from Social Sciences. I do not of course support that but I know many within my fraternity who do.”*

Absence of Civic Contestation Participants who had consulted for companies developing AI safety guardrails described evaluation tools and protocols designed without genuine engagement with the harms they were intended to detect. P3 described a red-teaming exercise in which the protocol covered explicit hate speech and child sexual abuse material but excluded caste-based discrimination, despite the system being deployed in India where such outputs were both likely and consequential. When he questioned this omission, the exclusion was described as a deliberate scoping decision rather than an oversight. P8 described working as a labeller for a major language model company where annotators received a spreadsheet with three broad categories—toxic, safe, and sexual—without cultural or contextual guidance. Bengali political slogans were flagged as unsafe while violent rhetoric in English passed through unreviewed because the reviewer overriding the flags did not read Bengali. Regional slurs directed at women that were not detected by the classifier were withheld from the client report, with the team lead explaining that *“if we focus so much on these issues we might not get another contract from them.”* Commercial considerations therefore shaped not only what was evaluated but also which findings were allowed to enter the record. P13 argued that this facade of safety persisted because affected communities lacked both awareness that these systems were making consequential decisions about them and mechanisms to contest those decisions. Most AI research produced in India, she noted, was published for technical audiences rather than engaging empirical and critical traditions that could surface these concerns for affected communities. P9 and P12 similarly characterised the absence of civic contestation as structural rather than incidental, describing the need for sustained infrastructures for dialogue between researchers, policymakers, and communities rather than episodic public education efforts. One practitioner situated this absence within a longer temporal horizon, warning that *“there is no point in doing Responsible AI if we let our personal values, especially those that are hurtful, get into AI systems. This is going to be a big challenge in the future when people find out how companies have mis-treated and misappropriated them.”*

Discussion

Addressing RQ3, our discussion proceeds in two parts. First, we theorise RAI practice in India as a doubly hegemonic system in which Western/Global North technical frameworks and local caste-gender hierarchies converge to determine legitimate ethical knowledge, operating through three interlocking mechanisms: epistemic silencing, technosolutionist channelling, and performative compliance. Second, we argue that countering this hegemony requires collective and enforceable infrastructures rather than stronger individual ethical commitment.

Hegemony, Technosolutionism, and Performative Responsible AI

We argue that RAI practice in India operates as a hegemonic system. Our findings show that practitioners did not simply lack better tools or clearer principles; rather, they worked within institutions that had already shaped which ethical concerns could be voiced, softened, or dismissed. Following Gramsci, hegemony describes how dominant ways of thinking cause specific ideologies to become accepted as common sense, often resulting in their institutionalisation through consent rather than coercion (Gramsci 1998). We identify two such dominant ideologies in our data. Globally, dominant RAI frameworks equated responsibility with audits, metrics, and safety procedures, making technical responses appear more legitimate than critiques of institutional power (Jobin, Ienca, and Vayena 2019). Locally, as participants described, caste, gender, class, and professional hierarchies shaped whose concerns were taken seriously within India's AI landscape. Together, global RAI vocabularies and local hierarchies did not merely coexist but actively legitimised each other, providing a respectable language of responsibility through which caste and gender authority could be exercised without appearing as such. Some practitioners, such as P3 and P18, however, challenged this dominance. For them, Responsible AI was not primarily about model error but a demand for accountability from policing, welfare, and bureaucratic systems that had already failed them. This corrective justice understanding of RAI represented a counter-hegemonic possibility, where RAI could discipline institutional power rather than optimise technical artefacts. The question our analysis pursues is why this possibility proved difficult to sustain within Indian AI organisations. This difficulty became apparent in how organisations received structural critiques from practitioners. Concerns about caste exclusion, gendered labelling, discriminatory welfare systems, and data misuse were routinely re-framed as personal grievance, political sensitivity, or obstruction. As our findings show, women practitioners were spoken over, caste and religious minority practitioners had to establish credibility before their arguments could be heard, and ethical concerns implicating clients or managers were treated as threats to delivery timelines and contract continuation. Drawing on Gramscian accounts of epistemic exclusion, Dotson describes this process as testimonial smothering, through which speakers are denied knower status within the very settings where ethical knowledge is most needed

