Jus Corpus Law Journal

Open Access Law Journal – Copyright © 2026 – ISSN 2582-7820
Editor-in-Chief – Prof. (Dr.) Rhishikesh Dave; Publisher – Ayush Pandey

AI Watermarking and Legal Implications of Its Removal

Vaibhavi Ambasta[a]

[a]National Law Institute University, Bhopal, India

Received 22 December 2025; *Accepted* 23 January 2026; *Published* 28 January 2026

With the rising proliferation of generative artificial intelligence, there is a significant increase in the blurred boundaries between authentic and synthetic digital content. It intensifies the risk associated with misinformation, impersonation and identity fraud. In response, regulators are increasingly relying on technical provenance mechanisms such as watermarking and metadata labelling to identify AI-generated content and preserve accountability. However, the ease with which AI watermarks can be deliberately removed or degraded fundamentally undermines these safeguards. This article examines the technical principles of AI watermarking, the vulnerabilities of their implementation and the growing practice of watermark circumvention. Focusing on India, it analyses how the deliberate removal of provenance markers interacts with legal frameworks addressing fraud, identity theft, intermediary liability, electoral integrity and digital evidence under Bharatiya Nyaya Sanhita, the Information Technology Act and recent MeitY and Election Commission advisories. The article contends that watermark stripping is not a neutral form of content modification but a regulatory circumvention that obstructs transparency-driven AI governance. It, therefore, argues for a shift in Indian regulatory strategy from a purely downstream harm-based approach to one that recognises deliberate watermark circumvention as a distinct legal and policy concern.

Keywords: *watermarks, AI-generated content, generative artificial intelligence, deepfakes, synthetic media.*

INTRODUCTION

The rapid advancement in generative artificial intelligence (GenAI) has transformed how content is created, be it text, images, audio or video. AI-generated content (AIGC) increasingly resembles human-produced material both in quality and in realism. This has led to challenging the traditional notion of authenticity and accountability. While these tools offer tremendous creative potential, they create new risks such as the spreading of misinformation and identity theft. A recent viral example illustrates the severity of the problem, where a video demonstrated how the new 'Nano Banana Pro', which is a consumer-level AI model, could allegedly generate highly realistic versions of Aadhar and PAN documents.[1] Although such illustrations oversimplify the security features of real identity documents, the fact that an AI model can visually approximate government IDs raises serious concerns related to identity fraud and the erosion of trust in routine verification processes.

An emerging solution to mitigate this risk is AI watermarking, a technique where a detectable signal is embedded in AIGC to indicate that it has a synthetic origin. Watermarking has become both a technical and regulatory focus. In the European Union, the recently adopted AI Act requires certain categories of AI systems to ensure that synthetic audio-visual and textual content is clearly labelled as artificially generated.[2] Similarly, in the United States, federal policy initiatives like the Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence (2023) have encouraged the development and deployment of robust watermarking and content authentication mechanisms, with agencies such as NIST tasked with issuing a technical standard for identifying AIGC.

In India, these parallel concerns are reflected in recent regulatory developments such as the MeitY advisories, the recent draft amendment to the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules and proposals under the forthcoming Digital India Act.[3] These together collectively place obligations on platforms to label AIGC and preserve metadata of the synthetic outputs. Despite the growing regulatory emphasis, watermarking cannot be considered foolproof. Techniques to remove, degrade, or bypass watermarks are now evolving rapidly, requiring minimal technical sophistication. This

[1] Aman Gupta, 'Nano Banana Pro is now generating fake Aadhaar and PAN cards: here's what we found' *Mint* (24 November 2025) <https://www.livemint.com/technology/tech-news/nano-banana-pro-is-now-generating-fake-aadhaar-and-pan-cards-heres-what-we-found-11764000416249.html> accessed 04 December 2025

[2] EU AI Act 2024, art 50

[3] 'Proposed Amendments to the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 in relation to Synthetically Generated Information' (*Ministry of Electronics and Information Technology*, 22 October 2025) <https://www.meity.gov.in/static/uploads/2025/10/8e40cdd134cd92dd783a37556428c370.p> accessed 04 December 2025

article examines the technical methods of AI watermarking, the vulnerabilities of using these methods and the emerging legal landscape governing the manipulation and circulation of AIGC.

