Deepfakes in the Global South: Understanding Stakeholder Perceptions, Harms, and Governance Challenges in Sri Lanka, India, and Bangladesh

Dilrukshi Gamage, Dilki Sewwandi, Shreeti Shubham, Sumaia Arefin, Dilaxshini Dunstan, Kartik Joshi, and Rashmi Gunawardana

With the rapid proliferation of AI-generated synthetic media across South Asia, communities in Sri Lanka, India, and Bangladesh face deepfake-related harms that existing Trust & Safety frameworks—designed predominantly for Global North contexts—are ill-equipped to address. Through three participatory workshops with 49 stakeholders including journalists, lawyers, fact-checkers, feminist advocates, and civil society practitioners, we investigate how deepfake harms are perceived, experienced, and contested across these distinct yet interconnected contexts. Our findings reveal four cross-cutting patterns: disproportionate gendered targeting of women and marginalized communities; systematic failure of platform moderation for South Asian languages; inaccessible institutional recourse for victims; and strong community preference for locally governed, participatory solutions over externally designed technical fixes. We contribute empirical evidence of deepfake harm in under-researched Global South contexts, a cross-country comparative framework for Trust & Safety analysis, and a set of design and policy implications that center community knowledge, survivor support, and linguistic justice as foundations for a reimagined deepfake governance paradigm.

1 Introduction

In the contemporary information ecosystem, seeing is no longer believing. Deepfakes—synthetic media in which a person's likeness, voice, or actions are fabricated or manipulated using deep learning techniques such as generative adversarial networks or diffusion

Journal of Online Trust and Safety 3 (3)
DOI: 10.54501/jots.v3i3.319
Research Article

models (Chesney and Citron 2019)—and AI-generated media more broadly, including voice cloning, text-to-image generation, and AI-generated video, have transformed the foundations of trust, authenticity, and accountability in digital communication. What began as a technical experiment in face-swapping has evolved into a global socio-political crisis that challenges the credibility of news, the integrity of elections, and the safety of individuals, especially women and marginalized communities (Kafayat 2025). Deepfakes unsettle one of the most basic human assumptions: that visual evidence corresponds to reality (Zhang 2024).

While scholarly, policy, and technology discourse around deepfakes has intensified in Western contexts, the Global South, home to the majority of the world's internet users, faces an equally urgent yet profoundly under-examined set of challenges (Paul and Dutta 2025). In South Asian countries such as Sri Lanka, India, and Bangladesh, deepfakes intersect with pre-existing structural vulnerabilities: fragile media ecosystems, low digital literacy, deep linguistic diversity, entrenched patriarchal norms, and volatile political climates. These conditions amplify deepfake harm, accelerate its spread, and shape who bears its consequences, yet the frameworks developed to understand, govern, and mitigate synthetic media harm are overwhelmingly designed in and for the Global North, rendering South Asian realities largely invisible in both research and policy.

In Sri Lanka, AI-generated content has been used to impersonate politicians (Atherton 2025), cricketers (Newswire 2025), and senior financial officials including the Governor of the Central Bank (Popowicz 2025)—deployed not for satire but for financial fraud and reputational destruction. In India, deepfakes have penetrated electoral discourse at unprecedented scale: candidates' speeches have been voice-cloned to appeal to specific linguistic communities, fabricated videos of political leaders have been weaponized for delegitimation, and AI-generated content targeting women in public life proliferates across platforms (Sunvy, Reza, and Al Imran 2023). In Bangladesh, a deepfake video depicting ousted Prime Minister Sheikh Hasina confessing to the creation of clandestine detention centers—known as *aynaghor*, or "house of mirrors"—was widely circulated, verified as AI-manipulated by fact-checkers, and credited with generating significant political turmoil (Chowdhury 2025). Across all three countries, synthetic pornography has been systematically weaponized against women, particularly politicians, journalists, and activists, revealing that deepfake harm is not only informational but emotional, cultural, and deeply gendered (Brieger 2024).

These harms unfold in a governance vacuum. The Trust & Safety (T&S) practices that structure major platforms—content moderation, automated labeling, policy enforcement—are rooted in Western liberal assumptions of stable democratic institutions, strong legal systems, and relatively homogeneous cultural expectations around speech, privacy, and individual harm (Karthikeyan and Roy 2026). In contexts where democratic institutions are fragile, where collective ethnic, linguistic, and religious identities frequently outweigh individual expression, and where a single manipulated video can precipitate commu-

nal violence with immediate offline consequences, these frameworks are structurally misaligned with the social contracts they are meant to protect. Compounding this, moderation systems trained predominantly on English-language data systematically misclassify or fail to process harmful content in Sinhala, Bengali, Tamil, and Hindi, leaving most South Asian internet users structurally under-protected (Shahid, Elswah, and Vashistha 2025).

Sri Lanka, India, and Bangladesh together offer a revealing lens on this governance gap: each has a rapidly expanding digital public sphere shaped by affordable smartphones, deep social media penetration, and a youthful population, alongside sharp asymmetries of power between technology companies and governments, urban and rural communities, and men and women. To understand how deepfakes are affecting these communities, we conducted three participatory workshops in Sri Lanka (April 2025), India (August 2025), and Bangladesh (September 2025), bringing together 49 stakeholders from journalism, law, fact-checking, civil society, feminist advocacy, and academia. Rather than treating these communities as populations to be studied, our participatory design positioned them as co-producers of knowledge about their own digital safety, centering local expertise as a primary analytical resource.

The communities we worked with possess substantial and largely untapped knowledge about how synthetic media harm propagates in their own contexts—knowledge with direct implications for global approaches to AI safety. By foregrounding these voices, this paper reframes the deepfake problem from a question of technology to be managed to a question of society to be understood, and calls for a re-imagined T&S paradigm built on that knowledge. The paper makes three contributions. First, we provide empirical evidence of deepfake harm in three under-researched Global South contexts, grounded in the lived experience and professional expertise of 49 local stakeholders. Second, we offer a cross-country comparative framework for T&S analysis that surfaces both cross-cutting patterns and country-specific distinctions. Third, we derive design and policy implications that center community knowledge, survivor support, and linguistic justice as foundations for any credible response to synthetic media harm in the Global South.

2 Literature Review

2.1 Deepfakes and the Erosion of Democratic Trust

Deepfakes threaten democratic institutions through several reinforcing mechanisms. Experimental evidence shows that even modest synthetic content such as fabricated footage of infrastructure failure produces measurable increases in distrust of government (Ahmed et al. 2025). Highly plausible deepfakes that subtly distort rather than wildly fabricate exert stronger delegitimizing effects on officeholders than crude fabrications,

precisely because their realism makes refutation more difficult (Hameleers, Van Der Meer, and Dobber 2024). Pawelec (2022) identifies the deeper democratic injury: by undermining citizens' capacity for evidence-based deliberation, deepfakes threaten the core functions of democratic life—empowered inclusion, agenda-formation, and collective decision making. These direct harms are compounded by the "liar's dividend," whereby political actors exploit public awareness of deepfakes to dismiss authentic recordings as synthetic, further eroding trust in journalism (Schiff, Schiff, and Bueno 2024). Synthetic media is thus not merely a vector for disinformation but an active force restructuring the epistemic foundations on which democratic legitimacy depends.

2.2 The Limits of Global North Trust & Safety Frameworks

Yet the T&S frameworks developed to govern synthetic media are structurally ill-equipped to address this crisis outside the contexts in which they were designed. Platform moderation systems are built overwhelmingly for English and a few high-resource languages, failing systematically when deployed for Tamil, Bengali, Sinhala, Hindi, Swahili, and other languages spoken by the majority of global internet users (Shahid, Elswah, and Vashistha 2025). This reflects design choices rather than mere resource constraints: platforms headquartered in the Global North optimize for Northern linguistic, legal, and cultural contexts, and disproportionately direct their misinformation and moderation budgets toward the U.S. market even though the Global South constitutes the majority of global internet traffic (De Gregorio and Stremlau 2023). The inequity extends into the labor of moderation itself: content labeling is outsourced to workers in Kenya, the Philippines, and elsewhere in the Global South, while modeling, policy design, and governance authority remain concentrated in Western tech centers—what Das et al. (2024) describe as a colonial division of labor. Content policies likewise embed Western liberal assumptions about speech, harm, and privacy that translate poorly into other legal and cultural regimes, suppressing legitimate local expression while failing to detect genuine harm (Balendra 2025; Shahid, Elswah, and Vashistha 2025). Users in the Global South consequently experience systematic mis-moderation, under-moderation, and policy mismatch. Addressing this requires decentralizing governance authority, incorporating local expertise into policy and model design, and treating moderation as a culturally situated practice rather than a technically universal one.

2.3 Deepfakes in the South Asian Context

South Asia combines some of the world's fastest-growing digital populations with fragile democratic institutions, deep linguistic heterogeneity, entrenched gender inequality, and platform governance systems not designed for its realities, yet empirical deepfake research grounded in the region remains sparse. The following subsections review what is known about deepfake perception, harm, legal response, and media literacy in India, Bangladesh, and Sri Lanka, moving from the most to the least studied context, before identifying the cross-cutting gaps this study addresses.

