Journal of Advances in Developmental Research (IJAIDR)

E-ISSN: 0976-4844 ● Website: www.ijaidr.com ● Email: editor@ijaidr.com

The Regional Language Challenge: Artificial Intelligence, Disinformation and the Vulnerability of India's Vernacular Digital Public Sphere

Sushmita Kumari[1], Atish Prashar[2], Anand Kumar[3]

[1]Research Scholar, Department of Mass Communication and Media,
Central University of South Bihar, Gaya.
sushmitakri090@gmail.com
[2]Professor and Head, Department of Mass Communication and Media,
Central University of South Bihar, Gaya.
[3]Research Scholar, Department of Mass Communication and Media, Central University of South Bihar, Gaya.

Abstract:
India's digital public sphere is increasingly shaped by mobile-first communication, platform-mediated news consumption, and the growing use of regional-language content. Artificial intelligence (AI) has the potential to widen participation in this environment through machine translation, speech recognition, text-to-speech systems, automated subtitling, and generative content tools. At the same time, these technologies create new pathways for the production, translation, personalization, and circulation of misleading information. The risks are particularly complex in a multilingual society where language varieties, dialects, scripts, code-switching, local political contexts, and uneven digital-literacy levels influence how information is created and interpreted. This conceptual paper examines the relationship between AI, disinformation, and the vulnerability of India's vernacular digital public sphere. Drawing upon the theories of the public sphere, networked gatekeeping, information disorder, platformization, and sociotechnical systems, the paper develops a conceptual framework that links AI-enabled content production, linguistic and contextual vulnerabilities, platform distribution mechanisms, audience interpretation, and verification capacity. The framework identifies six interrelated dimensions: multilingual AI capability, vernacular content ecology, platform amplification, socio-cultural susceptibility, verification and media-literacy capacity, and institutional governance. It argues that vulnerability does not arise from regional-language communication itself, nor can it be explained simply by the presence of generative AI. Rather, it emerges from the interaction of language-resource inequalities, opaque recommendation systems, rapid circulation in closed and semi-closed networks, contextual manipulation, weak provenance signals, and uneven access to credible verification. The paper also emphasizes the dual role of AI: the same technologies that can scale deceptive content can strengthen local-language fact-checking, transcription, translation, and public-interest communication. The study contributes a context-sensitive framework for understanding vernacular disinformation in

India and offers implications for policymakers, digital platforms, news organizations, fact-checkers, AI developers, and citizens. As a conceptual study, its propositions require empirical validation across languages, regions, platforms, and political contexts.

Keywords: Artificial intelligence, disinformation, misinformation, vernacular media, regional languages, digital public sphere, and multilingual fact-checking.

1. Introduction

India's digital transformation has expanded the scale and diversity of public communication. Affordable smartphones, low-cost data, social networking platforms, short-video applications, messaging services, and digital news outlets have enabled citizens to access and circulate information beyond the traditional boundaries of print and broadcast media. A significant feature of this transformation is the increasing importance of regional-language and Indic-language communication. The Internet in India Report 2024, prepared by the Internet and Mobile Association of India (IAMAI) and Kantar, estimated 886 million active internet users in 2024 and reported that 98 per cent had accessed content in Indic languages. It also found that 57 per cent of urban internet users preferred accessing internet content in Indic languages (IAMAI & Kantar, 2025). These figures demonstrate that vernacular communication is not a marginal feature of India's internet; it is central to its contemporary growth.

The expansion of vernacular digital communication has important democratic implications. Citizens can participate in public debate in languages with which they are familiar, access information about government services, discuss local concerns, and engage with news that reflects regional experiences. Regional-language media also provide visibility to issues that may receive limited attention in national English-language discourse. However, the expansion of access does not automatically guarantee an equitable or reliable information environment. The same infrastructures that facilitate participation can also accelerate the circulation of rumours, manipulated images, fabricated quotations, misleading translations, communal narratives, and politically motivated claims.

Artificial intelligence has intensified this dual transformation. Natural language processing, machine translation, speech recognition, synthetic voice generation, image generation, and large language models can lower the cost of creating and adapting content across languages. They can also help local newsrooms translate reports, transcribe interviews, identify emerging claims, and produce accessible public information. Yet AI systems may generate inaccurate translations, misrepresent cultural references, reproduce stereotypes, or produce plausible but unsupported claims. Generative systems can further facilitate the production of synthetic audio, images, and video that appear credible to audiences unfamiliar with the source material or the conventions of digital manipulation.

The challenge is not simply that AI makes false content possible. Misleading information existed long before generative AI, and its circulation depends on political incentives, social trust, media institutions, platform architecture, and audience practices. Research by Garimella and Chauchard (2024), based on approximately two million WhatsApp messages collected in Uttar Pradesh, found that fewer than two dozen of 1,858 viral messages in their examined sample contained generative-AI-created content. Their finding cautions against treating generative AI as the dominant explanation for all contemporary political misinformation. It nevertheless demonstrates the need to examine how emerging technologies interact with established communication networks and political narratives.