UNDERSTANDING AI WATERMARKING

AI watermarking is a method of embedding a visible or hidden signature inside AIGC. The primary goal is the easy identification and distinction by humans and machines of AIGC from natural, human-created content.[4]

Watermarking methods can broadly be categorised into three types:

Visible Watermarks: These are symbols, labels or logos placed on top of images or videos. They are immediately recognisable, but their simplicity becomes a weakness. Basic editing, such as cropping, erasing or overlaying, can easily remove them.

Invisible Watermarks: These are signals embedded in the internal structure of the content, such as pixel values, frequency distribution or audio spectral features. These marks cannot be detected by humans, but machine algorithms can easily identify them. However, invisible watermarks are fragile. Repeated compression, filtering, noise addition or 'regeneration' via a secondary AI model can destroy the watermark, leaving the content visually or audibly indistinguishable from the original.

Statistical Watermarks: These rely on detectable patterns in the AI's output, such as preferred word choices in text or pixel distributions in images. These watermarks can be erased or obfuscated through paraphrasing, minor editing, or translation, rendering the content effectively untraceable.[5]

Watermarking plays a critical role for two primary reasons. First, it helps prevent individuals from passing off AIGC as their own intellectual property. Second, it protects against the dissemination of harmful or misleading information like deepfakes that could damage reputation or compromise social trust. Once watermarks are removed or bypassed, AIGC can cause significant personal, societal and institutional harm.

[4] Xuandong Zhao et al., 'SoK: Watermarking for AI-Generated Content' (2024) arXiv <https://doi.org/10.48550/arXiv.2411.18479> accessed 04 December 2025

[5] *Ibid*

WHY WATERMARK REMOVAL MATTERS

Watermark removal is not a narrow technical issue; it directly undermines safeguards meant to manage the risks of AIGC. As synthetic media is becoming increasingly realistic, watermarks serve as one of the few tools that allow the public to distinguish authentic materials from fabricated content. When these markers get stripped, multiple harms can occur across society, politics and individual lives.

One of the most visible risks is political manipulation. In September 2023, just two days before Slovakia's parliamentary elections, a deepfake audio clip was circulated on social media that showed a prominent candidate discussing alleged vote-rigging plans. This emerged during the legally mandated pre-election silence period; there was minimal opportunity to publicly debunk the clip before votes were cast. Fact-checkers could quickly identify signs of manipulation, but the audio had already been spread widely.[6] Similarly, during the 2024 Indian Lok Sabha elections, tens of millions of AI-generated robocalls and voice-cloned messages, which imitated politicians such as Rahul Gandhi, disseminated misleading information to voters across multiple regions and languages.[7] Watermarks, if applied and preserved, could have helped the voters in distinguishing AIGC from reality.

Watermark removal also fuels fraud and impersonation. Voice-cloned phone calls, manipulated video calls, and AI-generated documents can lead to financial loss or the revealing of sensitive personal information. The stakes are even higher with AI-generated identity documents. Cybersecurity assessments now warn that generative models, such as the Nano Banana Pro, can produce visually convincing replicas of Aadhaar cards, PAN cards, and other government IDs.[8] While these forgeries do not replicate backend verification systems such as biometrics or encrypted QR codes, they are highly effective in low-friction verification environments, including remote KYC processes or informal digital onboarding. Once the watermarks are removed, these AI-generated IDs appear virtually

[6] Mateusz Łabuz and Christopher Nehring, 'On the way to deep fake democracy? Deep fakes in election campaigns in 2023' (2024) 23 European Political Science <https://link.springer.com/article/10.1057/s41304-024-00482-9> accessed 04 December 2025

[7] Gulam Jeelani, 'How AI influenced Lok Sabha election 2024 campaign – from deepfake PM Rahul Gandhi to 50 million robocalls' *Mint* (14 April 2025) <https://www.livemint.com/technology/tech-news/deepfake-pm-rahul-gandhi-and-50-million-robocalls-ai-influence-on-lok-sabha-election-2024-campaign-revealed-11744561921605.html> accessed 04 December 2025

[8] 'Google's AI Tool Sparks Identity Fraud Fears in India' (*OECD AI Policy Observatory,* 25 November 2025) <https://oecd.ai/en/incidents/2025-11-25-423c> accessed 04 December 2025

indistinguishable from legitimate ones.[9] This drastically lowers the barrier for identity theft, financial fraud and impersonation. Without persistent watermark signals, proving that a document is AI-generated becomes exceedingly difficult, and the resulting harm spreads before any detection is possible.