2.3.1 India: Aspirational AI, electoral manipulation, and regulatory gaps

India's relationship with synthetic media is shaped by a paradox: widespread concern about deepfake misuse coexists with unusually high public trust in AI. A 2025 KPMG global survey found that approximately 76% of Indian respondents expressed confidence in using AI, well above the global average (KPMG 2025). Kapania et al. (2022) document a pronounced deference to AI authority among Indian users, who describe AI systems as inherently right and safe and extend them trust even in the absence of demonstrated reliability, a disposition with direct implications for deepfake susceptibility. Consistent with this disposition, a 142-country comparative study found South Asia among the regions least concerned about online misinformation, with only 32.2% of respondents expressing high concern (Newman et al. 2024).

This low concern sits in tension with documented harm at scale. The 2024 Lok Sabha election marked an unprecedented inflection point: AI-generated videos and voice clones impersonating political leaders and Bollywood celebrities—including fabricated endorsements attributed to Aamir Khan and Ranveer Singh—circulated widely across regional-language platforms (Kalra, Vengattil, and Pandya 2024; Das 2024), prompting formal warnings from the Election Commission of India to political parties (Joy 2024). In the months surrounding the vote, over 50 million AI-generated voice-clone calls were reportedly made to voters, part of a rapidly professionalizing synthetic-media campaign industry (Christopher and Bansal 2024). Political actors have thus not merely encountered deepfakes as a threat but actively adopted generative AI as an electoral instrument. Susceptibility is compounded by this deference to institutional and technological authority (Kapania et al. 2022), while the stronger delegitimizing power of subtle deepfakes (Hameleers, Van Der Meer, and Dobber 2024) and the liar's dividend (Schiff, Schiff, and Bueno 2024) carry particular salience in India's complex multi-party electoral environment.

India has no deepfake-specific legislation. Existing responses rely on the Information Technology Act, 2000 Section 66E (unauthorized publication of private images), Section 66D (online impersonation), and Sections 67–67B (circulation of obscene and sexually explicit material) (GOI 2000) supplemented by the IT Rules, 2021, which require platforms to remove deepfake or morphed content within 24 hours of a user complaint (MEITY 2021), and the 2022 Digital Personal Data Protection Bill's consent framework, grounded in the right to privacy upheld in *Justice K.S. Puttaswamy v. Union of India* (SCI 2017). Scholars nonetheless agree these frameworks are substantially inadequate: recent legal scholarship identifies compounding deficiencies of definitional vagueness, slow and resource-intensive enforcement, and the absence of victim-centered remedies for those targeted by synthetic sexual imagery (Kashyap 2025). Cybercrime against women in India has long left victims navigating multiple overlapping and often inaccessible statutes (Halder and Jaishankar 2016), a burden that AI-driven abuse now compounds. The scholarly consensus calls for a multi-pronged strategy integrating legislative reform, technological innovation, and public education, with digital literacy understood as a

rights-protective capacity rather than a consumer skill (Kashyap 2025).

2.3.2 Bangladesh: Vulnerable populations, literacy gaps, and governance failures

Research on deepfakes specific to Bangladesh remains scarce, but converging evidence identifies low-digital-literacy populations as the most acutely vulnerable. Experimental work shows that people are generally unable to detect deepfakes reliably while overestimating their own ability to do so (Köbis, Doležalová, and Soraperra 2021), and that exposure to political deepfakes increases uncertainty and erodes trust in news (Vaccari and Chadwick 2020). Older users combine lower technology engagement with high trust in video as an evidential medium and are disproportionately likely to believe and share false content (Brashier and Schacter 2020; Guess, Nagler, and Tucker 2019), a vulnerability visible in the health misinformation circulating on Bangladeshi social media (Al-Zaman 2021). More encouragingly, even brief media literacy interventions measurably improve users' ability to discern false and manipulated content, including in South Asian settings (Guess et al. 2020), and scholars agree that building such literacy capacity is a more durable defense than detection software, which the evolution of generative AI consistently outpaces (Mirsky and Lee 2022). Yet the populations most in need—rural women, elderly users, those with limited formal education—are precisely those least reached by existing programs, which tend to be urban, English-language, and device-dependent. Language compounds every other barrier: most detection and literacy resources are unavailable in Bengali, and research engaging Bengali-speaking users or Bangladeshi digital culture remains rare (Das et al. 2024), leaving the communities most at risk unable to access even nominally existing protections.

Bangladesh's regulatory landscape is underdeveloped even relative to the already-inadequate frameworks in India and Sri Lanka. The country has no deepfake-specific legislation, and the Digital Security Act, 2018 has been criticized not for failing to protect deepfake victims but for being actively weaponized against journalists and activists—the very populations most targeted by synthetic abuse (Article 19 2020; HRW 2020). Victims thus face a legal environment in which the primary digital safety statute may itself threaten their safety, prompting calls for reform that distinguishes content regulation as accountability from content regulation as repression (HRW 2020). On the technical side, automated detection and moderation systems built for high-resource languages perform substantially worse on Bengali content (Shahid, Elswah, and Vashistha 2025; Das et al. 2024); ordinary users cannot be assumed to detect synthetic content unaided, since people systematically overestimate their own detection ability (Köbis, Doležalová, and Soraperra 2021); and vulnerable groups receive almost no attention in existing technical research despite being disproportionately affected. Scholars consistently call for multi-stakeholder collaboration among technologists, policymakers, educators, journalists, and community leaders, but note that no coordinating institutional infrastructure for AI governance yet exists in the country.

2.3.3 Sri Lanka: Ethnic polarization, documented harm, and an absent evidence base

Sri Lanka has received the least scholarly attention of the three countries, despite presenting a harm configuration that is in some respects the most acute. Its post-conflict history of ethnic polarization between Sinhalese, Tamil, and Muslim communities creates conditions in which manipulated media targeting one community or impersonating a leader from another can escalate to offline communal harm with unusual speed, as documented in analyses of Facebook-borne hate speech in Sri Lanka and of the platform's role in the 2018 anti-Muslim violence (Samaratunge and Hattotuwa 2014; Taub and Fisher 2018)—a "combustibility" theorized in the broader deepfake literature (Pawelec 2022) but never empirically studied in the Sri Lankan context. Documented incidents span financial fraud, political manipulation, and reputational attack: AI-generated videos impersonating the Governor of the Central Bank of Sri Lanka promoted fraudulent investment schemes in early 2025 (Popowicz 2025); deepfakes of cricket legend Kumar Sangakkara falsely depicted him endorsing commercial products (Newswire 2025); and a synthetic video of President Anura Kumara Dissanayake endorsing a fraudulent government investment scheme confirmed as AI-generated by Sri Lanka CERT and the Presidential Media Division became one of the region's most high-profile political deepfake cases (Atherton 2025). The pattern is consistent: high-trust institutional and cultural figures are synthetically impersonated to exploit public confidence.

Sri Lanka's legal framework is similarly underdeveloped: there is no deepfake-specific legislation, and existing cybercrime provisions do not adequately cover AI-generated content used for financial fraud or reputational attack. Institutional capacity is concentrated in Sri Lanka CERT, which has played a reactive verification role in high-profile cases but lacks the mandate or resources for proactive governance (Atherton 2025). Digital literacy programs addressing synthetic media are nascent and largely confined to urban, English-medium contexts, leaving most Sinhala- and Tamil-speaking communities outside Colombo without access to protective information or tools. Scholarly research on deepfakes in Sri Lanka is almost entirely absent; what exists consists of journalistic documentation of individual incidents (Popowicz 2025; Newswire 2025; Atherton 2025) and civil society reports linking synthetic media to unresolved ethnic tensions and weak digital governance infrastructure (Balendra 2025). This absence is itself a form of governance failure: without research that systematically maps the harm landscape, the evidence base for policy reform and community intervention remains thin, a gap this study addresses directly.

2.4 Cross-Cutting Gaps and the Case for Participatory Research

Across the three contexts, this literature reveals four cross-cutting gaps that motivate the present study. First, existing research is predominantly quantitative and survey-based, measuring individual awareness and attitudes without capturing the structural, institutional, and community-level dynamics through which deepfake harm is encoun-

tered, interpreted, and navigated in practice. Second, the populations most acutely affected—rural women, queer individuals, linguistic minorities, low-literacy users—are systematically underrepresented in existing research. Third, governance research is fragmented across legal, technical, and literacy domains, with few studies examining how these domains interact and where their combined failures leave communities exposed. Fourth, and most fundamentally, the communities most affected by deepfake harm in South Asia have not been positioned as agents of knowledge production in the research that purports to document their experience (Shahid and Vashistha 2023; Birhane et al. 2022). This study addresses these gaps through a multi-country participatory design that centers stakeholder expertise, foregrounds vulnerable community perspectives, examines governance across legal, technical, and literacy dimensions simultaneously, and treats affected communities as co-producers of the knowledge needed to respond to deepfake harm.

3 Methods

To investigate how deepfakes are perceived, experienced, and addressed within South Asian digital ecosystems, we employed a multi-country participatory qualitative research design grounded in Human–Computer Interaction (HCI) and critical data studies. Because deepfake harms in Sri Lanka, India, and Bangladesh are shaped by culturally embedded vulnerabilities, linguistic diversity, and socio-political sensitivities, a top-down or solely technical approach would have been insufficient; our participatory workshop model instead foregrounds local knowledge, lived experience, and collective sense-making as primary sources of evidence. Two questions guided the research:

- **RQ1:** How do stakeholders in Sri Lanka, India, and Bangladesh perceive and experience the social, political, and personal harms of deepfakes?
- **RQ2:** What locally grounded governance, literacy, and design interventions do stakeholders propose to enhance Trust & Safety in their digital ecosystems?