The vernacular context introduces specific challenges. India's language landscape includes multiple scripts, dialects, registers, transliteration practices, and forms of code-mixing. A claim may originate in one language, be translated into another, and circulate as an image, voice note, short video, or text message. Translation can alter nuance, attribution, tone, or the meaning of culturally specific terms. Automated moderation and fact-checking systems may perform unevenly across languages, particularly where high-quality training data, domain-specific corpora, and linguistic expertise are limited. A content moderation system that works adequately in standardized Hindi or English may not interpret regional idioms, dialectal speech, sarcasm, or local political references with equal reliability.

The central problem addressed in this paper is therefore the vulnerability of India's vernacular digital public sphere under conditions of rapid AI development and platform-mediated communication. The term 'vulnerability' is used not to characterize vernacular-language users as inherently gullible, but to describe structural conditions that may make the production, circulation, detection, and correction of disinformation more difficult. These conditions include unequal language resources, limited local verification infrastructure, platform opacity, social and political polarization, and gaps between the speed of circulation and the speed of verification.

This paper has following objectives:

1. To examine how artificial intelligence is transforming the production and circulation of information within India’s vernacular digital public sphere, while identifying the linguistic, technological, institutional, and socio-cultural factors that contribute to its vulnerability to disinformation.
2. To develop an integrated conceptual framework, grounded in relevant theoretical perspectives, that explains these vulnerabilities and proposes directions for future research, alongside practical and policy measures to strengthen the integrity of India’s multilingual digital public sphere.

The paper is conceptual in nature. It synthesizes established scholarship and selected contemporary reports and research on Indian digital communication, AI-generated content, and multilingual fact-checking. It does not claim to measure the prevalence of disinformation in all Indian languages or to establish causal effects through original field data.

2. Literature Review

2.1 Evolution of India's Vernacular Digital Public Sphere

The concept of the public sphere is commonly associated with Habermas (1989), who described a domain of social life in which citizens engage in discussion about matters of common concern. Subsequent scholarship has questioned the assumption that a single, unified public sphere adequately represents societies marked by social, cultural, and linguistic inequalities. Fraser (1990) argued that subordinated groups may develop alternative or counter-publics in which they articulate experiences and interests that are marginalized in dominant forums. This perspective is relevant to India, where regional-language media and digital communities provide spaces for local debate, cultural expression, and political participation.

The vernacular digital public sphere is not a single platform or homogeneous audience. It consists of interconnected but distinct communication spaces shaped by language, region, caste, class, religion, gender, age, occupation, and access to technology. Local news portals, regional television channels, community pages, creator-led video channels, messaging groups, and short-video networks all contribute

to this ecology. These spaces can broaden participation and challenge metropolitan or English-language dominance, but they may also reproduce existing inequalities and become sites of targeted manipulation.
Digital platforms have altered the relationship between publishers, audiences, and intermediaries. News is increasingly encountered through recommendation feeds, search results, creator channels, messaging groups, and algorithmically curated video streams rather than solely through direct visits to established news organizations. Gillespie (2018) describes platforms as important institutional actors that shape visibility, moderation, and public communication while often presenting themselves as neutral intermediaries. Their ranking and recommendation systems influence which content receives attention, how quickly it travels, and which sources become visible.
The Reuters Institute Digital News Report 2025 notes the importance of social platforms and messaging services in Indian news consumption, while also reporting that more than half of its Indian survey respondents identified WhatsApp as the channel carrying the greatest threat in relation to false and misleading information. The report explicitly cautions that its India sample is primarily English-speaking, which limits the extent to which it can represent the full range of vernacular audiences (Reuters Institute, 2025). This methodological limitation is itself important: research about India's digital public sphere must avoid treating English-speaking respondents as a proxy for the country's multilingual population.

2.2 Artificial Intelligence and the Transformation of Vernacular Communication

AI technologies are increasingly embedded in the production and distribution of digital content. Machine translation and transliteration can move information between languages and scripts; automatic speech recognition can convert voice messages into text; text-to-speech systems can make written material accessible to listeners; and generative AI can produce summaries, captions, images, audio, and video. These capabilities can reduce language barriers and support participation by people who have historically been underserved by English-dominant digital systems.
India's BHASHINI initiative illustrates the public-interest potential of multilingual AI. Launched under the National Language Translation Mission, BHASHINI aims to enable access to digital services in Indian languages through translation and voice-based technologies (Government of India, n.d.). Government reporting describes its role in developing language resources and AI services across Indian languages. Such initiatives can contribute to inclusion, but their existence does not by itself ensure that all languages, dialects, accents, and communicative contexts receive equal technological support.
The same AI capabilities can also be used to localize misleading content. A fabricated claim can be translated into multiple languages, rewritten in locally familiar idioms, or converted into speech and video. Synthetic voice technologies may make a message appear to come from a recognizable public figure or community authority. Image and video generation may create persuasive visual material that is detached from the event or context it purports to depict. Large language models can produce fluent text that resembles news reporting, official communication, or expert explanation, even when the underlying information is inaccurate.
However, it is necessary to distinguish AI-generated content from disinformation. AI-generated content may be accurate, creative, satirical, or clearly fictional. Disinformation generally refers to false or misleading information produced or disseminated with an intention to deceive or cause harm, whereas misinformation may circulate without such intent (Wardle & Derakhshan, 2017). The technology used to

produce a message does not determine its truth status or intent. Research must therefore examine content, provenance, context, circulation, and effects rather than assume that synthetic content is necessarily deceptive.