Non-consensual deepfake pornography represents one of the most direct personal harms caused by AIGC. Watermarks serve as a critical signal that media is synthetic, enabling platforms to remove or label content quickly and victims to gather evidence for legal action. However, even if watermarks are present on the video, the content can nevertheless create severe psychological stress for the individual and damage their reputation. When watermarks are stripped, these videos or images can circulate widely on social media, messaging apps, and private channels.

Beyond pornography, manipulated images or AI-generated videos can create false narratives about individuals, such as attributing statements, actions, or endorsements that they never made, which can irreversibly harm one's professional and personal reputations. This underscores the need for AI companies to impose stricter restrictions on prompt inputs and outputs. Without a persistent watermark, victims are left with little proof that the content is synthetic, making legal redress, takedown requests, and accountability extremely difficult.

Watermark removal also poses significant risks to journalism and the integrity of legal evidence. Investigative reporters increasingly rely on digital media, such as videos, photographs, and audio recordings, to document events and substantiate claims. If these materials are AI-generated and stripped of watermarks, journalists may inadvertently report fabricated narratives, undermining public trust in their work.[10] Similarly, in legal or judicial contexts, digital evidence is often scrutinised for authenticity. Watermark-stripped AI content can complicate forensic verification, weaken evidentiary chains, and introduce reasonable doubt even in cases where malicious intent is clear. The inability to reliably distinguish synthetic content from authentic media creates challenges not only for courts and

---

[9] 'Internet Organised Crime Threat Assessment (IOCTA) 2024' (*Europol*, 2024) <https://www.europol.europa.eu/publication-events/main-reports/internet-organised-crime-threat-assessment-iocta-2024> accessed 04 December 2025

[10] 'AI and the Future of News' (*Reuters Institute*) <https://reutersinstitute.politics.ox.ac.uk/ai-journalism-future-news> accessed 04 December 2025

law enforcement but also for whistleblowers, human rights investigations, and academic research that relies on verified sources.

Finally, the erosion of watermark reliability contributes to broader public distrust. Humans are cognitively wired to pay more attention to negative, sensational, or emotionally charged content, and AI-generated material often exploits these tendencies to maximise engagement. As a result, fake or harmful content spreads far more rapidly than neutral or corrective information, even when it is clearly synthetic.[11] When people cannot confidently verify what they see or hear online, scepticism grows, and misinformation proliferates more easily. Societal institutions—from media outlets to electoral bodies—risk losing credibility, and disinformation campaigns gain a wider foothold. Watermark removal, therefore, represents not merely a technical vulnerability but a structural threat to information integrity, accountability, and social stability.

THE LEGAL STATUS OF STRIPPING AI WATERMARKS UNDER INDIAN LAW

Indian law does not yet contain an express statutory prohibition on the act of stripping AI watermarks. However, watermark removal increasingly intersects with multiple legal regimes governing fraud, identity theft, intermediary liability, and digital governance. As watermarking becomes central to regulatory efforts aimed at transparency and traceability of AI-generated content, its deliberate removal undermines not only individual safeguards but also the enforceability of public law obligations. Consequently, watermark stripping operates within a narrowing legal grey zone shaped by existing statutory provisions and recent regulatory directions.

Fraud, Cheating, and Misrepresentation under the BNS: Under Indian criminal law, deception that induces belief or reliance attracts liability irrespective of the technical means employed. The Bharatiya Nyaya Sanhita, 2023, criminalises cheating and deceptive conduct under Section 318, which addresses acts intended to dishonestly induce a person to deliver

[11] Paul Rozin and Edward B Royzman, 'Negativity Bias, Negativity Dominance, and Contagion' (2001) 5(4) Personality and Social Psychology Review 296 <https://doi.org/10.1207/S15327957PSPR0504_2> accessed 04 December 2025

property or act to their detriment.[12] Further, Section 319 specifically penalises cheating by personation, covering impersonation through false representation.[13]

When AI-generated content is presented as authentic after the deliberate removal of watermarks or labels that would otherwise signal its synthetic origin, such conduct may satisfy the elements of deception contemplated under these provisions. This concern has become particularly acute with the emergence of AI systems capable of generating hyper-realistic images and documents. Demonstrations of AI-generated Aadhaar and PAN cards illustrate how watermark removal can facilitate fraudulent misrepresentation. In this context, watermark stripping functions as a preparatory act that strengthens intent to deceive, thereby engaging the offence of cheating under the BNS.

