International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Ethical Challenges of Artificial Intelligence on Social Media

A Socio-Ethical Analysis with an Indian Perspective

Ms. A Yasaswini Reddy | Ms. A Vaishnavi Priya | Ms. A Supraja

BS-CS, Vishwa Vishwani Institute of Systems and Management, Hyderabad, Telangana, India Email: yasaswinireddy29@gmail.com | Mobile: 8341720229

Abstract

As artificial intelligence (AI) drives modern social media platforms through personalized recommendations, content moderation, and generative tools, it introduces profound moral and socio-ethical concerns. This paper investigates core ethical challenges, including algorithmic bias, data exploitation, deepfakes, commercial surveillance, and the rapid spread of misinformation that depends on ideological polarization. Evaluating these dimensions through a socio-ethical lens, the study highlights the disproportionate impact of automated moderation on marginalized groups and the unique digital complexities within the Indian context, marked by high linguistic diversity and varying media literacy. The research demonstrates that voluntary platform self-regulation is insufficient to counter systemic digital harms. Consequently, it advocates for robust governmental oversight, mandatory algorithmic transparency, independent audits, and human-centric governance. Ultimately, the study concludes that responsible AI deployment must prioritize human dignity, social justice, democratic values, and inclusive digital participation to establish a secure and trustworthy online ecosystem.

Keywords: Artificial intelligence, social media ethics, algorithmic bias, data privacy, deepfakes, digital governance

1. Introduction

Social media platforms are often described as communication spaces, but their everyday operation is increasingly shaped by automated systems. AI classifies content, predicts which posts a person may engage with, personalises feeds, identifies suspected violations and supports advertising decisions. More recently, generative AI has lowered the cost of producing convincing text, images,

voices and videos. These technologies may help users find relevant information, translate across languages, improve accessibility and moderate content at a scale that manual review alone cannot achieve. At the same time, they influence the conditions under which people speak, learn, organise and participate in public life.

The central ethical question is not whether AI is inherently beneficial or harmful. It is how systems are designed, trained, deployed and governed; which interests they serve; and how their effects are distributed. A ranking model that increases average engagement can still reduce exposure to minority viewpoints. A moderation classifier that performs well in one language may fail on local idiom, satire or reclaimed speech. A personalised advertisement may be relevant to one user while relying on intrusive inferences about another. The social consequences of AI therefore cannot be measured by technical performance alone.

This paper is a conceptual analysis rather than an empirical study. It draws on UNESCO's Recommendation on the Ethics of Artificial Intelligence, Indian data-protection and intermediary frameworks, and selected research on information curation and recommender systems. The focus is on ethical challenges associated with platform AI, rather than a legal opinion on any particular company or content decision. The paper asks: (a) what ethical risks arise when AI mediates visibility and interaction on social media; (b) how do these risks intersect with India's linguistic, social and institutional diversity; and (c) what governance measures can make platform AI more accountable and rights-respecting?

The argument is that ethical governance must look beyond the model as an isolated artefact. It must address the full sociotechnical system: data collection, objective functions, interface choices, human review, business incentives and avenues for redress. The following sections outline the approach and ethical foundations, examine five areas of concern, and offer a practical governance framework.

2. Method and Scope

The paper uses a structured conceptual review. It brings together normative guidance on AI ethics, relevant Indian policy instruments and peer-reviewed work that illustrates how recommender systems can shape information exposure. Sources were selected for relevance to social media,

human rights and platform accountability. The analysis is interpretive: it does not estimate the prevalence of harms, compare platforms, or claim causal effects for Indian users. The examples are used to clarify mechanisms and governance questions, not to attribute unverified conduct to a named service.

The term AI is used broadly for computational systems that infer, rank, predict, classify or generate content in ways that influence social media experiences. This includes machine-learning recommenders and moderation tools, as well as generative systems embedded in platform workflows. Not every automated rule is AI, and not every platform outcome is caused by AI. This distinction matters because simplistic claims of algorithmic causation can obscure user behaviour, social networks, external media and platform policy. Research on recommendation systems, for example, cautions that evidence of amplification depends on the audit method, user choices and the content being measured (Horta Ribeiro et al., 2023).