2.3 Disinformation, Platform Amplification and Networked Circulation

Research on information disorder emphasizes that misleading content spreads through a combination of producers, messages, interpreters, and distribution systems. Wardle and Derakhshan (2017) distinguish misinformation, disinformation, and malinformation, drawing attention to the different relationships between truth, intent, and harm. This framework helps avoid the loose use of 'fake news' as a label for all disputed or unwanted information.

Platform architecture influences the visibility and circulation of such material. Recommendation systems optimize engagement-related signals and can create incentives for emotionally salient, sensational, or identity-confirming content to attract attention. Tufekci (2017) argues that algorithmic systems shape public attention in ways that are not always visible to users. In the Indian context, messaging applications add another layer: content may circulate through private or semi-private groups where public monitoring is difficult and where trust in the sender can substitute for verification of the original source.

Research on WhatsApp misinformation in India has documented the role of forwarded images, memes, and political messages. Garimella and Chauchard (2024) offer an important corrective to exaggerated claims about AI-generated political content: in their Uttar Pradesh sample, generative AI appeared in only a small proportion of the viral messages examined. The result suggests that familiar forms of misinformation, strategic distribution, and social networks remain significant even as generative technologies develop. The implication for this paper is that AI should be understood as an enabling and potentially intensifying factor within a wider information ecosystem, not as a self-sufficient cause.

2.4 Language Inequality, Low-Resource Languages and Automated Systems

AI systems depend on data, computational resources, evaluation methods, and human expertise. Languages with large digital corpora and extensive commercial investment are more likely to receive mature language models and better evaluation coverage. Languages with fewer digitized texts, limited speech datasets, underrepresented dialects, or complex orthographic practices may receive less reliable performance. These differences are often described through the language-resource hierarchy, although the classification of a language as 'low-resource' depends on the task and the available data.

In multilingual India, language inequality is not limited to the distinction between English and scheduled Indian languages. Standardized written forms may be better represented than regional dialects, oral registers, code-mixed speech, and community-specific expressions. A translation system may produce grammatically fluent output while distorting honorifics, caste references, idiomatic expressions, or political terminology. A speech-recognition system may mis-transcribe names, places, and locally salient terms. These errors can become consequential when the output is republished as a news report or circulated as an authoritative translation.

UNESCO (2023) emphasizes the need to protect linguistic and cultural diversity in the deployment of generative AI, including through more inclusive language data and safeguards against the marginalization of minority and indigenous languages. This recommendation has a direct relevance to

India's public sphere: linguistic inclusion requires not only access to AI tools but also reliable performance, local evaluation, transparent limitations, and meaningful participation by language communities.

2.5 Multilingual Fact-Checking and Verification Capacity

Fact-checking organizations perform an important corrective function by investigating claims, tracing original sources, checking images and videos, consulting domain experts, and publishing evidence-based explanations. In multilingual environments, however, verification requires more than literal translation. Fact-checkers must understand local political histories, cultural references, dialects, and the context in which a claim is circulating. They may also need to identify whether a piece of content has been altered during translation or recontextualized after crossing language communities.

Roy, Laskar, and Reyaz (2025), in their study of fact-checkers during India's 2024 general election, describe fact-checkers as 'retroactive gatekeepers' who intervene after misleading material has already circulated. Their analysis highlights the linguistic and infrastructural demands placed on fact-checking organizations, including the need to translate claims, work across vernacular contexts, and engage audiences through participatory mechanisms. Shahi and Hale (2025), examining WhatsApp tipline submissions across English, Hindi, and Telugu during the 2021 assembly elections, also demonstrate the value of studying claims and verification processes across languages rather than treating language as a superficial variable.

The speed asymmetry between circulation and verification is a recurring problem. A misleading message may be forwarded rapidly through trusted networks, while verification requires source tracing, translation, expert consultation, and careful publication. Corrections may therefore arrive after the original claim has reached a much larger audience. The effectiveness of correction also depends on whether it is delivered in the language and format in which the original claim was encountered.

2.6 Research Gap

Existing scholarship provides important insights into the public sphere, platform governance, information disorder, generative AI, and Indian fact-checking practices. Yet these research streams are often examined separately. There remains a need for an integrated conceptual account of how AI capability, language-resource inequality, platform distribution, socio-cultural context, and verification infrastructure jointly shape vulnerability in India's vernacular digital public sphere. In particular, the literature needs to move beyond a technology-centred narrative in which generative AI is treated as the sole driver of disinformation. This paper addresses that gap by proposing a framework that places AI within a broader multilingual sociotechnical ecosystem.

3. Theoretical Foundations

3.1 Public Sphere Theory and Counterpublics

Public Sphere Theory provides a foundation for understanding digital communication as a space in which citizens encounter information, debate shared concerns, and form public opinion (Habermas, 1989). Fraser's (1990) critique draws attention to the unequal conditions under which different groups enter public debate and to the importance of subaltern counterpublics. Applied to India, this perspective

suggests that vernacular digital spaces can expand democratic participation by enabling communities to communicate in familiar languages and articulate locally grounded concerns.