(Dotson 2011). Under these conditions, corrective justice understandings of RAI were not rejected through deliberation but prevented from entering organisational reasoning as legitimate knowledge. Technosolutionism can therefore be understood as emerging from this narrowing of ethical expression, not as a freely chosen ideological position but as the only form in which critique could survive institutional reception. Participants' concerns became legible to their organisations only when recast in technical terms—arbitrary institutional power became a problem of system improvement, gendered harm became a question of cultural scope, and discriminatory outputs became entries in red-team reports rather than evidence of deeper institutional failure. Morozov's concept of technosolutionism, which identifies the conversion of political problems into optimisation problems and resonates with Gramscian transformismo (Morozov 2013), is reflected in our findings, where this conversion was less a deliberate choice than a structural consequence of environments in which critique had already been made extraordinarily difficult to voice.

The conversion of political problems into optimisation problems also worked to the advantage of hegemonic actors, since framing harm as a system defect kept scrutiny focused on technical artefacts rather than on the people and institutions deciding what systems were built, whose data was used, and which harms were ultimately counted. Performative RAI completed and sustained this arrangement by allowing organisations to affirm their commitment to responsibility without any corresponding redistribution of authority. P2 described RAI appearing in formal materials while being addressed only reactively, P9 described evaluation firms being warned against criticising sovereign AI models too strongly, and P16 described a team declining to formally flag harmful outputs in a mental health chatbot for fear of jeopardising future contracts. These accounts reflect a pattern that Metcalf and Moss describe as the corporatisation of ethics, through which responsibility language is adopted while institutional priorities remain insulated from scrutiny (Metcalf, Moss, and boyd 2019), and which Green extends by showing how organisations can sustain an appearance of accountability while preventing precisely the reforms that genuine accountability would require (Green 2022). This arrangement was further stabilised by the absence of civic contestation described by our participants, where affected communities lacked both awareness that consequential decisions were being made about them and mechanisms through which to challenge them, removing the external pressure that might otherwise have made performative compliance costly. Gramsci observed that hegemony is maintained not through outright suppression but through the incorporation of dissent into forms that leave the dominant order undisturbed (Gramsci 1998). In this sense, performative RAI functioned effectively in the organisations our participants described, where the language of responsibility circulated freely while the practices it represented remained oriented less towards ethical transformation than towards managing ethical appearance.

Countering RAI Hegemony Through Collective and Enforceable Infrastructures

Countering RAI hegemony cannot rely on asking individual practitioners to become more ethical. Our findings show that practitioners regularly recognised harm but lacked the authority, protection, or collective support to act on that recognition. Ali et al. describe this as the individualisation of risk, where organisations adopt ethical language while leaving individuals to carry the burden of critique without the structural power to produce change (Ali et al. 2023). In the Indian context, this burden was further shaped by caste, gender, religion, and professional hierarchy. Meaningful intervention therefore requires infrastructures that redistribute ethical labour, make organisational irresponsibility costly, and create space for vernacular accounts of harm to enter technical and governance practice.

A first requirement is to shift ethical responsibility from isolated individuals to collective mechanisms. Since Indian companies already face obligations to allocate resources to Corporate Social Responsibility (CSR) initiatives (Jumde and du Plessis 2020), CSR-linked AI ethics residencies could embed academics, civil society organisations, and NGO-based practitioners within industrial AI development from the outset, using the legal weight of CSR to reduce dependence on voluntary corporate goodwill (McLennan et al. 2020, 2022). Ethics bounty programmes could serve a related function by treating the identification of safety failures, discriminatory outputs, data misuse, and social harm as remunerated expertise rather than unpaid activist labour (Walsh and Simpson 2020; Dennler et al. 2023). These mechanisms would help surface risks that are currently invisible or suppressed within corporate hierarchies while lowering the personal cost of speaking up. Practitioners also need spaces for collective deliberation outside immediate workplace hierarchies. Local chapters of professional organisations focused on ethical AI, alongside mentorship programmes analogous to Athena or MATS, could allow women and caste-marginalised practitioners to engage with techno-ethical concerns without fear of trivialisation.

