Jurnal Hukum Siber
dan Regulasi Teknologi

SAH PUBLISHER

e-ISSN 3123-8114
Vol. 1, No. 1, January 2026

The Urgency of Mandatory Labelling of Artificial Intelligence-Generated Content to Prevent Disinformation and Digital Fraud in Indonesia

Muhammad Rizal Alief Ramadhan[1]*, Allya Putri Sukamto[1]
[1]Department of Business Law, Universitas Sugeng Hartono, SUkoharjo, Indonesia
Email: muhrifzalar@sugenghartono.ac.id

Abstract. Generative artificial intelligence enables ordinary users and organised actors to create text, images, audio, and video that can imitate real persons, institutions, and events at low cost. In Indonesia, this capability intersects with existing problems of hoaxes, impersonation, consumer deception, and personal-data misuse. This article argues that Indonesia needs a binding obligation to label AI-generated and AI-manipulated content, especially when such content is distributed to the public, used in electronic transactions, or capable of affecting public trust. The research uses a normative juridical method with statutory, conceptual, and comparative approaches. It examines the Electronic Information and Transactions Law, the Personal Data Protection Law, rules on electronic system providers, and Indonesia's AI ethics policy, then compares them with transparency models in the EU AI Act, C2PA provenance standards, NIST guidance, and Chinese synthetic-content labelling rules. The analysis finds that Indonesian law can sanction false information, fraud, and unlawful data processing after harm occurs, but it does not yet impose a clear ex ante duty to disclose synthetic origin. A labelling regime would close this gap without turning every synthetic work into prohibited content. The article proposes a layered model: creator-side disclosure, provider-side machine-readable provenance, platform-side label preservation, heightened duties for political and financial-risk content, and due-process safeguards for lawful expression. Mandatory labelling should therefore be framed as a transparency and consumer-protection obligation, not as a blanket restriction on innovation.

Keywords: Artificial intelligence, AI-generated content, content labelling, disinformation, digital fraud, Indonesian cyber law

INTRODUCTION

Generative artificial intelligence has altered the evidentiary value of digital content. Text generators can produce persuasive narratives, image systems can create realistic photographs of events that never occurred, and audio or video systems can simulate a person's voice or appearance. Legal scholarship on deepfakes has warned that synthetic media may intensify harms to privacy, democracy, reputation, and national security because the technology lowers both the cost and skill needed to fabricate believable evidence [1]. Public policy reports describe the same risk in practical terms: a deepfake can make a person appear to say or do something that never happened [2]. The problem is not limited to technically advanced deepfakes. Cheap fakes, recontextualised footage, and AI-assisted editing can also mislead audiences because truth is produced through social processes, not merely through pixels and metadata [3].

Received May 2026 / Revised May 2026 / Accepted May 2026
*Corresponding author.
Email addresses: muhrifzalar@sugenghartono.ac.id (Ramadhan)

Indonesia already has several legal tools that may respond to harmful online conduct. Law No. 1 of 2024, as the second amendment to the Electronic Information and Transactions Law, adjusts the framework for prohibited electronic information and electronic transactions [4]. Law No. 27 of 2022 on Personal Data Protection establishes obligations for controllers and processors, data-subject rights, sanctions, and criminal provisions concerning personal data [5]. Government Regulation No. 71 of 2019 regulates electronic system and transaction governance [6], while Ministerial Regulation No. 5 of 2020 sets duties for private electronic system providers and has become central to platform accountability [7]. Indonesia has also issued Ministerial Circular Letter No. 9 of 2023 on Artificial Intelligence Ethics as a policy reference for public and private electronic system providers [8], and the National AI Strategy 2020-2045 places ethics and policy as one of the pillars of national AI development [9].

Despite those instruments, none of them gives users a simple and enforceable right to know when publicly circulated content was generated or materially manipulated by AI. This gap matters because many AI harms arise before a court can classify a message as defamation, fraud, hate speech, consumer deception, or unlawful processing of personal data. A synthetic video of a public official, a cloned voice in a banking scam, or an AI-generated advertisement using a celebrity likeness may cause loss before takedown or criminal enforcement begins. Labelling can reduce that temporal gap by giving users, platforms, journalists, investigators, and courts a visible signal at the point of encounter.

This article asks whether Indonesia should impose a mandatory labelling obligation for AI-generated content and how such an obligation can be designed consistently with freedom of expression, innovation, and existing cyber-law architecture. The central argument is that mandatory labelling is urgent, but it should be narrow, risk-based, and technically realistic. It should not prohibit lawful satire, artistic expression, accessibility tools, or routine editing. Instead, it should require disclosure when synthetic origin is material to audience trust, transaction safety, personal-data protection, or public-interest communication.

METHODS

This research uses a normative juridical method. The statutory approach examines Indonesian positive law governing electronic information, electronic system providers, personal data, and AI ethics, especially Law No. 1 of 2024, Law No. 27 of 2022, Government Regulation No. 71 of 2019, Ministerial Regulation No. 5 of 2020, and Ministerial Circular Letter No. 9 of 2023 [4]-[8]. The conceptual approach evaluates labelling as a transparency obligation, a consumer-protection instrument, and a form of risk governance. The comparative approach is limited to regulatory models that directly address synthetic content, provenance, or platform responsibility.

The comparative materials include OECD principles on transparency and responsible disclosure [10], UNESCO's Recommendation on the Ethics of Artificial Intelligence [11], the NIST AI Risk Management Framework [12], NIST's Generative AI Profile with its focus on content provenance and incident disclosure [13], the EU AI Act's transparency obligations for synthetic content [14], and the C2PA content provenance standard [15]. These sources are used not as binding law for Indonesia, but as references for designing a domestic rule that is administrable, proportionate, and compatible with Indonesia's existing electronic-system-provider regime.