At the same time, the democratic value of a public sphere depends on more than access. Participants require opportunities to encounter reliable information, challenge claims, and receive meaningful responses. Where language-specific verification is weak or content visibility is shaped by opaque systems, expanded participation may coexist with unequal informational protection. Public Sphere Theory therefore explains why vernacular disinformation is not merely a content-quality issue but a question of democratic inclusion and communicative equality.

3.2 Information Disorder Framework

The information disorder framework distinguishes misinformation, disinformation, and malinformation by considering the truthfulness of content, the intent behind its circulation, and the nature of the harm involved (Wardle & Derakhshan, 2017). This distinction is particularly useful in AI-mediated communication because synthetic media, translated text, and automated summaries may be accurate, inaccurate, misleading, or intentionally deceptive. The framework encourages researchers to examine the full lifecycle of a claim: creation, manipulation, distribution, interpretation, and correction.

3.3 Networked Gatekeeping and Platformization

Traditional gatekeeping theory focused on the selection and filtering of news by editors and media institutions. Digital communication distributes gatekeeping functions across journalists, influencers, platform algorithms, moderators, group administrators, fact-checkers, and ordinary users. Gillespie (2018) demonstrates that platforms exercise governance through decisions about ranking, moderation, and visibility. In vernacular settings, automated moderation and recommendation systems may have uneven linguistic competence, while local influencers and group administrators may become influential informal gatekeepers.

Networked gatekeeping explains how AI-generated or AI-translated content may gain visibility without passing through conventional editorial checks. It also helps explain why correction requires coordination among platforms, news organizations, community actors, and fact-checkers rather than reliance on a single institutional gatekeeper.

3.4 Sociotechnical Systems Perspective

A sociotechnical perspective treats technology and society as mutually shaping rather than independent domains. AI systems are designed using particular datasets, languages, commercial priorities, and assumptions about user behaviour. Their outputs are interpreted within social contexts shaped by political identities, trust relationships, historical grievances, and local knowledge. Accordingly, the vulnerability of a vernacular public sphere cannot be attributed solely to an algorithm or to audience behaviour. It emerges from interactions among technological design, institutional arrangements, social practices, and power relations.

3.5 Media Literacy and Contextual Resilience

Media and information literacy emphasizes the ability to access, evaluate, interpret, create, and share information responsibly. In multilingual settings, literacy must include the ability to evaluate translated

content, recognize missing context, assess source credibility, and understand the limits of synthetic media detection. It should not be reduced to teaching users to identify visual artifacts in deepfakes, since many forms of disinformation use authentic material with misleading captions or context. Contextual resilience therefore depends on critical habits, trusted local information sources, and accessible verification mechanisms.

4. Development of the Conceptual Framework

Based on the literature and theoretical foundations, this paper proposes the AI–Vernacular Information Integrity Framework (AVIIF). The framework conceptualizes vulnerability as the outcome of interactions among six components: (1) AI-enabled content capabilities; (2) vernacular language and content ecology; (3) platform distribution and amplification; (4) socio-cultural and audience conditions; (5) verification and media-literacy capacity; and (6) institutional and platform governance. The framework treats information integrity and democratic participation as the principal outcomes, while recognizing that the same AI infrastructure can support both inclusion and manipulation.

4.1 AI-Enabled Content Capabilities

The first component includes generative text, image, audio and video systems; machine translation; speech recognition; voice synthesis; automated subtitling; transliteration; and content-recommendation tools. These capabilities reduce the cost and time required to create and adapt content. They may also create risks through hallucinated claims, inaccurate translation, synthetic impersonation, fabricated visual evidence, and the mass production of superficially plausible messages. The framework does not assume that AI-generated content is inherently deceptive; it focuses on the potential for these capabilities to be used in misleading or manipulative ways.

4.2 Vernacular Language and Content Ecology

The second component concerns the linguistic environment in which content is produced and interpreted. Relevant characteristics include the availability and quality of language datasets, dialect diversity, code-mixing, transliteration, oral communication, local idioms, script variation, and the presence of trusted regional media. Content may move between languages and formats, with meaning altered through translation, summarization, captioning, or recontextualization. Unequal language-resource availability can increase the likelihood that automated systems misinterpret content or provide weaker moderation and verification support.

4.3 Platform Distribution and Amplification

The third component includes recommendation algorithms, engagement-based ranking, forwarding functions, group networks, creator ecosystems, search systems, and content moderation. Platform design influences the speed, reach, and visibility of information. Closed and semi-closed messaging groups may make public observation difficult, while recommendation feeds can expose users to content selected partly through behavioural signals. The framework proposes that platform amplification can intensify the effects of AI-enabled content capabilities when sensational or identity-relevant material is rewarded with attention and circulation.