A second requirement is enforceable governance. Performative compliance persists when organisations can display RAI commitments without facing consequences for irresponsible deployment. Anonymous reporting mechanisms should allow practitioners to flag system-level risks directly to national AI enforcement bodies when internal hierarchies suppress dissent. Data misuse requires stronger user-centred governance: systems collecting sensitive data should provide clear opt-out controls at the point of collection, and India's Digital Personal Data Protection Act should be updated to address the specific risks posed by intelligent systems (Naithani 2024). The aim is a governance environment in which irresponsibility has organisational rather than only individual consequences.

A third requirement is to build RAI around vernacular accounts of harm rather than imported universalised risk categories. Justice-oriented approaches to AI design offer one route for translating the lived experiences of marginalised communities into system development and governance prac-

tice (Krafft et al. 2021; Santoni de Sio, Almeida, and van den Hoven 2021). Vernacular harm taxonomies could document how communities understand misclassification, exclusion, surveillance, welfare denial, and caste and gender-based discrimination in their own terms (Kennedy and Vargas Campos 2024), with civil society organisations mediating between grassroots documentation and formal governance processes. For instance, P3's account of a ration official manipulating weights against a caste-minority relative captures a form of discretionary institutional harm that no existing risk category addresses, yet precisely the kind of harm P3 understood RAI as being equipped to discipline.

These interventions are not without friction. Anonymous reporting mechanisms presuppose enforcement bodies insulated from the political protection networks our participants described. CSR-linked residencies depend on compliance infrastructure that remains uneven in practice, and ethics bounty programmes risk commodifying the very labour they aim to redistribute. As Wright (2020) notes, no structural intervention is free of implementation caveats. What matters, therefore, is that these mechanisms shift ethical responsibility away from isolated individuals towards collective and enforceable structures. Educational institutions must also challenge the positioning of RAI as secondary or activist. Rather than confining ethics to optional modules, universities should embed RAI within core technical training, with regulatory guidance from bodies such as the All India Council for Technical Education normalising this integration across engineering curricula.

Conclusion

Our findings show how global Western RAI frameworks and local hierarchies jointly shape legitimate RAI practice in India. Caste marginalised and women practitioners most clearly articulated counter hegemonic perspectives, often framing RAI as corrective justice against unaccountable institutional power. These perspectives were frequently silenced or redirected into technosolutionist responses that preserved existing frameworks and organisational hierarchies, while performative compliance allowed organisations to signal responsibility without institutional consequences. Addressing these conditions requires collective and enforceable infrastructures that redistribute ethical labour, impose consequences for organisational irresponsibility, and bring vernacular accounts of harm into technical and governance practice. RAI scholarship should therefore treat hegemony as a structural condition shaping the principle practice gap. As a qualitative study, our findings are not statistically generalisable beyond our purposive sample (Lincoln and Guba 1985), although variation in organisational contexts, professional roles, and locations supports situated understanding and analytical transferability (Flyvbjerg 2006). Representation across gender, caste, religious, and ethnic minorities remained limited. Future work should examine these dynamics across more diverse and intersectional practitioner populations.