3.1 Research Design and Methodological Rationale

Our multi-sited, participatory qualitative design draws on participatory action research and co-design traditions within HCI (Hayes 2011; Sanders and Stappers 2008), for three reasons. First, deepfake harm ecosystems in South Asia are deeply context-dependent: political configurations, legal frameworks, linguistic landscapes, and cultural norms vary significantly across the three countries, making single-site or survey-based approaches ill-suited to capture local nuance. Second, Global South communities have historically been positioned as objects of study rather than agents of knowledge production in technology research (Irani et al. 2010; Dourish and Mainwaring 2012), making a participatory framework both methodologically appropriate and ethically necessary.

Table 1: Overview of participatory workshops.

Country	Participants	Stakeholder Groups Represented	Date
Sri Lanka	15	Journalists, NGOs, educators, technologists	April 2025
India	22	Fact-checkers, lawyers, academics, activists	August 2025
Bangladesh	12	Lawyers, media activists, students, feminists	September 2025

Third, the interventions most likely to succeed are those developed with, rather than for, local stakeholders. Each workshop combined guided discussion, experiential learning, scenario-based provocation, and collaborative ideation, following the same methodological scaffold for cross-country comparability while adapting language, cultural references, and scenario content to each national context.

3.2 Participant Recruitment and Sampling

Participants were recruited through purposive and snowball sampling, prioritizing diversity of professional background, gender, and institutional affiliation. In each country we targeted stakeholders who navigate digital information environments professionally and bear direct or indirect exposure to deepfake harms: journalists and fact-checkers, lawyers and policy advocates, educators, civil society practitioners and activists, and academics studying media, law, or technology. This multi-stakeholder approach was intentional: understanding deepfake governance requires perspectives from those who detect harm, those who litigate it, those who are victimized by it, and those who educate against it. Initial participants were identified through the University of Colombo's established HCI Lab networks across South Asia, with subsequent recruitment via professional referrals within each country's journalism, legal, and civil society ecosystems. Participation was voluntary and unincentivized; all participants provided informed consent and later received a debrief and summary of cross-country findings. A total of 49 participants took part: 15 in Sri Lanka, 22 in India, and 12 in Bangladesh, with 60% female and 40% male representation; demographics are reported at a general level to protect anonymity. Variation in sample size reflects the differing depth of established research networks and the logistics of the online format; despite the smaller Bangladesh sample, participants there included feminist scholars and student activists whose voices are frequently absent from formal policy discussions on deepfake governance. Table 1 summarizes the participant distribution.

3.3 Workshop Design and Procedures

Each workshop was conducted online via video conferencing, enabling participation from geographically dispersed locations and reducing the power asymmetries that can arise in in-person academic venues. Sessions ran approximately three hours and were facilitated by two to three researchers with expertise in HCI, media studies, or digital governance, at least one fluent in the dominant local language (Sinhala and

Tamil for Sri Lanka; Hindi and English for India; Bengali and English for Bangladesh). All workshops followed the research team's participatory research protocol, structured around five sequential activities held constant across sites for analytical comparability and summarized in Table 2. Activities 1 and 2 addressed RQ1, generating evidence of how stakeholders perceive and experience deepfake harms; Activity 3 extended RQ1 by surfacing stakeholders' reasoning about institutional responsibilities and governance failures under realistic crisis conditions; Activities 4 and 5 addressed RQ2, shifting from problem diagnosis to solution generation and collective prioritization. This sequence was deliberate: grounding participants in concrete experiences and scenarios before co-design ensured that proposed interventions were anchored in lived and professional realities rather than abstract preferences.

3.3.1 Activity 1: Spot the Fake

Each workshop opened with a structured media-evaluation exercise in which participants assessed a curated set of real and AI-generated images, audio clips, and short videos, localized to each country's media landscape (Sri Lankan political figures, Indian electoral content, Bangladeshi social media posts), judging authenticity and articulating their reasoning aloud or in a shared digital workspace. The activity established baseline awareness and detection skill, surfaced the intuitive heuristics participants use when evaluating digital media—which later formed an inductive coding category—and grounded subsequent discussion in concrete, shared experience rather than abstract theorizing.

3.3.2 Activity 2: Experience Sharing

The second activity was an open facilitated forum in which participants shared personal or professionally observed incidents involving deepfakes or AI-generated misinformation. Facilitators used broad opening prompts (e.g., "Have you or someone you know encountered a deepfake in your professional or personal life? What happened?") and probed for narrative depth, with participants encouraged to speak in their preferred language and real-time interpretation provided where needed. This activity generated some of the richest data in the study: journalists described receiving synthetic audio clips purporting to be from politicians and grappling with whether to report on or debunk them; lawyers recounted the legal dead-ends faced by women whose faces were used in non-consensual synthetic pornography; educators described students circulating deepfake content as entertainment, unaware of its political origins; and fact-checkers discussed synthetic content spreading through encrypted messaging platforms faster than verification was possible. These accounts, recorded with informed consent and transcribed, were the primary data source for RQ1.

3.3.3 Activity 3: Scenario Probing

In the third activity, small groups of three to five participants deliberated on hypothetical but contextually realistic crisis scenarios tailored to each country's political and social environment, each designed to surface tensions between competing values—free expression versus harm prevention, platform accountability versus state censorship, individual privacy versus collective safety. Scenarios included:

- Sri Lanka: A deepfake video showing a prominent politician confessing to election fraud circulates on WhatsApp days before a national election. What should journalists, platforms, and election authorities do?
- India: A synthetic voice clone of a Bollywood celebrity is used in a robocall campaign to endorse a political candidate without the celebrity's consent. Who bears responsibility, and what legal recourse is available?
- Bangladesh: A non-consensual deepfake pornographic video targeting a female journalist circulates on Telegram and Facebook. She contacts police, a lawyer, and a civil society organization. What happens next?

Groups discussed their scenarios for approximately 20 minutes before reporting back to the full workshop. Facilitators pressed participants to identify not only ideal responses but also the legal, institutional, technical, and social barriers that would prevent such responses in practice, generating data on perceived trust and safety gaps and existing coping strategies that experience sharing alone did not always surface.

3.3.4 Activity 4: Co-Design Session

The co-design activity moved from problem identification to solution generation. Working in mixed-stakeholder groups on a shared digital whiteboard (Miro), participants collaboratively designed interventions addressing the harms and gaps identified in the preceding activities, loosely structured around three domains—legal and regulatory frameworks, media and digital literacy, and technological tools—while being encouraged to combine domains and to ground proposals in their country's specific legal, cultural, and linguistic conditions. Facilitators documented all proposals in real time, ensuring that all voices were captured and no single stakeholder group dominated. This activity was the primary data source for RQ2.

3.3.5 Activity 5: Collective Prioritization

Finally, participants collectively prioritized the co-designed interventions through a structured dot-voting exercise guided by three criteria: urgency, feasibility, and local relevance. A plenary discussion then examined the highest-ranked interventions in depth, with participants articulating the reasoning behind their choices. This prioritization triangulated and validated the co-design outputs, distinguishing proposals that generated

Table 2: Workshop structure and purpose.

Activity	Description	Purpose
Spot the Fake	Participants evaluated real vs. AI-generated media (images, audio, and video).	Assess the understanding and intuitive detection skills; surface heuristics.
Experience Sharing	Open forum to recount personal or observed incidents of deepfake harms.	Identify context-specific harms and lived experiences.
Scenario Probing	Groups responded to hypothetical crises (e.g., election deepfake, harassment case).	Understand reasoning, risk perception, and decision paths.
Co-Design Session	Participants generated intervention ideas across law, literacy, and technology.	Elicit community-driven solutions.
Collective Prioritization	Ideas were ranked based on feasibility, urgency, and local relevance.	Identify actionable, high-impact interventions.

initial enthusiasm from those participants judged genuinely actionable within their socio-institutional contexts; the resulting data identified the most policy-relevant, community-endorsed interventions for each country, as reported in the Findings.

3.4 Data Collection

Each workshop generated multiple overlapping data streams. Where participants spoke in Sinhala, Tamil, Hindi, or Bengali, native-speaker members of the research team produced direct translations alongside the original-language transcripts, preserving linguistic specificity and cultural resonance; transcripts were reviewed against recordings for accuracy. Facilitators maintained detailed contemporaneous field notes capturing small-group dynamics, moments of consensus or disagreement, and power dynamics not fully legible in the audio record. Digital whiteboard exports from Miro preserved the visual outputs and voting distributions of the co-design and prioritization activities, and chat logs captured the links, comments, and examples participants typed in parallel with discussion. Together, these multi-modal sources formed a rich corpus for analysis.