4.4 Socio-Cultural and Audience Conditions

The fourth component includes political polarization, community trust, local histories, identity narratives, religious and caste relations, economic uncertainty, media habits, and varying levels of digital and media literacy. These factors influence whether a message is believed, questioned, forwarded, or rejected. Familiar language and recognizable cultural symbols may make a message feel credible, but this does not mean vernacular audiences are inherently more susceptible. The framework treats susceptibility as contextual and relational, shaped by source credibility, social trust, prior beliefs, and the availability of corrective information.

4.5 Verification and Media-Literacy Capacity

The fifth component includes regional-language fact-checking organizations, newsroom verification desks, public-interest databases, tip lines, local experts, community reporting networks, media-literacy initiatives, and accessible correction formats. Verification capacity is strengthened when organizations can work across languages and formats, preserve provenance information, and publish corrections where the original claim circulated. AI can support this component through transcription, translation, clustering of similar claims, image and video analysis, and multilingual search. Human review remains essential for contextual interpretation and accountability.

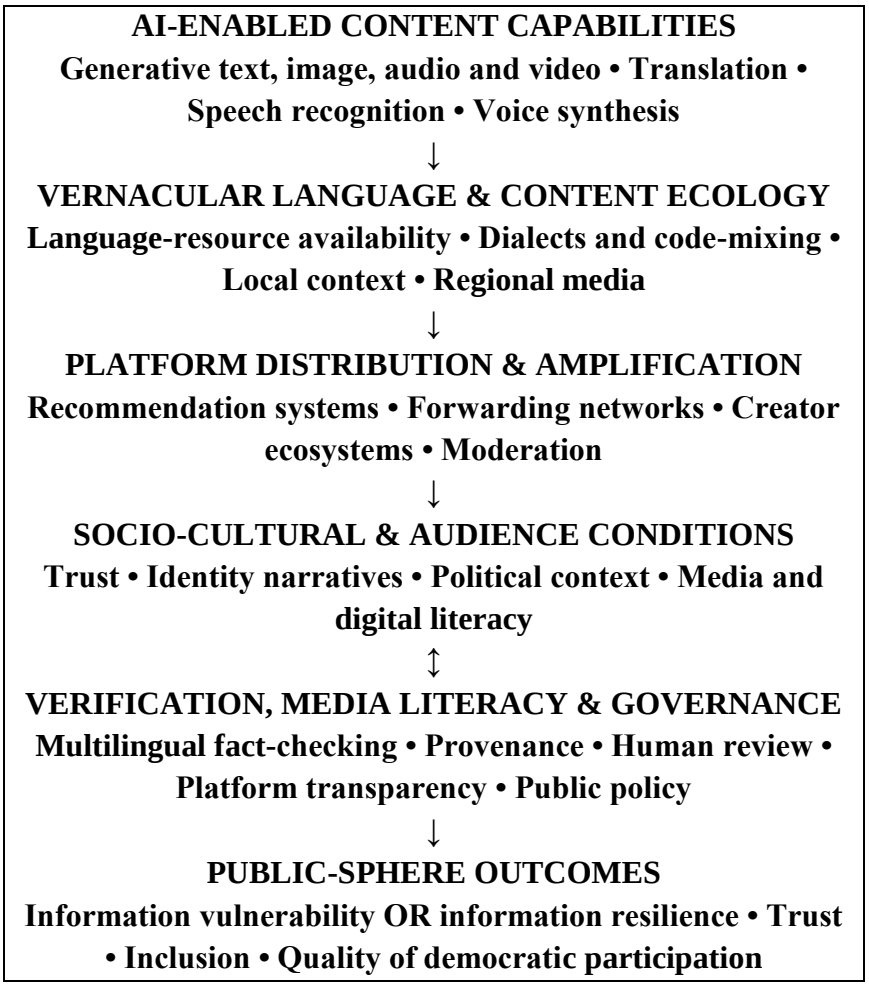
4.6 Institutional and Platform Governance

The sixth component includes platform transparency, language-specific safety evaluation, content provenance, data access for independent research, appeal mechanisms, privacy safeguards, election integrity protocols, and coordination among public institutions, media organizations, civil society, and technology developers. Governance should address not only harmful content but also unequal performance across languages. A system that removes content accurately in English but inconsistently in other languages may create unequal levels of protection and unequal risks of over-removal.

4.7 Outcomes of the Framework

The framework identifies two broad sets of outcomes. The first is vulnerability-related: increased circulation of misleading content, reduced trust in credible sources, delayed correction, exclusion of underrepresented language communities from effective verification, and distortion of public debate. The second is resilience-related: improved access to reliable local-language information, stronger verification networks, greater transparency, more equitable AI performance, and informed participation in democratic discussion. These outcomes are not mutually exclusive; the same digital ecosystem may display both vulnerability and resilience.

Figure 1. Proposed AI–Vernacular Information Integrity Framework



Source: Authors' conceptualization.

The framework is dynamic rather than strictly linear. AI capabilities shape the forms in which content can be created and translated; vernacular contexts influence the accuracy and meaning of that content; platforms shape its visibility and circulation; and audience conditions influence interpretation and sharing. Verification capacity and governance operate across the entire process by limiting harmful circulation and strengthening the conditions for correction. Feedback loops are expected: public controversies may alter platform policies, while the success or failure of verification efforts may affect audience trust and future sharing behaviour.