Ethical Statement

Our institution’s Research Ethics Committee reviewed and approved this study. Although we collected organisational affiliations as part of participant screening to determine company size and sector, we excluded this information from our findings to protect the privacy of both participants and their organisations. Personal contact information (e.g., email addresses, LinkedIn profiles, and WhatsApp numbers) was collected solely for recruitment and communication purposes and was securely deleted once recruitment was complete. All interviews were audio-recorded with participants’ consent and subsequently transcribed. Transcripts were anonymised by removing identifying details, such as names, organisational affiliations, and project-specific information. Only members of the research team directly involved in data collection and analysis had access to participant contact information. Our study addresses sensitive topics, including caste, gender discrimination, and structural inequality as experienced by participants in our study. We report these perspectives to accurately represent participants’ lived experiences.

Positionality Statement

We acknowledge our positionality as researchers working at the intersection of robotics, AI, and human-computer interaction (HCI), with experience spanning early-career researchers to decades of engagement across these fields. All five authors are currently based in Western Europe. Atmadeep Ghoshal is of South Asian origin, with lived experience in South Asia, and brings regionally grounded perspectives to his research in robotics, AI, and HCI. Arshia Dutta is also of South Asian origin, with lived experience in India and research expertise in information security and Responsible AI. Martim Brandão is from Western Europe and has no lived experience of the Global South or research expertise focused on the region. He has expertise in AI, with extensive research on evaluating and implementing fairness, explainability, and other aspects of Responsible AI. Ruba Abu-Salma is from the Middle East (West Asia) and specialises in HCI and human-centred privacy research, with particular focus on at-risk populations. Sanjay Modgil is of South Asian ethnicity and has lived experience in South Asia. He has expertise in the logical foundations of AI and in the philosophy and ethics of AI.

Acknowledgements

The authors thank all participants for their time and valuable contributions. This paper was also inspired by Sunil Mitra’s Bangla work, *Antonio Gramsci: Marxbadi Oitijhorashir Sakriya Sampad*, the first Bangla-language study to introduce Gramsci’s life and intellectual contributions to a general readership and help popularise his ideas beyond academic circles and into the professional lives of workers and practitioners.

References

Adams, R. 2021. Can artificial intelligence be decolonized? *Interdisciplinary Science Reviews*, 46(1–2): 176–197.

Agarwal, D.; Naaman, M.; and Vashistha, A. 2025. AI Suggestions Homogenize Writing Toward Western Styles and Diminish Cultural Nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI ’25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713941.

Ali, S. J.; Christin, A.; Smart, A.; and Katila, R. 2023. Walking the Walk of AI Ethics: Organizational Challenges and the Individualization of Risk among Ethics Entrepreneurs. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’23, 217–226. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701924.

Anuyah, O.; Wan, R.; Adejoro, C.; Yeh, T.; Metoyer, R.; and Badillo-Urquiola, K. 2024. Cultural Considerations in AI Systems for the Global South: A Systematic Review. In *Proceedings of the 4th African Human Computer Interaction Conference*, AfriCHI ’23, 125–134. New York, NY, USA: Association for Computing Machinery. ISBN 9798400708879.

Ashktorab, Z.; Hoover, B.; Agarwal, M.; Dugan, C.; Geyer, W.; Yang, H. B.; and Yurochkin, M. 2023. Fairness Evaluation in Text Classification: Machine Learning Practitioner Perspectives of Individual and Group Fairness. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI ’23. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394215.

Bellamy, R. K. E.; Dey, K.; Hind, M.; Hoffman, S. C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilovic, A.; Nagar, S.; Ramamurthy, K. N.; Richards, J.; Saha, D.; Sattigeri, P.; Singh, M.; Varshney, K. R.; and Zhang, Y. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. arXiv:1810.01943.

Bhatt, S.; Dev, S.; Talukdar, P.; Dave, S.; and Prabhakaran, V. 2022. Re-contextualizing Fairness in NLP: The Case of India. In He, Y.; Ji, H.; Li, S.; Liu, Y.; and Chang, C.-H., eds., *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 727–740. Online only: Association for Computational Linguistics.