3. Ethical Foundations

UNESCO's 2021 Recommendation treats human dignity and human rights as central to AI ethics and identifies fairness, transparency, accountability, privacy, human oversight, inclusion and sustainability as guiding principles. It specifically recognises that AI-driven curation and moderation can affect access to information, freedom of expression, discrimination and cultural diversity. These principles are useful for social media because platforms do not simply host communication: their systems structure what is discoverable, prominent, restricted or monetised.

A rights-based approach is complemented by social justice. Equal treatment in code does not necessarily produce equitable outcomes where data and social conditions are unequal. Ethical assessment should ask whether design choices burden groups differently, whether affected people can participate in decisions, and whether harms can be remedied. It should also recognise the asymmetry between platforms that possess extensive behavioural data and users who cannot see or negotiate the systems shaping their experience.

Critical scholarship helps explain why technical systems should be examined alongside the institutions and commercial models that shape them. Noble (2018) shows how information technologies can reproduce social hierarchies through the categories, data and rankings they use.

Zuboff (2019) analyses the extraction of behavioural data as an economic model that supports prediction and influence. Applied to social media, these perspectives direct attention to whose knowledge is represented, how personal data become commercially valuable and whether the pursuit of engagement conflicts with users' autonomy and the public interest.

This paper brings those critical perspectives into conversation with research on information curation. Lazer et al. (2024) propose evaluating curation algorithms at both individual and societal levels, including the quality of information and the effects on agency. Horta Ribeiro et al. (2023) caution that recommender-system audits can produce different impressions depending on how user behaviour is modelled. Together, this literature supports a balanced approach: examine structural incentives and unequal effects, while avoiding unsupported claims that a single algorithm mechanically determines belief or behaviour.

4. Key Ethical Challenges

4.1 Privacy, surveillance and data power

Personalisation depends on data. Platforms may process searches, interactions, location signals, device information and inferred interests to predict what a user will view or buy. The ethical problem is not limited to a single data field. Combining seemingly ordinary signals can produce intimate inferences, and users may not understand how those inferences affect recommendations, advertising or visibility. Consent may be formally obtained yet practically weak when services are difficult to avoid, notices are complex, or opting out reduces core functionality.

Large-scale profiling can also make surveillance routine. Persistent observation may chill participation, particularly for people discussing sensitive health, political or identity-related topics. Data collected for one purpose can be reused, shared or retained in ways that exceed users' expectations. Generative AI introduces further questions where personal content is reused in training or generation workflows, and where synthetic outputs imitate an individual's likeness or voice. Ethical safeguards should therefore include purpose limitation, data minimisation, secure retention practices and clear controls over profiling, alongside legally required protections.

India's Digital Personal Data Protection Act, 2023 establishes a statutory framework for processing digital personal data, with provisions taking effect as notified by the Central Government. Its enactment makes privacy and data governance particularly salient, but the precise obligations applicable to a platform depend on commencement notifications, rules and the circumstances of processing. Ethical practice should not wait for minimum compliance: platforms should explain meaningful data uses in accessible language, provide workable choices and test whether users can exercise their rights in practice.

4.2 Bias, representation and unequal visibility

AI systems learn from data that reflect patterns of past participation and institutional power. If some communities are underrepresented, their language, humour and concerns may be poorly understood. If abusive content is common in the training data, a moderation model may normalise it or fail to distinguish between a threat and a report condemning a threat. Conversely, systems may disproportionately flag dialect, political dissent or reclaimed expressions as unsafe. These failures can suppress legitimate speech or leave targeted communities exposed to abuse.

Bias also operates through ranking. Engagement models can favour creators who already have large audiences, content that provokes strong reactions, or languages for which the platform has better tools. This may marginalise regional-language creators, smaller publishers and minority perspectives. The issue is not that every ranking system must show every view equally; that would be neither feasible nor necessarily desirable. The ethical requirement is to identify systematic disparities, understand their causes and provide meaningful ways to correct them.