Identity Theft, Personation, and Cyber Offences: Watermark stripping also enables identity-based offences regulated under the Information Technology Act, 2000. Section 66C penalises identity theft involving the fraudulent use of another person's electronic signature, password, or unique identification feature. Section 66D addresses cheating by personation using computer resources.[14] The removal of watermarks from AI-generated identity documents or synthetic profiles facilitates these offences by allowing fabricated content to circulate without provenance markers.

This risk has been acknowledged in regulatory responses. The CERT-In Advisory of November 2024 framed deepfakes primarily as threats to cybersecurity and financial integrity, recommending watermarking, source verification mechanisms, and adoption of provenance standards such as C2PA. In this regulatory framing, watermarking is treated as a preventive security measure. Its removal, therefore, weakens safeguards designed to mitigate identity theft, fraud, and impersonation in digital environments.

Intermediary Liability under the IT Act and IT Rules: The legality of watermark removal must also be assessed through the lens of intermediary responsibility. Section 79 of the IT Act provides intermediaries with conditional safe-harbour protection, contingent upon compliance with due diligence obligations.[15] These obligations are operationalised through

[12] Bharatiya Nyaya Sanhita 2023, s 318
[13] Bharatiya Nyaya Sanhita 2023, s 319
[14] Information Technology Act 2000, s 66C
[15] Information Technology Act 2000, s 79

the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, which require platforms to act against unlawful or harmful content upon actual knowledge or notification.

Recent regulatory interventions have strengthened these expectations. The MeitY Advisory of November 2023 directed intermediaries to remove deepfake content within thirty-six hours, explicitly warning that failure to do so could jeopardise safe-harbour protection. The subsequent March 2024 MeitY Advisory went further by requiring platforms to ensure that synthetic content carries persistent labels or embedded metadata capable of tracing the originator, accompanied by compliance reporting obligations. These directions implicitly treat watermarking and metadata embedding as components of intermediary due diligence.

In this evolving regulatory environment, systematic watermark stripping, particularly where foreseeable or facilitated by platform architecture, raises questions about whether intermediaries are enabling the circumvention of statutory obligations. Watermarks thus transition from voluntary attribution tools to elements of compliance infrastructure.

Electoral Integrity and Democratic Harm: The risks posed by unmarked synthetic content are especially pronounced in electoral contexts. During the enforcement of the Model Code of Conduct for the 2024 general elections, the Election Commission of India (ECI) issued directives requiring political parties and candidates to remove deepfake or AI-manipulated content within three hours of identification. The misuse of AI tools for misinformation or impersonation was expressly prohibited.

These directives reflect an institutional recognition that even brief circulation of unlabelled deepfakes can cause irreversible harm to democratic processes. In such contexts, watermark removal undermines not only individual rights but also the integrity of elections, strengthening the case for heightened regulatory scrutiny.

Watermarking, Labelling, and Emerging Obligations: India's regulatory trajectory further reinforces the legal significance of watermarking. Proposed amendments to the IT Rules seek to mandate the labelling of synthetically generated information through visible disclosures and persistent metadata identifiers.[16] These proposals indicate a shift from advisory-based governance toward enforceable obligations.

Once the law expects AI-generated content to carry indicators of synthetic origin, the deliberate removal of such markers would resemble the circumvention of regulatory safeguards rather than benign modification. In this emerging framework, watermark stripping frustrates regulatory intent and weakens the enforceability of laws aimed at protecting consumers, voters, and victims of impersonation.