Biju, P. R.; and Gayathri, O. 2023. The Indian approach to Artificial Intelligence: an analysis of policy discussions, constitutional values, and regulation. *AI & SOCIETY*, 39(5): 2321–2335.

Birhane, A. 2023. *Algorithmic Colonization of Africa*, 247–260. Oxford University Press Oxford. ISBN 9780191955815.

Birhane, A. 2025. The False Promise of “AI for Social Good”. Project Syndicate.

Birhane, A.; Isaac, W.; Prabhakaran, V.; Diaz, M.; Elish, M. C.; Gabriel, I.; and Mohamed, S. 2022. Power to the People? Opportunities and Challenges for Participatory AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO ’22, 1–8. ACM.

- Birhane, A.; Steed, R.; Ojewale, V.; Vecchione, B.; and Raji, I. D. 2024. AI auditing: The Broken Bus on the Road to AI Accountability. *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 612–643.
- Braun, V.; and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2): 77–101.
- Braun, V.; and Clarke, V. 2019. To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales. *Qualitative Research in Sport, Exercise and Health*, 13(2): 201–216.
- Brown, V.; Larasati, R.; Third, A.; and Farrell, T. 2025. A Qualitative Study on Cultural Hegemony and the Impacts of AI. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '24, 226–238. AAAI Press.
- Cao, Y.; and Zhai, J. 2022. A survey of AI in finance. *Journal of Chinese Economic and Business Studies*, 20(2): 125–137.
- Chen, T.; Guo, W.; Gao, X.; and Liang, Z. 2021. AI-based self-service technology in public service delivery: User experience and influencing factors. *Government Information Quarterly*, 38(4): 101520.
- Chowdhary, S.; Ingber, A. S.; and Andalibi, N. 2025. Technical Solutions to Emotion AI's Privacy Harms: A Systematic Literature Review. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, 1119–1144. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.
- Corbett, E.; and Denton, R. 2023. Interrogating the T in FAccT. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 1624–1634. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701924.
- Deng, W. H.; Guo, B.; Devrio, A.; Shen, H.; Eslami, M.; and Holstein, K. 2023a. Understanding Practices, Challenges, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394215.
- Deng, W. H.; Yildirim, N.; Chang, M.; Eslami, M.; Holstein, K.; and Madaio, M. 2023b. Investigating Practices and Opportunities for Cross-functional Collaboration around AI Fairness in Industry Practice. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 705–716. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701924.
- Dennler, N.; Ovalle, A.; Singh, A.; Soldaini, L.; Subramonian, A.; Tu, H.; Agnew, W.; Ghosh, A.; Yee, K.; Peradejordi, I. F.; Talat, Z.; Russo, M.; and Pinhal, J. D. J. D. P. 2023. Bound by the Bounty: Collaboratively Shaping Evaluation Processes for Queer AI Harms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 375–386. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702310.
- Dotson, K. 2011. Tracking Epistemic Violence, Tracking Practices of Silencing. *Hypatia*, 26(2): 236–257.
- Elahi, T.; Srinivasa, H.; and Emruli, B. 2025. Ethical AI in Unregulated Environments: A Qualitative Empirical Study.
- Floridi, L. 2019. Translating principles into practices of digital ethics: Five risks of being unethical. *Philos. Technol.*, 32(2): 185–193.
- Flyvbjerg, B. 2006. Five misunderstandings about case-study research. *Qual. Inq.*, 12(2): 219–245.
- Friedman, B. 1996. Value-sensitive design. *interactions*, 3(6): 16–23.
- Friedman, B.; and Nissenbaum, H. 1996. Bias in computer systems. *ACM Transactions on information systems (TOIS)*, 14(3): 330–347.
- Ge, X.; Xu, C.; Misaki, D.; Markus, H. R.; and Tsai, J. L. 2024. How Culture Shapes What People Want From AI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703300.
- Ghoshal, A.; Brandão, M.; Abu-Salma, R.; and Modgil, S. 2025. Embodied AI at the Margins: Postcolonial Ethics for Intelligent Robotic Systems. In *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*.
- Gramsci, A. 1998. *Prison notebooks*. London, England: Lawrence & Wishart.
- Green, B. 2022. Technology ethics in action: Critical and interdisciplinary perspectives.
- Hao, K. 2019. In 2020, let's stop AI ethics-washing and actually do something. *MIT Technology Review*.
- Holmes, W.; and Tuomi, I. 2022. State of the art and practice in AI in education. *European Journal of Education*, 57(4): 542–570.
- Holstein, K.; Wortman Vaughan, J.; Daumé, H.; Dudik, M.; and Wallach, H. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, 1–16. New York, NY, USA: Association for Computing Machinery. ISBN 9781450359702.
- India, E. 2025. Aidea of India 2025: Policy agenda for India. Technical report, EY India.
- Ismail, A.; Thakkar, D.; Madhiwalla, N.; and Kumar, N. 2023a. Public Health Calls for/with AI: An Ethnographic Perspective. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2).
- Ismail, A.; Thakkar, D.; Madhiwalla, N.; and Kumar, N. 2023b. Public Health Calls for/with AI: An Ethnographic Perspective. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2).
- Jakesch, M.; Buçinca, Z.; Amershi, S.; and Olteanu, A. 2022. How Different Groups Prioritize Ethical Values for Responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 310–323. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.
- Jalote, P.; and Natarajan, P. 2019. The growth and evolution of India's software industry. *Commun. ACM*, 62(11): 64–69.