Fairness assessment must include context. Aggregate accuracy can conceal poor performance for a small language group. Evaluation should therefore be disaggregated where lawful and safe, conducted with community input, and attentive to the risks of collecting sensitive attributes. Diverse review teams and local-language expertise are essential, but they cannot substitute for ongoing measurement and public accountability.

4.3 Misinformation, synthetic media and information integrity

Generative systems can produce persuasive content at low cost and high volume. Fabricated images, cloned voices or synthetic video can be used for satire and creative expression, but also to impersonate public figures, harass individuals or mislead audiences during crises and elections. Automation may allow coordinated campaigns to vary their messages and reach users quickly. The resulting challenge concerns both authenticity and attribution: people need ways to assess whether content is synthetic, manipulated or distributed in coordination.

Recommendation systems can influence the circulation of low-credibility or polarising material, but claims about their effects require care. An audit that follows recommendations mechanically may reveal what a system can serve, while real users may choose differently. Horta Ribeiro et al. (2023) show why recommendation effects can appear paradoxical across audit methods. A responsible analysis should distinguish exposure, engagement, reshares and belief, and should not equate a platform's recommendation with a proven change in opinion.

Platforms can reduce risk through provenance signals, friction before resharing, contextual notices and rapid responses to coordinated manipulation. These measures should be calibrated: labels can be wrong, synthetic media can be used for legitimate purposes, and excessive removal can itself distort public debate. Provenance information is most useful when it is understandable, reliable and accompanied by an appeal route. Media and information literacy should complement platform interventions by helping users evaluate sources, context and uncertainty.

4.4 Content moderation, due process and freedom of expression

Automated moderation can detect known spam, child sexual abuse material and other clearly defined violations quickly. It can also help prioritise reports for human review. Yet contextsensitive decisions are difficult to automate. Meaning may depend on language, local law, political context, irony, quotation or the relationship between speaker and target. A system may remove a journalist's documentation of violence, leave a disguised threat online, or apply a rule inconsistently across languages.

The ethical stakes include both under-enforcement and over-enforcement. Under-enforcement can leave people exposed to threats, harassment and incitement. Over-enforcement can silence users,

including those who rely on social media to report abuses or participate in public debate. The platform's decision process should therefore provide notice of consequential actions, a comprehensible reason, a timely appeal and human reconsideration for high-impact or contested cases. Users should be able to correct errors without needing specialised technical knowledge.

India's Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, as amended, set out duties for intermediaries and grievance mechanisms. The framework is evolving, and platform duties depend on the rules and facts applicable at the time. Whatever the legal position in a specific case, ethical governance requires transparent policies, consistent application, accurate reporting and independent scrutiny. Legal compliance alone does not answer whether moderation is fair, proportionate or culturally competent.

4.5 Autonomy, well-being and engagement-oriented design

Personalised feeds can help users navigate abundant information. But when systems optimise repeated engagement, they may exploit attention and behavioural vulnerabilities. Notifications, autoplay, infinite scroll and emotionally salient recommendations can make it difficult for users to disengage. For some users, including children and people experiencing distress, continuous exposure to particular content may intensify anxiety, social comparison or harmful behaviour. The strength and direction of these effects vary; ethical assessment should avoid assuming a uniform outcome for all users.

Autonomy concerns arise when people cannot understand why they are seeing content or cannot meaningfully adjust the system. A recommendation that appears neutral may reflect commercial priorities, inferred traits or past behaviour. Users should have accessible controls to reset or diversify recommendations, reduce sensitive personalisation, and understand the principal factors shaping their feed. Such controls should be usable without a substantial penalty or obscure settings journey.

Platforms should assess well-being alongside engagement, including the impacts on children and vulnerable groups. This does not imply that platforms can or should make clinical judgments about users. It does mean that known design risks should be examined, and that interventions should be

tested for unintended effects. Independent research access, with safeguards for privacy, can help establish evidence beyond platform self-reporting.

4.6 Accountability, explainability and labour

The complexity of modern AI systems can make accountability diffuse. A platform may point to an automated classifier, a vendor or a user report, while the person affected cannot identify who can correct the result. Ethical responsibility should remain with the organisations that select objectives, deploy systems and benefit from them. Documentation, traceability and audit logs are necessary to establish what happened and to investigate patterns of harm.