4.8 Conceptual Propositions

P1: Greater AI-enabled content-generation and localization capability increases the potential scale and speed of vernacular content production, including the production of misleading content.

P2: Unequal language-resource availability and weak dialect-sensitive evaluation increase the likelihood of inaccurate translation, transcription, classification, and moderation in underrepresented Indian languages.

P3: The relationship between AI-enabled content capability and disinformation exposure is strengthened by platform distribution mechanisms that reward rapid engagement and facilitate repeated forwarding.

P4: Socio-cultural context, including community trust, political polarization, and local identity narratives, influences the interpretation and circulation of misleading vernacular content.
P5: Strong multilingual verification capacity and accessible media-literacy support reduce the persistence and perceived credibility of disinformation in vernacular digital spaces.
P6: Platform transparency, language-specific safety testing, and accountable governance strengthen the relationship between AI deployment and information integrity.
P7: AI-enabled tools used by fact-checkers and local newsrooms improve verification capacity when their outputs are subject to linguistic expertise, source checking, and human editorial oversight.
P8: The combined effect of language inclusion, platform accountability, and community-based verification positively influences trust and meaningful participation in India's vernacular digital public sphere.

5. Discussion

The proposed framework advances the understanding of vernacular disinformation by locating AI within a wider sociotechnical environment. A technology-centred explanation may suggest that the main threat arises from the ability to generate convincing synthetic media. Yet the available evidence indicates that the relationship is more complex. Garimella and Chauchard's (2024) analysis found a relatively small share of generative-AI-created items among the viral messages they examined in Uttar Pradesh. This finding does not establish that AI-generated disinformation is insignificant in every context; rather, it cautions against equating the broader problem of disinformation with the prevalence of synthetic media. Manipulated context, recycled images, false captions, impersonation, and strategic distribution can remain influential without sophisticated generative tools.

A second contribution is the recognition that linguistic inequality is a structural dimension of information integrity. When AI systems perform unevenly across languages, the consequences extend beyond inconvenience. Inaccurate transcription may change the identity of a speaker or the meaning of a statement; poor translation may remove qualifications or intensify a claim; and weak moderation may allow harmful content to circulate or cause legitimate speech to be incorrectly removed. These risks require language-specific evaluation and participation by speakers who understand local usage. The framework therefore treats linguistic inclusion as a condition of safety and democratic equality, not merely as a product feature.

Third, the framework emphasizes the importance of platform-mediated distribution. The speed and scale of circulation are shaped by forwarding features, recommender systems, creator networks, and social trust. Content that appears in a familiar language or arrives through a trusted contact may be accepted without checking its origin. This does not imply that users lack critical capacity; rather, it reflects the ordinary role of social trust in communication. Platform interventions should therefore address the design conditions that facilitate repeated circulation, while preserving legitimate political discussion and community expression.

Fourth, verification must be understood as a multilingual public infrastructure. Fact-checking is not simply the publication of a correction in English followed by automated translation. Effective verification requires local-language expertise, contextual knowledge, source tracing, and distribution through channels that reach the affected audience. Roy et al. (2025) show how Indian fact-checkers operated as retroactive gatekeepers during the 2024 election, including through linguistic translation and

audience participation. This suggests that fact-checking capacity should be strengthened before high-risk periods rather than activated only after misleading claims become viral.

Fifth, AI has a dual role in the information ecosystem. The same technologies that can translate and adapt deceptive claims can help identify recurring narratives, transcribe voice notes, compare claims across languages, search public records, and generate accessible corrections. However, automated tools should not be treated as final arbiters of truth. AI outputs may reproduce errors, miss sarcasm, or misread local context. Human editorial judgement, transparent sourcing, and appeal mechanisms remain essential.

Finally, the framework links information integrity to democratic inclusion. A vernacular public sphere is not resilient merely because more citizens can access content in their preferred languages. It is resilient when citizens can also assess credibility, encounter plural perspectives, challenge misleading claims, and access timely corrections. This requires collaboration among platforms, government institutions, regional newsrooms, civil society, academic researchers, and language communities.

6. Theoretical Contributions

The paper makes several theoretical contributions to scholarship on AI, media, and digital democracy. First, it extends Public Sphere Theory by positioning linguistic access and verification capacity as interdependent conditions of meaningful participation. Vernacular digital spaces can function as counterpublics that expand the range of voices in public debate, but their democratic potential is constrained when information integrity protections are distributed unequally.

Second, the paper integrates the information disorder framework with networked gatekeeping. It shows how the creation, translation, ranking, forwarding, interpretation, and correction of content are distributed across AI systems, platforms, news organizations, fact-checkers, influencers, and ordinary users. This approach moves beyond a simple producer-message-receiver model and recognizes that gatekeeping is performed throughout the circulation process.

Third, the paper contributes a language-sensitive sociotechnical perspective. It conceptualizes language-resource inequality as a structural condition that influences the performance of AI systems and the capacity of institutions to detect and correct misleading content. This expands the analysis of AI bias beyond demographic fairness to include dialect, script, register, code-mixing, and local contextual competence.