3.5 Analytical Approach

The corpus—session transcripts, chat logs, whiteboard exports, facilitator field notes, and prioritization voting data—was analyzed using reflexive thematic analysis (RTA) (Braun et al. 2022), selected for its epistemological alignment with our participatory, interpretivist design: RTA treats themes as actively constructed through the researcher's engagement with the data, shaped by theoretical commitments and reflexive awareness of positionality, rather than as latent structures awaiting discovery. Three researchers, each with

primary expertise in a different national context, independently familiarized themselves with and open-coded the full corpus, grounding codes in participants' own language wherever possible; codes were organized into candidate themes through iterative team meetings attending to convergence and productive disagreement. Coding combined inductive strategies, capturing emergent and locally specific insights, with deductive codes guided by existing literature on deepfake harms, T&S frameworks, and digital governance in the Global South. Analysis was organized around four dimensions corresponding to the research questions and workshop activities: perceived threats and harms (political manipulation, gendered violence, communal tensions, identity spoofing, reputational damage); trust and safety gaps (platform inaction, linguistic invisibility, limited detection tools, weak legal infrastructures); existing coping mechanisms (informal verification, peer-to-peer literacy, trusted community intermediaries); and proposed interventions (legal reforms, media literacy programs, provenance technologies, community support systems). Iterative theme-refinement sessions sharpened theme boundaries and resolved interpretive disagreements, supported by analytical memos maintained by each researcher for transparency and reflexivity. Themes were then compared across Sri Lanka, India, and Bangladesh to identify cross-cutting patterns and country-specific distinctions.

3.6 Ethical Consideration

The study received institutional ethical approval from the University of Colombo ethics review committee. All participants provided informed consent via Google Forms before the workshops, having received a plain-language information sheet describing the study's aims, data collection methods, intended outputs, and their rights, including withdrawal at any point without consequence; all agreed to recording, transcription, and the use of anonymized excerpts in publications. Given the sensitivity of the subject matter, facilitators were trained to recognize and respond to participant distress; ground rules for respectful, confidential sharing were established at the outset, and participants could decline to answer or speak off the record at any time. No participant withdrew or removed contributions from the corpus. To protect confidentiality, all direct quotations are attributed only by country, professional role, and gender (e.g., 'female journalist, Sri Lanka'), and—given the political sensitivity of deepfake discourse in all three countries—every quotation was reviewed to ensure it could not be traced back to an individual through contextual inference.

3.6.1 Positionality and reflexivity

The design, facilitation, and analysis of participatory research are never neutral acts. Our team includes scholars based in Sri Lanka with collaborative ties to institutions in India, Bangladesh, and the Global North—simultaneously partial insiders, with linguistic competence, professional networks, and cultural familiarity, and academics whose practices are shaped by Global North disciplinary norms and publication incentives. We

navigated this tension by centering participants' own framings and vocabularies throughout analysis, having country-specific findings reviewed for accuracy and resonance by co-investigators from each context, and resisting the impulse to flatten contextual variation into universally applicable frameworks. We also acknowledge that our stakeholder groups, while deliberately diverse, are not a representative sample of the populations affected by deepfakes in South Asia: professionals with digital access are likely to have more developed vocabularies for describing deepfake harm than the general public, particularly rural and low-literacy communities. The findings reflect the perspectives of stakeholders in governance and advocacy roles rather than the experiences of all affected individuals, and should be interpreted accordingly.

4 Results

Findings are organized by research question and country. RQ1 (perceived and experienced harms) draws on open-discussion transcripts, facilitator notes, and activity outputs, analyzed through reflexive thematic analysis; RQ2 (proposed governance, literacy, and design interventions) draws on co-design outputs and collective prioritization. Participant-proposed interventions are consolidated in Table 3, and a cross-country comparison of harm themes is presented in Table 4.

4.1 RQ1: Perceived Harms and Experiences

Sri Lanka: Participants articulated deepfake harm primarily through the lens of democratic fragility and ethnic polarization. Journalists and technologists described AI-generated videos of politicians—including simulated confessions, fabricated policy announcements, and synthetic entertainment content—circulating widely on WhatsApp and Facebook in Sinhala and Tamil, languages poorly served by automated moderation systems. A recurring concern was the weaponization of deepfakes in pre-election periods: synthetic content targeting female candidates had both a psychological silencing effect and the practical consequence of suppressing voter trust. The "Spot the Fake" activity revealed that even digitally experienced participants struggled to identify AI-generated political imagery when it was contextualized within familiar media narratives. Participants stressed that Sri Lanka's post-conflict ethnic tensions create a uniquely volatile environment in which manipulated content portraying members of one ethnic group as making provocative statements can escalate to offline communal harm with unusual speed, and noted that low-cost smartphone AI applications have democratized synthetic content production, making reactive moderation inadequate as a primary defense.

Structural trust and safety gaps centered on platform inaction: content takedown requests were reported as routinely ignored or rejected on grounds that did not account for Sri Lankan cultural and linguistic contexts. A technologist demonstrated during the workshop how clearly manipulated political content in Sinhala had remained online for

weeks despite multiple reports, while superficially similar English-language content was removed within hours—experienced not as a technical limitation but as a manifestation of structural power imbalances in Global North–centric platform governance. Participants also noted the absence of deepfake-specific legislation in Sri Lanka and the insufficient localization of detection tools for Sinhala-language content.

Existing coping strategies relied on established journalist networks and civil society coalitions including organizations such as Sri Lanka CERT and Watchdog for informal verification and escalation of egregious cases to platform trust and safety teams via personal contacts, alongside individual heuristics of cross-referencing sources, checking metadata, and consulting known fact-checkers as primary tools of epistemic defense.

India: The largest workshop, with 22 participants from fact-checking organizations, legal practice, academia, and media literacy NGOs, produced findings notable for their breadth across harm domains. Participants shared detailed experience of the 2024 Lok Sabha election cycle, during which AI-generated robocalls, voice clones of political leaders, and fabricated celebrity endorsements circulated at unprecedented scale across regional-language ecosystems; a fact-checker from a national verification organization described the practical impossibility of responding to synthetic content in Hindi, Marathi, and Tamil simultaneously with detection tools trained primarily on English-language data. Health misinformation emerged as particularly urgent: participants described a proliferation of AI-generated videos featuring synthetic doctors dispensing dangerous medical advice to elderly and low-literacy populations through WhatsApp forward chains, including one case in which a deepfake video of a medical professional recommending an unproven treatment led a patient to abandon prescribed medication. Participants also documented the "liar's dividend": political actors instrumentalizing deepfake awareness to dismiss legitimate recordings as synthetic in adversarial contexts.

Trust and safety gaps included platform inaction on verified reports of synthetic election content—one fact-checker received automated rejections citing insufficient evidence despite submitting forensic documentation; IT Act provisions described as slow and difficult to invoke in practice; and detection tools that were either not localized for South Asian languages and visual contexts or not publicly accessible at the needed scale. Distrust in AI tools and data privacy concerns were also prominently raised, a theme largely absent in the other two workshops.

Coping strategies included collaborative verification networks in which multiple fact-checking organizations shared intelligence about suspicious content in real time, with the Deepfakes Analysis Unit (DAU) cited as a notable institutional effort combining technical detection with policy advocacy; several participants were already collaborating with it. Participants acknowledged these practices, along with individual cross-referencing and consultation of known fact-checkers, as fragile and inaccessible to rural and low-literacy populations.

Bangladesh: The workshop generated the most concentrated and emotionally charged discussion of the three sessions, centering on women as disproportionate targets of deepfake harm. Participants including a lawyer, feminist researchers, and civil society activists described non-consensual synthetic pornography, manipulated intimate images used for extortion, and fake social media accounts as routine vectors of harassment against women in public life, including journalists, activists, and students. A lawyer recounted multiple cases in which clients presented evidence of deepfake abuse only to encounter a complete absence of applicable legal provisions, as Bangladesh at the time had no legislation addressing AI-generated content or image-based sexual abuse. Victims from marginalized communities such as queer individuals, indigenous women, and rural users faced compounded vulnerabilities: more likely to be targeted, with fewer institutional channels for redress. The psychological dimension was particularly prominent: participants described clients experiencing severe anxiety, social withdrawal, and in extreme cases suicidal ideation following victimization. Health misinformation also surfaced, with AI-generated health videos spreading through Bengali WhatsApp chains among older users who place high trust in video content featuring apparent medical authority.

Trust and safety gaps included the complete absence of platform contacts to receive takedown complaints—described by a lawyer participant as making even initiating a removal process impossible—and a digital safety framework offering victims little practical recourse; distrust in AI tools and data privacy concerns were strongly foregrounded as a primary theme, and victims from marginalized communities faced structural invisibility in both legal and platform-facing redress channels. Coping strategies centered on feminist civil society organizations that had developed peer support networks for victims of image-based abuse, providing legal referrals, emotional support, and practical guidance through encrypted channels participants regarded as safer than formal reporting mechanisms.

4.2 RQ2: Proposed Interventions

Across all three workshops, legal reform ranked as the most urgent intervention domain and media and digital literacy as the second; Table 3 consolidates the full set of proposals by domain and country, and we summarize the country-specific emphases here.

Sri Lankan participants prioritized deepfake-specific electoral legislation with penalties for political misuse and victim protection provisions; gamified detection tools for students and community awareness programs attentive to ethnic and linguistic diversity; locally built verification technology, including LLM-based multimodal detection, browser extensions, and a volunteer-sourced deepfake database; victim-support platforms connecting NGOs with affected individuals and mental health referral pathways; and faster, more transparent takedown processes with community-based flagging.