A Contracting Legal Grey Zone: Taken together, provisions of the Bharatiya Nyaya Sanhita, the Information Technology Act, intermediary due diligence requirements, and recent regulatory directions suggest that watermark stripping is not legally neutral. While courts have yet to articulate a standalone offence, existing legal frameworks already capture many of the harms enabled by the removal of provenance markers. As regulatory expectations harden into binding rules, the legal grey zone surrounding watermark stripping is likely to contract further, transforming watermark removal from a technical vulnerability into a matter of legal accountability.

REGULATORY AND INSTITUTIONAL CHALLENGES IN ADDRESSING DEEPFAKES AND WATERMARK REMOVAL

Despite increasing regulatory attention, governing deepfakes and AI-generated content remains institutionally and legally challenging. These difficulties are not merely technological; they stem from structural limitations in attribution, enforcement capacity, judicial processes, constitutional constraints, and cross-border regulation. The ease with which AI watermarks can be stripped significantly exacerbates each of these challenges, weakening the effectiveness of both existing laws and emerging regulatory frameworks.

Difficulty in Attribution and Traceability: A central challenge in regulating deepfakes lies in identifying their origin. AI-generated content can be produced anonymously, disseminated through encrypted platforms, and hosted on servers located outside Indian jurisdiction. Watermarking and embedded metadata are intended to mitigate this problem by preserving a signal of provenance and synthetic origin. However, the ease with which

[16] Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules 2025

such markers can be removed undermines this regulatory assumption. When watermarks are stripped, investigators and platforms lose one of the few scalable tools available to link harmful content to its creator or distributor. This severely weakens enforcement under cybercrime and intermediary liability frameworks, which depend on traceability to function effectively.

Institutional Capacity and Enforcement Limitations: Even where legal obligations exist, enforcement is constrained by limited institutional capacity. Detecting deepfakes and verifying whether content originally carried a watermark requires specialised forensic tools and technical expertise that many law enforcement agencies currently lack.[17] While regulatory advisories have encouraged the adoption of AI-based detection mechanisms, watermark removal ensures that even technically detectable signals are absent at the investigation stage. As a result, enforcement becomes reactive and complaint-driven rather than preventive, leading to delayed investigations and weak deterrence.[18] This gap between regulatory ambition and enforcement capability highlights the fragility of governance models that rely on technical safeguards without ensuring their resilience.

Evidentiary Challenges in Judicial Proceedings: The judiciary faces increasing difficulty in assessing the authenticity of digital evidence in cases involving impersonation, misinformation, or reputational harm. Indian courts have consistently emphasised the importance of reliability and authenticity in electronic evidence, particularly under the framework of the Indian Evidence Act 1872.[19] Watermarks can serve as preliminary indicators of synthetic origin, assisting courts in evaluating credibility. Their removal, however, increases uncertainty and places greater reliance on expert testimony and complex forensic analysis. This not only prolongs proceedings but also risks inconsistent evidentiary standards across cases, undermining confidence in judicial fact-finding in the digital age.

Balancing Regulation with Freedom of Speech: Any regulatory response to deepfakes must operate within the constitutional protection of freedom of speech under Article 19(1)(a) of the Constitution of India.[20] The Supreme Court has repeatedly cautioned against vague or

[17] 'Facing Reality? Law Enforcement and the Challenge of Deepfakes' (*Europol Innovation Lab*, 2022) <https://www.europol.europa.eu/cms/sites/default/files/documents/Europol_Innovation_Lab_Facing_Reality_Law_Enforcement_And_The_Challenge_Of_Deepfakes.pdf > accessed 05 December 2025

[18] *Ibid*

[19] *State (NCT of Delhi) v Navjot Sandhu@ Afsan Guru* (2005) 11 SCC 600

[20] Constitution of India 1950, art 19(1)(a)

overbroad restrictions on online expression. In Shreya Singhal v Union of India, the Court struck down Section 66A of the IT Act for its chilling effect on free speech, underscoring the need for precision in regulating digital content.[21] Similarly, in Anuradha Bhasin v Union of India, the Court emphasised proportionality and necessity in restrictions affecting online communication.[22]

Watermarking offers a less speech-restrictive regulatory tool by enabling differentiation between deceptive synthetic content and lawful expression, such as satire or artistic creation. However, when watermarks are easily removed, regulators and platforms may resort to blunt measures such as mass takedowns or pre-emptive censorship, increasing the risk of constitutional overreach. In this way, watermark removal indirectly distorts the balance between regulation and free speech by making targeted enforcement more difficult.