- Jobin, A.; Ienca, M.; and Vayena, E. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9): 389–399.
- Jumde, A.; and du Plessis, J. 2020. Legislated Corporate Social Responsibility (CSR) in India: The Law and Practicalities of its Compliance. *Statute Law Review*, 43(2): 170–197.
- Kalyanakrishnan, S.; Panicker, R. A.; Natarajan, S.; and Rao, S. 2018. Opportunities and Challenges for Artificial Intelligence in India. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, 164–170. New York, NY, USA: Association for Computing Machinery. ISBN 9781450360128.
- Kapania, S.; Siy, O.; Clapper, G.; SP, A. M.; and Sambasivan, N. 2022. "Because AI is 100% right and safe": User Attitudes and Sources of AI Authority in India. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22. New York, NY, USA: Association for Computing Machinery. ISBN 9781450391573.
- Kay, J.; Kasirzadeh, A.; and Mohamed, S. 2024. Epistemic injustice in generative AI. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 684–697.
- Kennedy, W. M.; and Vargas Campos, D. 2024. Vernacularizing Taxonomies of Harm is Essential for Operationalizing Holistic AI Safety. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7: 698–710.
- Khanuja, S.; Ruder, S.; and Talukdar, P. 2023. Evaluating the Diversity, Equity, and Inclusion of NLP Technology: A Case Study for Indian Languages. In Vlachos, A.; and Augenstein, I., eds., *Findings of the Association for Computational Linguistics: EACL 2023*, 1763–1777. Dubrovnik, Croatia: Association for Computational Linguistics.
- Krafft, P. M.; Young, M.; Katell, M.; Lee, J. E.; Narayan, S.; Epstein, M.; Dailey, D.; Herman, B.; Tam, A.; Guetler, V.; Bintz, C.; Raz, D.; Jobe, P. O.; Putz, F.; Robick, B.; and Barghouti, B. 2021. An Action-Oriented AI Policy Toolkit for Technology Audits by Community Advocates and Activists. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 772–781. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.
- Lincoln, Y. S.; and Guba, E. G. 1985. *Naturalistic Inquiry*. Thousand Oaks, CA: SAGE Publications.
- Lu, Y.; Zhang, C.; Zhang, I.; and Li, T. J.-J. 2022. Bridging the Gap Between UX Practitioners' Work Practices and AI-Enabled Design Support Tools. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22. New York, NY, USA: Association for Computing Machinery. ISBN 9781450391566.
- Madaio, M.; Egede, L.; Subramonyam, H.; Wortman Vaughan, J.; and Wallach, H. 2022. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1).
- Madaio, M. A.; Stark, L.; Wortman Vaughan, J.; and Wallach, H. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, 1–14. New York, NY, USA: Association for Computing Machinery. ISBN 9781450367080.