Indian participants focused on strengthening existing frameworks: clearer definitional

Table 3: Participant-proposed interventions by domain and country.

Intervention Domain	Sri Lanka	India	Bangladesh
Legal & Regulatory Reform	Deepfake-specific legislation; penalties for political misuse; victim protection provisions in electoral law	Clearer definitional frameworks within IT Act; faster enforcement of IT Rules 2021; stricter platform takedown obligations	Comprehensive new cyber law modeled on EU AI Act; explicit provisions for gender-based synthetic abuse and image-based violence
Media & Digital Literacy	Gamified detection tools for students; community awareness programs accounting for ethnic and linguistic diversity	School curriculum integration as sustained component; multilingual public campaigns; training for healthcare professionals to identify synthetic medical content	Literacy programs targeting rural and low-literate groups; feminist advocacy networks; approaches adapted for students without device access
Technology & Detection Tools	Automated verification platform for Sri Lankan context; LLM-based multimodal detection; browser extensions for social media monitoring; volunteer-sourced deepfake database	Deepfake content labeling standards; collaboration between fact-checkers and Deepfake Analysis Unit (DAU); AI Incident Database integration	South Asia–specific detection tools; Bengali-language model training; tools capable of processing regional script and audio
Victim Support & Community Infrastructure	Victim-centric response platforms connecting NGOs with affected individuals; mental health referral pathways; debunking tools for rapid correction	Grassroots initiatives in local languages; multi-stakeholder accountability coalitions; survivor-centered legal aid networks	Safe and encrypted reporting channels for marginalized groups (queer, indigenous, rural women); cross-disciplinary support networks; peer counseling programs
Platform Accountability	Faster content takedown processes; community-based content flagging mechanisms; transparency in moderation decisions for local-language content	Stricter enforcement of existing IT Rules; platform transparency requirements; independent oversight of AI moderation systems	Platform liability for non-removal of deepfake content; mandatory content provenance labeling; engagement with local civil society in policy design

provisions within the IT Act and faster enforcement of the IT Rules 2021; media literacy integrated into school curricula as a sustained component—with specific attention to students without device access, particularly girls—multilingual public campaigns, and training for healthcare professionals to identify synthetic medical content; deepfake content labeling standards, fact-checker collaboration with the DAU, and AI Incident Database integration; survivor-centered legal aid networks and multi-stakeholder accountability coalitions; and platform transparency requirements with independent oversight of AI moderation.

Bangladeshi participants called for building legal infrastructure from the ground up:

Table 4: Cross-country comparison of key harm themes identified in workshops.

Theme	Sri Lanka	India	Bangladesh
Gendered harms / non-consensual synthetic media	✓	✓	✓✓
Political misinformation & electoral manipulation	✓✓	✓✓	✓
Health misinformation via AI-generated content	—	✓✓	✓
Financial scams / celebrity endorsement fraud	✓	✓✓	✓
Low digital literacy among vulnerable populations	✓	✓✓	✓✓
Weak / absent national legal frameworks	✓	✓	✓✓
Platform under-moderation in local languages	✓✓	✓✓	✓
Psychological harm to victims (trauma, anxiety)	✓	✓	✓✓
Ethnic / communal tensions amplified by deepfakes	✓✓	✓	—
Distrust in AI tools & data privacy concerns	—	✓	✓✓

✓✓ = primary theme; ✓ = present; — = absent or marginal.

comprehensive new cyber law modeled on international frameworks such as the EU AI Act, with explicit provisions for gender-based synthetic abuse and image-based violence; literacy programs targeting rural and low-literate groups and feminist advocacy networks; South Asia–specific detection tools with Bengali-language model training; safe, encrypted reporting channels for marginalized groups—queer, indigenous, and rural women—with peer counseling programs; and platform liability for non-removal of deepfake content alongside mandatory provenance labeling and engagement with local civil society in policy design.

4.3 Cross-Country Thematic Comparison

Table 4 presents a structured comparison of the key harm themes identified across the three workshops, where ✓✓ indicates primary concerns that occupied significant discussion time and generated strong consensus, ✓ indicates themes present and substantively discussed, and — indicates themes absent or only briefly mentioned. The comparison reveals both the cross-cutting nature of certain harms—gendered targeting, platform under-moderation in local languages, and weak legal frameworks were salient in all three contexts—and the country-specific primacy of particular harm types.

4.4 Triangulated Findings: Cross-Cutting Patterns and Country-Specific Distinctions

Four cross-cutting patterns emerged across the three workshops. First, deepfake harm in all three contexts is fundamentally gendered: women, particularly in public-facing professional roles, are disproportionate targets across harm types—the most consistent cross-country signal in the data. Second, the inadequacy of platform governance for non-English content is experienced not as a technical inconvenience but as structural exclusion that amplifies harm in communities already under-resourced for digital safety. Third, institutional channels for redress—legal, regulatory, and platform-facing—

are perceived across all three contexts as effectively inaccessible to victims, creating dependency on informal community networks whose sustainability is precarious. Fourth, and most significantly for intervention design, participants in all three countries preferred community-embedded, locally governed solutions over externally designed technical fixes, reflecting a broader critique of top-down trust and safety paradigms that do not account for the knowledge, agency, and social organization of affected communities.

Country-specific distinctions matter equally. Sri Lanka's unique vulnerability lies in the intersection of deepfake harm with pre-existing ethnic polarization, creating the potential for synthetic content to catalyze rapid communal violence. India's challenge is one of scale and institutional complexity: the diversity of languages, legal jurisdictions, and harm types across over a billion internet users makes any single-point intervention inadequate. Bangladesh's most pressing need is the construction of basic legal infrastructure, accompanied by culturally sensitive support systems for victims who currently have no institutional recourse. These distinctions carry direct implications for governance frameworks, literacy interventions, and technological tools—a point to which we return in the Discussion.

5 Discussion

Our participatory workshops surfaced findings that both corroborate and extend the existing literature, while revealing structural gaps that Western-centric deepfake research has largely failed to address. We reflect on four interconnected themes that cut across our findings—the gendered architecture of deepfake harm, the structural failure of platform governance in multilingual contexts, the tension between technological solutionism and community-led resilience, and the limits of individual literacy as a defensive framework—before turning to implications for design, policy, and future work.

5.1 Deepfake Harm as a Gendered and Intersectional Problem

That women were the disproportionate targets of deepfake abuse in all three countries aligns with and deepens the literature on technology-facilitated gender-based violence (Henry et al. 2020; Chesney and Citron 2019). Across all three countries, participants described non-consensual synthetic pornography, fabricated images used for extortion, and reputational attacks on female political candidates as routine. This resonates with the foundational analysis by Chesney and Citron (2019) of deepfakes as tools of "democratic collapse and personal destruction," but our findings extend that account: the harms we document are not primarily about societal-level epistemic uncertainty but about intimate, targeted violence against individual women whose lives are materially disrupted.

Critically, deepfake harm is not uniformly gendered but intersectionally structured. In Bangladesh particularly, participants emphasized that queer individuals, indigenous

women, and rural users are more likely to be targeted, less likely to have access to legal recourse, and less visible to platform governance systems. This intersectional dimension is largely absent from existing deepfake literature, which has tended to treat gender as a binary variable rather than one axis of a more complex matrix of vulnerability (Scheuerman et al. 2020). Deepfake research and governance frameworks should adopt an explicitly intersectional lens, attending to how caste, class, sexuality, and geographic marginality compound the harms synthetic media can inflict.

The psychological harms participants described also demand attention from the HCI community. Design responses to deepfake harm have focused predominantly on detection and removal, but our findings suggest victim support infrastructure is at least as urgent a design challenge, echoing calls in the technology-facilitated abuse literature for "survivor-centered design" that centers the safety, agency, and emotional needs of those harmed rather than the technical properties of the harmful content itself (Freed et al. 2018; Havron et al. 2019).

5.2 The Structural Failure of Platform Governance in Multilingual Contexts

A second major theme concerns the systematic failure of platform T&S infrastructure to serve users in South Asian languages. Participants in all three workshops described content in Sinhala, Bengali, Tamil, and Hindi that they had personally verified as synthetic remaining online despite multiple reports, while comparable English-language content was removed quickly. This asymmetry is not merely anecdotal: the audit by Shahid, Elswah, and Vashistha (2025) of automated moderation pipelines for low-resource languages demonstrates that such systems exhibit systematic colonial biases. Our findings show the human consequences of this structural bias: the moderation gap is not a neutral technical limitation but an active mechanism through which platform governance reproduces the power asymmetries of the Global North.

This connects to what De Gregorio and Stremlau (2023) call the "inequalities of content moderation"—platform policies designed for high-resource linguistic and legal environments consistently failing users in the Global South. Our participants' accounts make the lived experience of this failure visible: journalists unable to have verified synthetic political content removed before an election; lawyers unable to identify a platform contact to receive a victim's takedown request; fact-checkers receiving automated rejections of forensically documented reports. These are not edge cases but systemic conditions shaping the information environment experienced by the majority of the world's internet users.