Democratic Integrity and Electoral Harm: The threat posed by unmarked synthetic content is particularly severe in democratic contexts. Deepfakes capable of impersonating political actors or fabricating events can influence voter perception within short timeframes. Recognising this risk, the Election Commission of India adopted a zero-tolerance approach during the 2024 general elections, mandating the rapid takedown of AI-manipulated political content. These measures reflect the understanding that even brief circulation of deceptive media can cause irreversible harm to electoral integrity. Watermark removal undermines such safeguards by allowing synthetic political content to circulate without detection, thereby weakening democratic resilience.

Cross-Border and Platform-Level Constraints: Deepfake regulation is further complicated by the transnational nature of AI systems and digital platforms. Many content-generation tools and hosting services operate outside Indian jurisdiction, limiting the effectiveness of traditional enforcement mechanisms. Watermarks and provenance standards have the potential to operate across borders by embedding origin signals directly into content, irrespective of where it is hosted.[23] However, the ease with which these markers can be stripped allows harmful content to circulate globally without traceability, constraining international cooperation and platform-level accountability.

---

[21] *Shreya Singhal v Union of India* (2015) 5 SCC 1
[22] *Anuradha Bhasin v Union of India* AIR 2020 SC 1308
[23] *Global Risks Report 2024* (WEF, 2024)

Watermark Removal as a Regulatory Fault Line: Taken together, these challenges demonstrate that watermark removal is not merely a technical weakness but a structural regulatory fault line. By eroding traceability, weakening enforcement capacity, complicating judicial processes, and distorting constitutional balances, the stripping of AI watermarks amplifies existing governance failures. As Indian regulators increasingly rely on labelling, metadata, and provenance-based mechanisms to govern synthetic media, the fragility of these tools raises serious questions about the adequacy of current regulatory strategies. Addressing watermark removal is therefore essential not only to mitigate individual harms but to preserve the effectiveness and legitimacy of AI governance itself.

TOWARD A LEGAL FRAMEWORK FOR REGULATING AI WATERMARK CIRCUMVENTION

The preceding analysis demonstrates that while Indian law increasingly relies on watermarking and provenance mechanisms to govern AI-generated content, it lacks an adequate response to their deliberate circumvention. This section argues that effective AI governance requires legal recognition of watermark stripping as a distinct regulatory concern. Rather than treating such conduct as legally irrelevant until downstream harm occurs, Indian law should adopt a framework that addresses circumvention upstream, while remaining consistent with constitutional safeguards and proportionality principles.

Recognising Watermark Circumvention as a Distinct Regulatory Harm: A central weakness of the current legal approach is its exclusive focus on downstream misuse. By contrast, many regulatory regimes intervene earlier, recognising that the disabling of safeguards itself undermines regulatory objectives. AI watermarking functions as an infrastructure of accountability: it enables traceability, supports intermediary compliance, and facilitates enforcement. Deliberate watermark removal frustrates these objectives even where concrete harm is difficult to quantify.

Indian law should therefore recognise watermark circumvention as a legally cognisable harm in itself—not because watermarking creates proprietary rights, but because it performs a public regulatory function. Such recognition would align AI governance with existing regulatory logic in areas such as financial disclosure and environmental monitoring, where evasion of transparency mechanisms attracts liability independent of proven downstream damage.

Doctrinal Pathways for Legal Recognition: There are multiple doctrinal routes through which Indian law could address watermark circumvention without creating overbroad offences. First, courts could treat deliberate watermark removal as evidence of mens rea in offences involving fraud, impersonation, or misinformation. The act of stripping provenance markers could be used to infer intent to deceive or evade detection, lowering evidentiary barriers in appropriate cases.

Second, watermark circumvention could be recognised as an aggravating factor in sentencing or regulatory penalties, particularly where the conduct demonstrably undermines enforcement or public trust. This approach avoids criminalising benign conduct while signalling heightened concern for intentional evasion.