Transparency should be meaningful rather than merely technical. Publishing model details that expose private information or enable manipulation may be counterproductive, but users and oversight bodies need enough information to understand system purposes, key factors and error patterns. Public transparency reports can describe moderation volumes, appeal outcomes, language coverage and known limitations. Independent audits can examine distributional impacts and whether mitigation works.

Human review also depends on working conditions. Moderators may encounter graphic violence, abuse and other disturbing material, often under time pressure. A policy that promises human oversight is not credible if reviewers lack training, language expertise, time or psychological support. Responsible deployment includes adequate staffing, safe workloads and access to support for all workers involved in content review, including contracted workers.

5. India-Specific Considerations

India's social media environment is marked by linguistic plurality, uneven digital access and sharply differentiated social positions. A system trained or evaluated primarily in English may perform poorly in languages and dialects used by large communities. Transliteration, codeswitching, local idiom and regional political references complicate both moderation and misinformation detection. Limited local-language support can create a double harm: users may receive less protection against abuse while also facing greater risk that legitimate speech is incorrectly restricted.

Digital inequalities influence who can benefit from AI and who can challenge its decisions. Users with reliable connectivity, digital literacy and institutional support are better placed to report errors and seek redress. Marginalised communities may face higher exposure to surveillance and harassment, while having fewer resources to contest profiling or removal. A national policy discussion should therefore include civil society organisations, linguistic communities, educators, journalists, child-rights advocates and researchers, not only platforms and state institutions.

India's evolving governance framework combines data protection rules, intermediary obligations and broader constitutional and statutory protections. The interaction among these frameworks is complex and may change through notifications, judicial interpretation and amendments. A conference paper should consequently treat this analysis as an ethical discussion, not as a substitute for jurisdiction-specific legal advice. The durable policy point is that accountability mechanisms should be accessible in Indian languages, responsive to local context and consistent with fundamental rights.

6. Limits of Voluntary Self-Regulation

Platform policies, internal safety teams and voluntary ethics commitments can contribute to responsible practice. They also have structural limits: platforms set many of the objectives, control the operational data and decide what evidence to disclose. Commercial incentives may conflict with changes that reduce intrusive profiling, slow the circulation of sensational material or make it easier for users to disengage. This does not establish that every platform deliberately promotes harm. It does mean that internal commitments alone cannot provide independent assurance or an effective remedy in every case.

External oversight is needed to address information asymmetry and conflicts of interest. Public standards, independent audits, accessible appeals and transparent reporting can test whether safeguards work in practice. Regulation should remain proportionate and rights-based: rules addressing harmful content must be clearly defined, and state intervention must be subject to due process and independent review so it cannot become a tool for censorship or political surveillance.

7. A Framework for Responsible Social Media AI

The following framework translates ethical principles into practical obligations across the AI lifecycle. It is intended as a governance proposal for platforms and a reference point for policy, research and civil society oversight.

1. **Assess rights and social impacts before deployment.** Platforms should identify affected groups, foreseeable harms and affected rights before introducing high-impact ranking, profiling or moderation systems. Assessments should include privacy, freedom of expression, equality, child safety and access to information, with mitigation plans and periodic review.
2. **Build for representative performance.** Training and evaluation should account for linguistic and cultural variation. Platforms should measure error rates across relevant language communities and user groups, document gaps, and involve affected communities in designing corrective measures. Data collection for fairness must itself be privacy-protective.
3. **Give users intelligible control and explanation.** People should be told when AI materially shapes a decision affecting their account, content or feed. They should receive concise reasons, accessible options to adjust personalisation, and notice when content is labelled or restricted.
4. **Provide effective contest and remedy.** Appeals should be timely, available in supported languages and reviewed by a suitably trained person when the decision is consequential or contested. Platforms should track reversal rates and recurring error patterns, then use that evidence to improve systems and policies.
5. **Enable independent scrutiny.** Qualified researchers and auditors should have access to appropriate data under safeguards that protect privacy, security and platform integrity. Regulators and the public should receive regular, comparable reporting on system purposes, significant changes, performance limits and mitigation outcomes.
6. **Align incentives with public interest.** Platforms should test whether engagement objectives create predictable risks and should include safety, diversity and user well-being in product evaluation. Governance should apply to the business model and interface choices as well as the underlying model.