Fourth, the proposed AVIIF framework integrates vulnerability and resilience in a single model. Rather than treating AI only as a threat, it recognizes the possibility of AI-enabled verification and public-interest communication. This dual orientation offers a more balanced theoretical foundation for examining multilingual digital ecosystems.

Finally, the conceptual propositions provide a basis for empirical research. They can be operationalized through measures of language-specific AI performance, platform exposure, verification latency, trust, correction reach, and perceived information integrity. The framework can therefore guide comparative studies across languages, states, platforms, and election and non-election periods.

7. Practical and Policy Implications

7.1 Implications for Policymakers and Public Institutions

Public institutions should treat multilingual information integrity as part of digital inclusion policy. Investments in language technology should be accompanied by independent evaluation of accuracy across dialects, accents, scripts, and code-mixed communication. Public-interest language datasets should be developed with consent, appropriate privacy safeguards, and participation from linguistic communities. Government communication should be made available in accessible local-language formats, with clear source attribution and correction procedures.

Policy frameworks should encourage transparency and accountability from platforms while protecting freedom of expression and legitimate political debate. Requirements for risk assessment, meaningful reporting, independent research access, and accessible appeals should be designed to avoid arbitrary censorship. Election-related information integrity measures should be developed in collaboration with independent fact-checkers, regional media, researchers, and civil society rather than relying exclusively on centralized content removal.

7.2 Implications for Digital Platforms

Platforms should evaluate automated moderation, ranking, and recommendation systems separately across major languages and relevant dialects. Aggregate performance scores can conceal severe gaps in specific language communities. Platforms should provide clearer information about the origin and provenance of synthetic or materially altered media where technically feasible, and should make reporting and appeal mechanisms accessible in users' preferred languages.

Forwarding friction, contextual labels, provenance indicators, and links to reliable sources may help users pause before sharing uncertain claims. Such interventions should be proportionate and should not presume that all political or controversial content is false. Platforms should also support independent researchers with privacy-preserving access to data necessary to study the circulation of public-interest information.

7.3 Implications for News Organizations and Regional Media

Regional newsrooms should develop verification workflows that cover text, images, audio, video, and translated material. Journalists require training in source tracing, synthetic-media awareness, reverse image search, metadata limitations, and the responsible use of AI-assisted transcription and translation. News organizations should disclose when AI has materially contributed to content production and ensure that a human editor checks claims, names, quotations, and context before publication.

Collaboration among regional newsrooms can reduce duplication and improve coverage. Shared verification desks, language-specific claim databases, and partnerships with universities can support smaller organizations that lack specialist resources. Corrections should be published in the same language and, where possible, the same format and distribution channels as the original misleading content.

7.4 Implications for Fact-Checking Organizations and Civil Society

Fact-checking organizations should strengthen local-language capacity through partnerships with community journalists, language experts, universities, and civil society groups. Tip lines and audience-

submission systems can help identify emerging claims, but organizations should communicate clearly about verification timelines and uncertainty. AI tools may assist with clustering similar claims, transcription, translation, and prioritization, provided that final assessments remain accountable to human researchers.

Civil society organizations can support media and information literacy through practical programmes that teach users to examine source identity, date, original context, corroboration, and evidence. Training should be locally designed and delivered in accessible formats, including audio and visual materials. It should avoid stigmatizing communities or portraying regional-language users as uniquely vulnerable.

7.5 Implications for AI Developers and Researchers

AI developers should prioritize inclusive data collection, transparent model documentation, dialect-sensitive evaluation, and clear communication of system limitations. Benchmarking should include code-mixed text, transliterated messages, voice notes, regional accents, and culturally specific references. Researchers should avoid reporting a single aggregate accuracy figure as evidence that a system performs reliably across India's linguistic diversity.

Academic researchers should develop comparative datasets and ethically governed methods for studying public and private communication. Research involving messaging groups must respect informed consent, privacy, and data minimization. Collaboration with native speakers and regional institutions is essential for annotation quality and contextual interpretation.

7.6 Implications for Citizens and Community Networks

Citizens can strengthen information resilience by checking the original source, date, location, and context of claims; comparing reports from credible sources; and being cautious about forwarded material that asks for urgent action or relies on emotional pressure. Community leaders, educators, libraries, and local journalists can act as trusted intermediaries by sharing verified information and explaining uncertainty without humiliating people who have encountered misleading content.

8. Conclusion

The expansion of regional-language internet use is transforming India's digital public sphere. It has widened access to information, enabled local expression, and created new opportunities for participation in public debate. Artificial intelligence can accelerate this process through translation, speech technologies, generative tools, and multilingual interfaces. Yet the same technological developments can be used to create, adapt, and circulate misleading content, while uneven language resources and platform systems may make detection and correction more difficult.

This conceptual paper has argued that the vulnerability of India's vernacular digital public sphere cannot be attributed to AI alone. It emerges through the interaction of AI-enabled content capabilities, language-resource inequalities, platform distribution, socio-cultural context, verification capacity, and institutional governance. The proposed AI–Vernacular Information Integrity Framework provides an integrated account of these relationships and advances eight conceptual propositions for future research.