The colonial division of labor Das et al. (2024) identify in natural language processing—Global South workers performing the labeling labor that trains moderation systems while Global North institutions retain design and policy control—was echoed in participants' own relationships to platform governance: several India participants contributed to fact-checking databases that fed platform moderation pipelines while having no meaningful

influence over how that data was used. Decentralizing platform governance is thus not simply a technical challenge but a political one, requiring deliberate redistribution of decision-making authority toward the communities most affected by synthetic media harm.

5.3 Beyond Technological Solutionism: Community Knowledge as Infrastructure

A recurring tension in our workshops was between participants' awareness of technological detection tools and their skepticism that such tools could adequately address the harms they were experiencing—a tension that maps onto the wider HCI debate about technological solutionism, the tendency to frame complex social problems as technical puzzles amenable to engineering fixes (Morozov 2013). Participants were not opposed to technology; the Sri Lanka and India workshops generated sophisticated proposals for detection platforms, browser extensions, and LLM-based multimodal verification systems. But they consistently embedded these proposals within a larger vision of community-governed infrastructure in which technology serves human judgment rather than replacing it.

This vision resonates with the call by Sengers et al. (2005) for "reflective design," which surfaces and interrogates the values embedded in technical systems and makes space for communities to shape them, and with recent work on participatory AI governance arguing that effective oversight of AI systems in the Global South requires genuine co-production of governance frameworks, not merely consultation (Birhane et al. 2022). Our co-design findings suggest that South Asian stakeholders possess substantial, largely untapped expertise about the social dynamics of deepfake harm in their own contexts—expertise that neither platform governance systems nor academic research programs have systematically engaged.

The informal coping strategies our participants described—collaborative journalist verification networks in India, feminist peer-support systems in Bangladesh, civil society coalitions in Sri Lanka—are not merely workarounds for institutional failure. They are forms of community knowledge infrastructure that have evolved in direct response to the inadequacy of formal governance, embodying sophisticated local understanding of how synthetic media harm propagates in specific cultural and linguistic environments. New tools and platforms should therefore augment and formalize these networks rather than replace them: interoperating with existing community channels (including the encrypted messaging platforms that serve as primary communication infrastructure in all three contexts), supporting local-language interfaces, and keeping governance of the tools themselves in community hands.

5.4 Rethinking Literacy: From Individual Skill to Collective Resilience

Media and digital literacy emerged as a top-ranked intervention in all three workshops, but participants consistently reframed the concept: not as an individual cognitive skill of detecting manipulation in a given piece of content, but as a collective, relational, and culturally embedded form of resilience inseparable from community trust, institutional accountability, and political power. This reframing aligns with recent scholarship critiquing the "deficit model" of media literacy, which locates the problem of misinformation in individual cognitive failures and its solution in individual detection training (Pennycook and Rand 2021; Wineburg and McGrew 2019).

The deficit model is particularly inadequate in South Asian contexts for two reasons. First, the scale and speed of deepfake circulation through encrypted platforms like WhatsApp and Telegram—largely beyond the scope of platform moderation—means individual detection skill is insufficient as a primary defense: as India participants noted, by the time a single user identifies fabricated content, it may already have reached millions through forward chains. Second, the social dynamics of trust, in which content shared by a trusted family member or religious authority is believed regardless of verifiability, mean that detection skill alone does not translate into resistance to harm (Banaji et al. 2019). What is needed are community-level structures that make trusted verification accessible at the point of encounter with synthetic content, rather than requiring each individual to develop independent forensic competence.

Participants in all three workshops expressed strong interest in gamified, narrative-based tools that build critical evaluation habits through contextually familiar scenarios—an approach supported by inoculation theory research demonstrating that pre-emptive exposure to manipulation techniques can reduce susceptibility to misinformation (Roozenbeek et al. 2022). But such tools must be deeply localized: not merely translated into regional languages but built around culturally resonant scenarios, recognizable social dynamics, and the harm types most prevalent in each context. Generic literacy materials developed for Global North audiences, participants argued, may even be counterproductive, implicitly modeling a relationship to media and authority that does not map onto South Asian information ecosystems.

5.5 Implications for Design

Our findings carry several concrete implications for HCI researchers and designers building Trust & Safety tools for the Global South.

Design for community governance, not individual users. The most consistently prioritized interventions across all three workshops were collective rather than individual: community verification networks, multi-stakeholder coalitions, and peer-support systems. Detection tools and literacy platforms should treat community governance as a first-class requirement, supporting collective sensemaking and distributed trust rather than

optimizing for individual accuracy scores.

Treat linguistic localization as a safety-critical requirement. The failure of existing moderation and detection systems to serve South Asian languages is a primary harm amplifier, not a secondary limitation; linguistic and cultural localization must be a core engineering requirement from the outset, not a post-hoc translation task.

Build survivor-centered support infrastructure. The HCI community's expertise in designing for safety in intimate partner violence and online harassment (Freed et al. 2018) should be applied to deepfake victimization, with attention to the particular challenge of content that can never be fully removed once circulated.

Engage existing informal networks as design partners. The journalist coalitions, feminist peer-support systems, and civil society verification networks that have emerged in all three contexts are both repositories of local expertise and potential distribution channels for new tools; co-design processes should engage them as genuine partners rather than as end-user populations.

5.6 Implications for Policy

Our findings extend and specify existing calls for Global South–sensitive platform governance (Shahid and Vashistha 2023; De Gregorio and Stremlau 2023). Platform content moderation policies must be evaluated on their performance in the languages spoken by the majority of global internet users. Regulatory frameworks in Sri Lanka, India, and Bangladesh should prioritize survivor protection and platform accountability over general definitional debates about synthetic media, given that the most urgent harms documented here—gender-based violence, electoral manipulation, health misinformation—are already occurring at scale in the absence of adequate legal responses. And international policy dialogue on AI governance should systematically include civil society representatives from the Global South, whose communities are most affected by synthetic media harm yet least represented in the institutional spaces where governance frameworks are designed.

5.7 Limitations and Future Work

Our participant samples, while deliberately diverse in professional background, are drawn predominantly from urban, digitally connected professionals; the perspectives of rural users, elderly populations, and individuals with low digital literacy—groups identified by participants themselves as particularly vulnerable—are present only indirectly, as reported by advocates and practitioners. Future work should employ ethnographic and community-based participatory methods to engage these populations directly.

The online format enabled cross-regional participation but may have introduced selection bias toward participants with reliable internet access and comfort with video-based interaction, and may have dampened embodied and affective expression; future iterations

should explore hybrid formats combining online accessibility with the relational depth of in-person co-design. Our study also covers only three South Asian contexts: Pakistan and Nepal, among others, face deepfake harm dynamics shaped by distinct political and linguistic configurations that merit dedicated investigation. The research team has planned follow-up workshops in Pakistan, and future work should extend systematic comparison to a wider range of Global South contexts, including sub-Saharan Africa and Southeast Asia, where analogous dynamics of platform under-governance and linguistic marginalization are well documented (Shahid, Elswah, and Vashistha 2025).

Finally, our study is cross-sectional, capturing stakeholder perceptions at a specific moment in a rapidly evolving technology landscape. Longitudinal research is needed to track how deepfake harm patterns, community coping strategies, and institutional responses evolve as generative AI capabilities become more accessible and regulatory frameworks develop—in particular, the behavioral and psychological effects of sustained exposure to deepfake-saturated information environments remain a critical and underexplored research frontier (Ahmed et al. 2025).

Taken together, these limitations point toward a research agenda that is more longitudinal, more geographically inclusive, and more directly engaged with affected communities.

6 Conclusion

Deepfakes are not arriving into a neutral information environment in South Asia. They are landing in societies already shaped by fragile democratic institutions, linguistic diversity, entrenched patriarchal norms, and platform governance systems that were never designed with their realities in mind. Through three participatory workshops with 49 stakeholders—journalists, lawyers, fact-checkers, feminist advocates, civil society practitioners, and academics—in Sri Lanka, India, and Bangladesh, we asked how communities perceive and experience the harms of deepfakes, and what locally grounded interventions they propose to strengthen Trust & Safety in their own digital ecosystems.

On the first question, participants described a harm landscape that is structurally gendered, linguistically marginalized, and institutionally abandoned: in Sri Lanka, deepfakes targeting politicians, public figures, and senior officials deployed for fraud and political manipulation, amplified by post-conflict ethnic tensions that make synthetic content uniquely combustible; in India, voice clones, fabricated celebrity endorsements, and AI-generated health misinformation circulating at unprecedented scale across regional-language platforms that automated moderation cannot serve; in Bangladesh, harm concentrated on women—non-consensual synthetic pornography, extortion through fabricated intimate images, and the silencing of journalists and activists—in a legal en-

vironment offering victims almost no recourse. Across all three countries the pattern is consistent: deepfake harm falls disproportionately on women and intersectionally marginalized communities, platform moderation structurally fails users whose languages are not English, and institutional channels for redress are effectively inaccessible. On the second question, participants did not call primarily for better detection technology. They called for deepfake-specific legal frameworks that treat synthetic abuse as a rights violation, community-governed verification infrastructure in the languages communities actually use, media literacy embedded in sustained community practice, and survivor-centered support platforms—in every domain preferring solutions developed *with* affected communities rather than *for* them, a principle about where governance authority should reside, not merely about design method.