7. **Invest in human capacity and literacy.** Human oversight needs adequate staffing, training, local expertise and support for moderators. Public programmes should strengthen media and AI literacy so users can assess synthetic content, recommendations and privacy choices.
8. **Coordinate proportionate, adaptive regulation.** Regulatory duties should be proportionate to the scale and risk of a system, while preserving fundamental rights. Public authorities should consult affected communities, publish guidance and review rules as technology and evidence evolve.

8. Discussion

The framework rests on three conclusions. First, fairness and safety are contextual. A system cannot be declared ethical because it meets a single accuracy threshold or because the same rule is applied to every user. Evaluation must examine who experiences errors, what opportunities are lost and whether remedies are accessible. Second, platform transparency must be tied to power. Information is valuable when users, researchers and oversight institutions can use it to identify problems and seek change. Third, accountability must follow the system through its lifecycle. A model update, interface change or new commercial objective may alter outcomes even if the underlying technology appears unchanged.

These conclusions also caution against technological determinism. AI systems shape information environments, but they operate alongside user choices, social networks, political actors, media institutions and offline inequalities. Ethical analysis should neither absolve platforms by attributing harms to users alone nor attribute every social problem to algorithms. Carefully designed research and transparent platform data are needed to distinguish correlation, exposure and causal impact.

A further challenge is balancing competing rights. More aggressive moderation may reduce some harms while restricting lawful expression; greater transparency may support accountability while exposing personal data or enabling adversarial gaming. Such tensions do not remove the need for action. They require proportional safeguards, documented reasoning, accessible contestation and periodic review by institutions with sufficient independence.

9. Conclusion

AI has become a consequential intermediary in social media, influencing what people see, how they are represented and whether their speech remains accessible. Its ethical challenges include intrusive profiling, unequal visibility, synthetic deception, inconsistent moderation, weakened autonomy and opaque responsibility. These harms are intensified when systems operate across social and linguistic contexts that are poorly represented in their design or evaluation.

A rights-respecting response should make platforms responsible for the systems they deploy while enabling meaningful scrutiny by regulators, researchers and affected communities. This paper therefore calls for robust governmental oversight, mandatory algorithmic transparency proportionate to risk, independent audits and public reporting, privacy-protective data practices, language-sensitive evaluation, accessible appeals and humane conditions for content reviewers. These measures should be developed with community participation and reviewed as evidence and technology change. In India, accountability must reflect linguistic diversity and unequal capacity to exercise digital rights. Responsible social media AI should advance human dignity, social justice, democratic values and inclusive participation. The quality of governance will be measured by whether people can understand, influence and challenge systems that mediate their public and private lives.

References

Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. New York University Press.

Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. PublicAffairs.

Government of India. (2023). The Digital Personal Data Protection Act, 2023 (Act No. 22 of 2023). Ministry of Law and Justice. https://www.meity.gov.in/writereaddata/files/Digital%20Personal%20Data%20Protection%20Act%202023.pdf

Government of India, Ministry of Electronics and Information Technology. (2021, as amended).

The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021. https://www.meity.gov.in/content/information-technology-intermediateguidelines-and-digital-media-ethics-code-rules-2021

Horta Ribeiro, M., Veselovsky, V., & West, R. (2023). The amplification paradox in recommender systems. Proceedings of the International AAAI Conference on Web and Social Media, 17(1), 1138–1142. https://doi.org/10.1609/icwsm.v17i1.22223

Lazer, D., Swire-Thompson, B., & Wilson, C. (2024). A normative framework for assessing the information curation algorithms of the internet. Perspectives on Psychological Science, 19(5). https://doi.org/10.1177/17456916231186779

UNESCO. (2021). Recommendation on the ethics of artificial intelligence. https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence

United Nations. (1966). International Covenant on Civil and Political Rights. https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenantcivil-and-political-rights