The paper also cautions against overstating the prevalence of generative-AI-created political content. Available evidence from a large WhatsApp message sample in Uttar Pradesh found such content in a small proportion of viral items examined (Garimella & Chauchard, 2024). The wider disinformation

challenge includes older and less technologically complex practices, such as misleading captions, recycled images, false attribution, and strategic forwarding. AI should therefore be understood as a changing capability within a broader political and communicative ecosystem.

The central policy implication is that linguistic inclusion and information integrity must be advanced together. Multilingual AI infrastructure should be evaluated for fairness and accuracy; platforms should be accountable for language-specific safety performance; regional newsrooms and fact-checkers should receive institutional support; and citizens should have access to timely, comprehensible verification. AI can contribute to a more inclusive public sphere only when its deployment is transparent, context-sensitive, and subject to human oversight.

Ultimately, the strength of India's vernacular digital public sphere will depend not only on how many citizens can communicate online in their preferred languages, but also on whether they can participate in an environment where claims can be examined, corrections can be accessed, and diverse voices can be heard without unequal exposure to manipulation. The proposed framework offers a foundation for understanding and strengthening that environment.

9. Limitations and Future Research Directions

9.1 Limitations

First, the study is conceptual and does not include original empirical data. The relationships proposed in the framework are derived from theoretical literature and selected secondary research rather than statistically tested. Second, India's linguistic landscape is highly diverse, and the paper cannot provide equal depth of analysis for every language, dialect, state, and platform. Third, the paper focuses on the public-interest implications of AI and disinformation but does not undertake a technical audit of specific language models, moderation systems, or recommendation algorithms. Fourth, the availability of research on private messaging environments is limited by privacy, access, and ethical constraints. Fifth, AI capabilities and platform policies evolve rapidly; therefore, the framework should be treated as an evolving model rather than a definitive account of a stable technological environment.

9.2 Future Research Directions

Future studies should empirically test the proposed relationships through comparative research across languages such as Hindi, Bengali, Tamil, Telugu, Marathi, Malayalam, Kannada, Gujarati, Punjabi, Odia, Assamese, and selected tribal and minority languages. Researchers may examine differences in AI translation quality, speech recognition, content moderation, and fact-checking performance across dialects and code-mixed communication.

Quantitative studies could develop measures of AI-enabled content capability, language-resource inequality, platform amplification, verification capacity, audience trust, and information resilience. Structural equation modelling or multilevel modelling may be used to examine relationships among these constructs, provided that measurement instruments are validated for the relevant linguistic contexts.

Mixed-method research combining content analysis, interviews with journalists and fact-checkers, audience studies, platform audits, and digital ethnography would offer a more comprehensive understanding of how disinformation circulates and is corrected. Longitudinal studies could compare

election and non-election periods and examine whether synthetic media changes the volume, form, or perceived credibility of misleading content.

Future research should also investigate the role of AI in strengthening verification. Studies may assess whether automated transcription, translation, claim clustering, and image analysis improve verification speed without increasing false positives. Research should examine the conditions under which AI-assisted fact-checking is reliable, explainable, and trusted by local-language audiences.

Finally, comparative research across India and other multilingual democracies could help identify common patterns and context-specific differences. Such work would strengthen theoretical understanding of language inequality, platform governance, and information integrity in digitally diverse societies.

REFERENCES:

1. Fraser, N. (1990). Rethinking the public sphere: A contribution to the critique of actually existing democracy. Social Text, 25/26, 56–80.
2. Garimella, K., & Chauchard, S. (2024). How prevalent is AI misinformation? What our studies in India show so far. Nature, 630, 32–34. https://doi.org/10.1038/d41586-024-01588-2
3. Gillespie, T. (2018). Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.
4. Government of India. (n.d.). BHASHINI: National Language Translation Mission. Digital India BHASHINI Division. https://bhashini.gov.in/about-bhashini
5. Habermas, J. (1989). The structural transformation of the public sphere: An inquiry into a category of bourgeois society (T. Burger, Trans.). MIT Press. (Original work published 1962)
6. IAMAI & Kantar. (2025). Internet in India 2024. Internet and Mobile Association of India. https://www.iamai.in/research/internet-india-2024-kantariamai-report
7. Reuters Institute for the Study of Journalism. (2025). Digital News Report 2025: India. University of Oxford. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2025/india
8. Roy, E., Laskar, K. A., & Reyaz, M. (2025). Retroactive gatekeepers: Fact-checkers, electoral disinformation, and audience engagement in India. Social Media + Society. https://doi.org/10.1177/27523543251395190
9. Shahi, G. K., & Hale, S. A. (2025). WhatsApp tiplines and multilingual claims in the 2021 Indian assembly elections. Online Social Networks and Media.
10. Tufekci, Z. (2017). Twitter and tear gas: The power and fragility of networked protest. Yale University Press.
11. UNESCO. (2023). Guidance for generative AI in education and research. United Nations Educational, Scientific and Cultural Organization. https://unesdoc.unesco.org/ark:/48223/pf0000386693
12. Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. Council of Europe.