Three findings stand out. First, deepfake harm in the Global South is structurally gendered and intersectionally distributed; survivor-centered design, feminist advocacy infrastructure, and legal frameworks that treat synthetic abuse as a serious rights violation are not peripheral concerns but the foundation of any credible Trust & Safety response in this region. Second, the failure of platform governance to serve South Asian language users is not a technical limitation awaiting an engineering fix but a political condition that reproduces the power asymmetries of the global digital economy, requiring a fundamental redistribution of governance authority toward affected communities. Third, the communities most harmed by deepfakes are not passive victims: their verification networks, feminist peer-support systems, and civil society coalitions constitute community knowledge infrastructure embodying sophisticated local understanding of how synthetic media harm propagates, and the most valuable contribution this field can make is to build the conditions under which that knowledge shapes the solutions themselves. The challenge of deepfakes in the Global South is ultimately not a problem of technology to be managed but one of trust, governance, and justice to be addressed. This paper has foregrounded the voices of those most affected as a first step toward a reimagined Trust & Safety paradigm—one that acknowledges the cultural diversity of harm, amplifies local agency, and treats literacy as a form of community empowerment and democratic resilience. The work of building that paradigm remains urgently ahead.

References

Ahmed, Saifuddin, Muhammad Masood, Adeline Wei Ting Bee, and Kei Ichikawa. 2025. "False Failures, Real Distrust: The Impact of an Infrastructure Failure Deepfake on Government Trust." *Frontiers in Psychology* 16. https://doi.org/10.3389/fpsyg.2025.1574840.

Article 19. 2020. "Bangladesh: Increase in Charges under DSA as Government Seeks to Silence Criticism," July 3, 2020. https://www.article19.org/resources/bangladesh-increase-in-charges-under-dsa-as-government-seeks-to-silence-criticism/.

Atherton, Daniel. 2025. "Incident 1140: Purported Deepfake of Sri Lankan President Anura Kumara Dissanayake Promotes Alleged Fraudulent Government Investment Scheme." AI Incident Database, Responsible AI Collaborative, June 18, 2025. https://incidentdatabase.ai/cite/1140/.

Balendra, Soorya. 2025. "Meta's AI Moderation and Free Speech: Ongoing Challenges in the Global South." *Cambridge Forum on AI: Law and Governance* 1. https://doi.org/10.1017/cfl.2025.5.

Banaji, Shakuntala, Ramnath Bhat, Anushi Agarwal, Nihal Passanha, and Mukti Sadhana Pravin. 2019. *WhatsApp Vigilantes: An Exploration of Citizen Reception and Circulation of WhatsApp Misinformation Linked to Mob Violence in India.* London, UK: Department of Media, Communications, London School of Economics, and Political Science. https://eprints.lse.ac.uk/104316/.

Birhane, Abeba, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. "Power to the People? Opportunities and Challenges for Participatory AI." In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization,* 1–8. https://doi.org/10.1145/3551624.3555290.

Brashier, Nadia M., and Daniel L. Schacter. 2020. "Aging in an Era of Fake News." *Current Directions in Psychological Science* 29 (3): 316–23. https://doi.org/10.1177/0963721420915872.

Braun, Virginia, Victoria Clarke, Nikki Hayfield, Louise Davey, and Elizabeth Jenkinson. 2022. "Doing Reflexive Thematic Analysis." In *Supporting Research in Counselling and Psychotherapy: Qualitative, Quantitative, and Mixed Methods Research,* 19–38. Springer. https://doi.org/10.1007/978-3-031-13942-0_2.

Brieger, Alexandra Rose. 2024. "Empowerment or Exploitation: A Qualitative Analysis of Online Feminist Communities' Discussions of Deepfake Pornography." Dissertation, Uppsala University. https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-532059.

Chesney, Bobby, and Danielle Keats Citron. 2019. "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security." *California Law Review* 107 (6): 1753–820. https://doi.org/10.15779/Z38RV0D15J.

Chowdhury, Archis. 2025. “Sheikh Hasina’s Confession on Detention Centres in Bangladesh Is a Deepfake.” BOOM Live Fact Check, February 18, 2025. https://www.boomlive.in/fact-check/sheikh-hasina-false-bangladesh-aynaghor-confession-deepfake-analysis-27795.

Christopher, Nilesh, and Varsha Bansal. 2024. “Indian Voters Are Being Bombarded with Millions of Deepfakes. Political Candidates Approve.” *Wired,* May 20, 2024. https://www.wired.com/story/indian-elections-ai-deepfakes/.

Das, Dipto, Shion Guha, Jed R. Brubaker, and Bryan Semaan. 2024. “The “Colonial Impulse” of Natural Language Processing: An Audit of Bengali Sentiment Analysis Tools and Their Identity-Based Biases.” In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems,* 1–18. https://doi.org/10.1145/3613904.3642669.

Das, Srijit. 2024. “Video of Ranveer Singh Criticising PM Modi Is a Deepfake AI Voice Clone.” BOOM Live Fact Check, April 18, 2024. https://www.boomlive.in/fact-check/viral-video-bollywood-actor-ranveer-singh-congress-campaign-lok-sabha-elections-claim-social-media-24940.

De Gregorio, Giovanni, and Nicole Stremlau. 2023. “Inequalities and Content Moderation.” *Global Policy* 14 (5): 870–79. https://doi.org/10.1111/1758-5899.13243.

Dourish, Paul, and Scott D. Mainwaring. 2012. “Ubicomp’s Colonial Impulse.” In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing,* 133–42. https://doi.org/10.1145/2370216.2370238.

Freed, Diana, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. 2018. ““A Stalker’s Paradise”: How Intimate Partner Abusers Exploit Technology.” In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems,* 1–13. https://doi.org/10.1145/3173574.3174241.

Government of India, The Information Technology Act, 2000 (2000), https://www.indiacode.nic.in/handle/123456789/1999.

Guess, Andrew M., Michael Lerner, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. 2020. “A Digital Media Literacy Intervention Increases Discernment between Mainstream and False News in the United States and India.” *Proceedings of the National Academy of Sciences* 117 (27): 15536–45. https://doi.org/10.1073/pnas.1920498117.

Guess, Andrew M., Jonathan Nagler, and Joshua A. Tucker. 2019. “Less Than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook.” *Science Advances* 5 (1): eaau4586. https://doi.org/10.1126/sciadv.aau4586.

Halder, Debarati, and K. Jaishankar. 2016. *Cyber Crimes against Women in India.* New Delhi: SAGE Publications India.

Hameleers, Michael, Toni G. L. A. Van Der Meer, and Tom Dobber. 2024. "Distorting the Truth versus Blatant Lies: The Effects of Different Degrees of Deception in Domestic and Foreign Political Deepfakes." *Computers in Human Behavior* 152. https://doi.org/10.1016/j.chb.2023.108096.

Havron, Sam, Diana Freed, Rahul Chatterjee, Damon McCoy, Nicola Dell, and Thomas Ristenpart. 2019. "Clinical Computer Security for Victims of Intimate Partner Violence." In *Proceedings of the 28th USENIX Security Symposium,* 105–22. https://www.usenix.org/conference/usenixsecurity19/presentation/havron.

Hayes, Gillian R. 2011. "The Relationship of Action Research to Human-Computer Interaction." *ACM Transactions on Computer-Human Interaction* 18 (3): 1–20. https://doi.org/10.1145/1993060.1993065.

Henry, Nicola, Clare McGlynn, Asher Flynn, Kelly Johnson, Anastasia Powell, and Adrian J. Scott. 2020. *Image-Based Sexual Abuse: A Study on the Causes and Consequences of Non-consensual Nude or Sexual Imagery.* Routledge. https://doi.org/10.4324/9781351135153.

Human Rights Watch. 2020. "Bangladesh: Repeal Abusive Law Used in Crackdown on Critics," July 1, 2020. https://www.hrw.org/news/2020/07/01/bangladesh-repeal-abusive-law-used-crackdown-critics.

Irani, Lilly, Janet Vertesi, Paul Dourish, Kavita Philip, and Rebecca E. Grinter. 2010. "Postcolonial Computing: A Lens on Design and Development." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 1311–20. https://doi.org/10.1145/1753326.1753522.

Joy, Shemin. 2024. "Lok Sabha Election 2024: Remove Fake Content within 3 Hours, EC Tells Political Parties." *Deccan Herald,* May 6, 2024. https://www.deccanherald.com/elections/india/lok-sabha-election-2024-remove-fake-content-within-3-hours-ec-tells-political-parties-3010408.

Kafayat, Sobia. 2025. "Deepfakes' Challenge to International Politics: Sowing Political Discord." *Annals of Human and Social Sciences* 6 (2): 318–26. https://doi.org/10.35484/ahss.2025(6-II)27.

Kalra, Aditya, Munsif Vengattil, and Dhwani Pandya. 2024. "Deepfakes of Bollywood Stars Spark Worries of AI Meddling in India Election." *Reuters,* April 22, 2024. https://www.reuters.com/world/india/deepfakes-bollywood-stars-spark-worries-ai-meddling-india-election-2024-04-22/.

Kapania, Shivani, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. ""Because AI Is 100% Right and Safe": User Attitudes and Sources of AI Authority in India." In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems,* 1–18. https://doi.org/10.1145/3491102.3517533.