- Marjanovic, O.; Cecez-Kecmanovic, D.; and Vidgen, R. 2022. Theorising algorithmic justice. *European Journal of Information Systems*, 31(3): 269–287.
- Matthan, R.; and Javed, S. 2025. Global AI governance law and policy: India. <https://iapp.org/resources/article/global-ai-governance-india/>. Accessed: 2025-08-18.
- Mayeesha, T. T.; Islam, F.; and Ahmed, N. 2025. AI4Bangladesh: AI Ethics for Bangladesh - Challenges, Risks, Principles, and Suggestions. In *Proceedings of the 13th International Conference on Information & Communication Technologies and Development*, ICTD '24, 260–272. New York, NY, USA: Association for Computing Machinery. ISBN 9798400710414.
- McLennan, S.; Fiske, A.; Celi, L. A.; Müller, R.; Harder, J.; Ritt, K.; Haddadin, S.; and Buyx, A. 2020. An embedded ethics approach for AI development. *Nature Machine Intelligence*, 2(9): 488–490.
- McLennan, S.; Fiske, A.; Tigard, D.; Müller, R.; Haddadin, S.; and Buyx, A. 2022. Embedded ethics: a proposal for integrating ethics into the development of medical AI. *BMC Medical Ethics*, 23(1).
- Metcalfe, J.; Moss, E.; and boyd, d. 2019. Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics. *Social Research: An International Quarterly*, 86(2): 449–476.
- Morley, J.; Elhalal, A.; Garcia, F.; Kinsey, L.; Mökander, J.; and Floridi, L. 2021. Ethics as a Service: A Pragmatic Operationalisation of AI Ethics. *Minds and Machines*, 31(2): 239–256.
- Morozov, E. 2013. *To save everything, click here*. New York, NY: PublicAffairs.
- Muralikumar, M. D.; and McDonald, D. W. 2024. Analyzing Collaborative Challenges and Needs of UX Practitioners when Designing with AI/ML. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW2).
- Naithani, P. 2024. Analysis of India's Digital Personal Data Protection Act, 2023. *International Journal of Law and Management*, 67(5): 543–553.
- NITI Aayog. 2018. National Strategy for Artificial Intelligence. Technical report, Government of India, New Delhi. Accessed: 2025-08-18.
- Novelli, C.; Taddeo, M.; and Floridi, L. 2024. Accountability in artificial intelligence: What it is and how it works. *AI & Society*, 39(4): 1871–1882.
- Okolo, C. T. 2023. AI in the Global South: Opportunities and challenges towards more inclusive governance. <https://www.brookings.edu/articles/ai-in-the-global-south-opportunities-and-challenges-towards-more-inclusive-governance/>. Accessed: 2025-8-18.
- Okolo, C. T.; Dell, N.; and Vashistha, A. 2022. Making AI Explainable in the Global South: A Systematic Review. In *ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS)*, COMPASS '22, 439–452. ACM.