Third, legislative intervention could create a narrowly tailored offence or civil wrong focused specifically on the deliberate removal of mandated provenance markers, where such removal defeats statutory or regulatory safeguards. Such a provision would need to be carefully drafted to avoid capturing incidental or technically unavoidable degradation of watermarks.

Duties of AI Developers and Platforms: Any framework addressing watermark circumvention must also consider the responsibilities of AI developers and intermediaries. Where regulators mandate watermarking or labelling, developers should be expected to implement safeguards that make circumvention meaningfully difficult, rather than merely symbolic.[24] Platforms, in turn, should be required to preserve provenance indicators during ordinary processing and refrain from design choices that systematically strip metadata.

At the same time, liability should be calibrated to reflect control and foreseeability. Developers and platforms should not be held strictly liable for every instance of watermark removal by users, but they may bear responsibility where circumvention is predictable and unaddressed despite available technical measures. This approach aligns with existing intermediary liability principles under Indian law.

[24] Proposed Amendments to the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 in relation to Synthetically Generated Information (n 3)

Safeguarding Free Speech and Legitimate Use: Concerns about overreach and censorship must be addressed explicitly. Any legal response to watermark circumvention must be narrowly scoped to target intentional evasion, not creative transformation, research, or legitimate anonymisation. Importantly, watermark regulation should focus on provenance disclosure, not content suppression.

By targeting the removal of identifiers rather than the content itself, such a framework may, in fact, be less speech-restrictive than blunt takedown regimes. Effective provenance enables regulators and platforms to distinguish harmful deception from lawful expression, reducing pressure for overbroad content moderation that risks violating Article 19(1)(a).

From Soft Law to Enforceable Governance: Finally, addressing watermark circumvention requires a shift from advisory-driven governance to enforceable standards. Soft law has played a valuable role in signalling regulatory priorities, but its effectiveness is limited in adversarial contexts. As AI-generated content becomes more realistic and easier to weaponise, governance mechanisms must anticipate strategic evasion rather than assume compliance.

Recognising watermark circumvention as a regulatory wrong would not resolve all challenges posed by synthetic media. However, it would strengthen the integrity of existing safeguards and restore coherence to a governance framework that increasingly depends on provenance-based regulation. Without such recognition, watermarking risks becoming a formal requirement devoid of practical effect.

CONCLUSION

The rapid normalisation of AI-generated content has compelled regulators to rely on technical provenance mechanisms such as watermarking to preserve transparency, accountability, and public trust. In the Indian context, recent advisories and regulatory directions increasingly position watermarking and labelling as central tools in addressing the risks posed by deepfakes, misinformation, and impersonation. Yet, the legal effectiveness of these measures is fundamentally compromised by the ease with which AI watermarks can be deliberately removed.

Indian law currently approaches synthetic media harms through a downstream, harm-based lens, addressing deception only after it materialises. This framework is ill-suited to an ecosystem in which regulatory compliance depends on the persistence of technical safeguards that can be neutralised upstream. The deliberate stripping of AI provenance markers is not a neutral act of content modification; it is a form of regulatory circumvention that undermines the very architecture upon which contemporary AI governance relies.

By drawing analogies to technological protection measures in copyright law and to anti-evasion principles in securities and environmental regulation, Indian law already recognises the necessity of intervening against the defeat of regulatory safeguards, even in the absence of immediate harm. Extending this logic to AI watermarking does not require a radical departure from existing doctrine, but rather a recalibration of legal focus toward intent, circumvention, and systemic risk.

Recognising watermark circumvention as a distinct regulatory concern would strengthen the coherence of India's emerging AI governance framework while remaining consistent with constitutional constraints under Article 19. A narrowly tailored approach targeting deliberate and bad-faith evasion while excluding lawful editing, research, and privacy-preserving practices offers a proportionate path forward that avoids chilling legitimate speech or innovation.

Ultimately, the challenge posed by AI-generated content is not merely technological, but institutional. As generative systems become more capable and adversarial use becomes more common, regulation cannot rely solely on voluntary compliance or fragile technical markers. If watermarking is to function as more than a symbolic gesture, the law must be prepared to protect the safeguards it mandates. The treatment of AI watermark circumvention will therefore serve as an early test of whether Indian law can move beyond reactive enforcement toward anticipatory and resilient AI governance.