Karthikeyan, S., and Anindita Roy. 2026. “Digital Trust in the Age of Deepfakes and Synthetic Media.” *Journal of Cyber Risk and Digital Trust* 1 (2). https://admin.mantechpublications.com/index.php/JCRDT/article/view/2756.

Kashyap, Sommya. 2025. “The Digital Mirage: India’s Evolving Legal Battle against Deepfake Technology.” *SCRIPTed: A Journal of Law, Technology & Society* 22 (2): 162–226. https://doi.org/10.2218/scrip.22.2.2025.12004.

Köbis, Nils C., Barbora Doležalová, and Ivan Soraperra. 2021. “Fooled Twice: People Cannot Detect Deepfakes but Think They Can.” *iScience* 24 (11). https://doi.org/10.1016/j.isci.2021.103364.

KPMG. 2025. *Trust, Attitudes and Use of Artificial Intelligence: A Global Study 2025.* KPMG International. https://kpmg.com/xx/en/our-insights/ai-and-technology/trust-attitudes-and-use-of-ai.html.

Ministry of Electronics and Information Technology, Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 (2021), https://www.meity.gov.in/static/uploads/2024/02/Information-Technology-Intermediary-Guidelines-and-Digital-Media-Ethics-Code-Rules-2021-updated-06.04.2023-.pdf.

Mirsky, Yisroel, and Wenke Lee. 2022. “The Creation and Detection of Deepfakes: A Survey.” *ACM Computing Surveys* 54 (1): 1–41. https://doi.org/10.1145/3425780.

Morozov, Evgeny. 2013. *To Save Everything, Click Here: The Folly of Technological Solutionism.* PublicAffairs.

Newman, Nic, Richard Fletcher, Craig Robertson, Kirsten Eddy, and Rasmus Kleis Nielsen. 2024. *Reuters Institute Digital News Report 2024.* Oxford: Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2024-06/RISJ_DNR_2024_Digital_v10%20lr.pdf.

Newswire. 2025. “Kumar Sangakkara Warns Public of AI Deepfake Videos Misusing His Image.” *Newswire.lk,* May 6, 2025. https://www.newswire.lk/2025/05/06/kumar-sangakkara-warns-public-of-ai-deepfake-videos-misusing-his-image/.

Paul, Swapan, and Bidyarthi Dutta. 2025. “Unveiling Trends in Deepfake Research: A Global Bibliometric Perspective.” *DESIDOC Journal of Library & Information Technology* 45 (3): 224–37. https://doi.org/10.14429/djlit.20858.

Pawelec, Maria. 2022. “Deepfakes and Democracy (Theory): How Synthetic Audio-Visual Media for Disinformation and Hate Speech Threaten Core Democratic Functions.” *Digital Society* 1 (2). https://doi.org/10.1007/s44206-022-00010-6.

Pennycook, Gordon, and David G. Rand. 2021. “The Psychology of Fake News.” *Trends in Cognitive Sciences* 25 (5): 388–402. https://doi.org/10.1016/j.tics.2021.02.007.

Popowicz, Joasia E. 2025. "Sri Lanka's Governor Is Latest to Feature in Deepfake Scam: Video Uses Manipulated Image to Pretend Central Bank Chief Is Endorsing Financial Investments." *Central Banking,* March 1, 2025. https://www.centralbanking.com/technology/7972454/sri-lankas-governor-is-latest-to-feature-in-deepfake-scam.

Roozenbeek, Jon, Sander van der Linden, Beth Goldberg, Steve Rathje, and Stephan Lewandowsky. 2022. "Psychological Inoculation Improves Resilience against Misinformation on Social Media." *Science Advances* 8 (34): eabo6254. https://doi.org/10.1126/sciadv.abo6254.

Samaratunge, Shilpa, and Sanjana Hattotuwa. 2014. *Liking Violence: A Study of Hate Speech on Facebook in Sri Lanka.* Colombo, Sri Lanka: Centre for Policy Alternatives. https://www.cpalanka.org/liking-violence-a-study-of-hate-speech-on-facebook-in-sri-lanka/.

Sanders, Elizabeth B.-N., and Pieter Jan Stappers. 2008. "Co-creation and the New Landscapes of Design." *CoDesign* 4 (1): 5–18. https://doi.org/10.1080/15710880701875068.

Scheuerman, Morgan Klaus, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker. 2020. "How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis." In *Proceedings of the ACM on Human-Computer Interaction,* 4:1–35. CSCW1. https://doi.org/10.1145/3392866.

Schiff, Daniel S., Kaylyn Jackson Schiff, and Natalía Bueno. 2024. "Watch Out for False Claims of Deepfakes, and Actual Deepfakes, This Election Year." Brookings Institution, May 30, 2024. https://www.brookings.edu/articles/watch-out-for-false-claims-of-deepfakes-and-actual-deepfakes-this-election-year/.

Sengers, Phoebe, Kirsten Boehner, Shay David, and Joseph Kaye. 2005. "Reflective Design." In *Proceedings of the 4th Decennial Conference on Critical Computing,* 49–58. https://doi.org/10.1145/1094562.1094569.

Shahid, Farhana, Mona Elswah, and Aditya Vashistha. 2025. "Think Outside the Data: Colonial Biases and Systemic Issues in Automated Moderation Pipelines for Low-Resource Languages." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 8 (3): 2331–44. https://doi.org/10.1609/aies.v8i3.36719.

Shahid, Farhana, and Aditya Vashistha. 2023. "Decolonizing Content Moderation: Does Uniform Global Community Standard Resemble Utopian Equality or Western Power Hegemony?" In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems,* 1–18. https://doi.org/10.1145/3544548.3581538.

Sunvy, Ahmed Shafkat, Raiyan Bin Reza, and Abdullah Al Imran. 2023. "Media Coverage of DeepFake Disinformation: An Analysis of Three South-Asian Countries." *Informasi* 53 (2): 295–308. https://doi.org/10.21831/informasi.v53i2.66479.

Justice K.S. Puttaswamy v. Union of India, Writ Petition (Civil) No. 494 of 2012 (2017).

Taub, Amanda, and Max Fisher. 2018. “Where Countries Are Tinderboxes and Facebook Is a Match.” *The New York Times,* April 21, 2018. https://www.nytimes.com/2018/04/21/world/asia/facebook-sri-lanka-riots.html.

Vaccari, Cristian, and Andrew Chadwick. 2020. “Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News.” *Social Media + Society* 6 (1). https://doi.org/10.1177/2056305120903408.

Wineburg, Sam, and Sarah McGrew. 2019. “Lateral Reading and the Nature of Expertise: Reading Less and Learning More When Evaluating Digital Information.” *Teachers College Record* 121 (11): 1–40. https://doi.org/10.1177/016146811912101102.

Al-Zaman, Md. Sayeed. 2021. “COVID-19-Related Online Misinformation in Bangladesh.” *Journal of Health Research* 35 (4): 364–68. https://doi.org/10.1108/JHR-09-2020-0414.

Zhang, Weidi. 2024. “Seeing Is Not Believing: AI Face Based on Deepfake Technology.” In *2024 10th International Conference on Humanities and Social Science Research (ICHSSR 2024),* 1247–52. https://doi.org/10.2991/978-2-38476-277-4_139.

Authors

Dilrukshi Gamage is a senior lecturer at the University of Colombo School of Computing (UCSC) and the lead researcher located in Sri Lanka exploring societal impact of deepfakes. Email: dilrukshi.gamage@gmail.com.

Dilki Sewwandi is one of the research assistants for the project leading workshop facilitation and data collection in Sri Lanka. Email: dsewwandi2001@gmail.com.

Shreeti Shubham is one of the research assistants for the project leading the workshop facilitation and data collection in India. Email: shubhamshreeti67@gmail.com.

Sumaia Arefin is one of the research assistants for the project leading the workshop facilitation and data collection in Bangladesh. Email: sumaiaritu550@gmail.com.

Dilaxshini Dunstan is one of the research assistants for the project leading the workshop facilitation and data collection in Sri Lanka. Email: dilaxshini16dunstan@gmail.com.

Kartik Joshi is a master's student in India and helped the project in facilitating the workshop and collecting data from the India workshop. Email: kartik.joshi@iiitb.ac.in.

Rashmi Gunawardana is a final year undergraduate student at UCSC who helped with the project workshop facilitation and data collection in Sri Lanka. Email: nirashagunawardana9@gmail.com.

Acknowledgments

We thank all the participants from India, Sri Lanka, and Bangladesh for spending their valuable time sharing insights with the authors. We also thank all the contacts who helped us find the right participants for the workshops. We also thank Google Academic Research Award that helped us to finance our work. We thank the Director and team from University of Colombo School of Computing (UCSC) in Sri Lanka for facilitating collaborative research across multiple countries in South Asia.

Data Availability Statement

Workshop materials, specifically the activities used, can be shared upon request.

Funding Statement

This research was supported by the independent research grant GARA: Google Academic Research Award provided to the lead author.

Ethical Standards

We received the ethical review board approval for this research under the approval code: ERC-5-8-001.

Keywords

Deepfake harms; Global South; AI governance; digital literacy; coping mechanisms.