- Panch, T.; Mattie, H.; and Celi, L. A. 2019. The “inconvenient truth” about AI in healthcare. *npj Digital Medicine*, 2(1).
- Png, M.-T. 2022. At the Tensions of South and North: Critical Roles of Global South Stakeholders in AI Governance. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’22, 1434–1445. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.
- Ramjee, P.; Chhokar, M.; Sachdeva, B.; Meena, M.; Abdullah, H.; Vashistha, A.; Nagar, R.; and Jain, M. 2025. ASHABot: An LLM-Powered Chatbot to Support the Informational Needs of Community Health Workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–22. ACM.
- Rismani, S.; and Moon, A. 2023. What does it mean to be a responsible AI practitioner: An ontology of roles and skills. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’23, 584–595. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702310.
- Ruster, L. P.; and Davis, J. L. 2025. The Gaps that Never Were: Reconsidering Responsible AI’s Principle-Practice Problem. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’25, 350–360. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.
- Sadek, M.; Kallina, E.; Bohné, T.; Mougenot, C.; Calvo, R. A.; and Cave, S. 2024. Challenges of responsible AI in practice: scoping review and recommended actions. *AI Soc.*, 40(1): 199–215.
- Sambasivan, N.; Arnesen, E.; Hutchinson, B.; Doshi, T.; and Prabhakaran, V. 2021. Re-imagining Algorithmic Fairness in India and Beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’21, 315–328. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.
- Sambasivan, N.; Arnesen, E.; Hutchinson, B.; and Prabhakaran, V. 2020. Non-portability of Algorithmic Fairness in India. *ArXiv*, abs/2012.03659.
- Santoni de Sio, F.; Almeida, T.; and van den Hoven, J. 2021. The future of work: freedom, justice and capital in the age of artificial intelligence. *Critical Review of International Social and Political Philosophy*, 27(5): 659–683.
- Sarkar, A. 2023. Enough With “Human-AI Collaboration”. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI ’23, 1–8. ACM.
- Sheikh, H.; Prins, C.; and Schrijvers, E. 2023. *AI as a System Technology*, 85–134. Springer International Publishing. ISBN 9783031214486.
- Solano-Kamaiko, I. R.; Mishra, D.; Dell, N.; and Vashistha, A. 2024. Explorable Explainable AI: Improving AI Understanding for Community Health Workers in India. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI ’24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703300.
- Toy, T. 2024. Transparency in AI. *AI & SOCIETY*, 39(6): 2841–2851.
- UNESCO. 2021. Recommendation on the Ethics of Artificial Intelligence: Key Facts. Technical report, UNESCO, Paris. Accessed: 2025-08-18.
- Vakkuri, V.; Kemell, K.-K.; Kultanen, J.; Siponen, M.; and Abrahamsson, P. 2019. Ethically Aligned Design of Autonomous Systems: Industry viewpoint and an empirical study. arXiv:1906.07946.
- van Teijlingen, E.; and Hundley, V. 2002. The importance of pilot studies. *Nursing Standard*, 16(40): 33–36.
- Vasileiou, K.; Barnett, J.; Thorpe, S.; and Young, T. 2018. Characterising and justifying sample size sufficiency in interview-based studies: systematic analysis of qualitative health research over a 15-year period. *BMC Medical Research Methodology*, 18(1).
- von Braun, J.; and Baumüller, H. 2021. *AI/Robotics and the Poor*, 85–97. Springer International Publishing. ISBN 9783030541736.
- Wagner, B. 2019. *Ethics As An Escape From Regulation. From “Ethics-Washing” To Ethics-Shopping?*, 84–89. Amsterdam University Press. ISBN 9789048550180.
- Walshe, T.; and Simpson, A. 2020. An Empirical Study of Bug Bounty Programs. In *2020 IEEE 2nd International Workshop on Intelligent Bug Fixing (IBF)*, 35–44.
- Widder, D. G. 2024. Epistemic Power in AI Ethics Labor: Legitimizing Located Complaints. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’24, 1295–1304. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.
- Widder, D. G.; Zhen, D.; Dabbish, L.; and Herbsleb, J. 2023. It’s about power: What ethical concerns do software engineers have, and what do they (feel they can) do about them? In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’23, 467–479. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701924.
- Winfield, A. F. T.; and Jirotko, M. 2018. Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133): 20180085.
- Wright, E. O. 2020. *Envisioning real utopias*. The Real Utopias Project. London, England: Verso Books.