RESULT AND DISCUSSION

1. AI-generated content as infrastructure for disinformation and fraud

AI-generated content should be treated as infrastructure, not merely as a category of speech. It enables speed, scale, personalisation, and plausible impersonation. Empirical research on political deepfakes suggests that such content may create uncertainty even when it does not fully deceive audiences, and that uncertainty can reduce trust in news on social media [16]. Philosophical analysis similarly warns that deepfakes threaten the social role of recordings as an epistemic backstop, because audiences may become less willing to treat audio-visual evidence as a reliable check on testimony [17]. Election-focused scholarship has also shown how synthetic media can affect democratic integrity by distorting the information environment in which voters form judgments [18].

The fraud dimension is equally important. Deepfake and voice-cloning technologies can be used to imitate family members, corporate officers, government agencies, or financial-service providers. Studies on political and social deepfakes show that synthetic media is not always more persuasive than textual misinformation, but it remains harmful because it can blur source identity and exploit cognitive shortcuts [19], [20]. In the terminology of information disorder, AI-generated deception may combine misinformation, disinformation, and malinformation, depending on falsity, intent, and harm [21]. Labelling is therefore not a complete solution, but it is a necessary first layer for reducing ambiguity about origin.

2. The gap in Indonesian positive law

The ITE Law can address certain harmful outputs after they are distributed, especially where electronic information causes consumer loss, hatred, or other prohibited effects [4]. The PDP Law can respond when a person's identity, biometric features, voice, image, or other personal data are processed unlawfully [5]. PP 71/2019 and Ministerial Regulation No. 5/2020 create obligations for electronic system reliability, registration, and platform governance [6], [7]. These instruments, however, do not create a general duty to disclose that content is AI-generated before harm occurs.

The AI Ethics Circular is important because it introduces values such as accountability, transparency, security, and personal-data protection into Indonesia's AI-policy vocabulary [8]. Yet a circular letter is not enough for a public-facing labelling regime. Its normative force is weaker than a statute or government regulation, and it does not specify who must label content, what counts as sufficient disclosure, whether labels must be machine-readable, how platforms must preserve labels, or what sanctions apply for deliberate removal. As a result, Indonesia has a post-harm enforcement structure but lacks a pre-harm transparency duty.

3. Why mandatory labelling is legally urgent

Mandatory labelling is urgent for three reasons. First, it supports autonomy. Users cannot evaluate a message, advertisement, voice note, or video properly if they do not know whether it is a representation of a real event or a synthetic construction. OECD and UNESCO principles both connect trustworthy AI with transparency, explainability, and human accountability [10], [11]. Second, labelling supports prevention. NIST's AI RMF treats AI risk as a socio-technical problem requiring governance, mapping, measurement, and management across the lifecycle [12]. Its Generative AI Profile identifies content provenance as one of the key governance areas for generative AI [13].

Third, labelling is less restrictive than takedown or criminalisation. It allows synthetic content to remain online when it is lawful, educational, satirical, artistic, or assistive, while still warning users that the content is not an unmediated record. This distinction is crucial in Indonesia, where cyber-law enforcement has often raised concerns about overbreadth and legal uncertainty. A well-drafted labelling duty can reduce reliance on criminal enforcement by giving regulators and platforms a preventive, administrative, and auditable compliance tool.

4. Comparative regulatory models

The EU AI Act provides one of the clearest statutory models. Article 50 requires providers of AI systems that generate synthetic audio, image, video, or text to ensure that outputs are marked in machine-readable format and detectable as artificially generated or manipulated, subject to limited exceptions [14]. This model is valuable because it distinguishes between provider-side technical marking and deployer-side disclosure to affected persons. It also recognises that technical solutions must be effective, interoperable, robust, and reliable as far as technically feasible.

Technical provenance standards complement legal duties. C2PA provides an open standard for recording the origin and edit history of digital content through content credentials [15]. Provenance does not prove that content is true, but it can show where the content came from, what edits were made, and whether a label has been stripped. China's deep synthesis and AI-generated content measures are more prescriptive: they require providers to label synthetic content and regulate deep synthesis services as part of broader platform governance [23], [24]. The EU's code of practice on transparency of AI-generated content, adopted to support AI Act compliance, also treats marking and labelling as a practical compliance ecosystem rather than a single notice [25].

Indonesia need not copy any of these models wholesale. The EU approach may be too technical if adopted without local capacity building; the Chinese approach may be too state-centred if imported without constitutional safeguards. The more appropriate lesson is layered governance. Technical marking should begin at the model or service-provider level, visible labelling should be preserved by platforms, and high-risk public distribution should trigger clearer human-readable disclosure.

5. Proposed Indonesian model

Indonesia should adopt a binding rule, preferably through a government regulation or ministerial regulation grounded in the ITE Law and PP 71/2019, while preparing a future AI statute. The rule should define AI-generated content as content materially generated or manipulated by an AI system so that a reasonable user may believe it records a real person, event, statement, transaction, or institutional communication. The rule should exclude minor editing, spelling correction, assistive translation, compression, and accessibility functions when they do not alter meaning or identity.

The obligation should apply to four actor groups: AI service providers, deployers or users who publish content at scale, electronic system providers that distribute public content, and advertisers or merchants who use AI content in electronic transactions. The proposed model is summarised in Table 1.

Table 1. Proposed layered obligation model for AI-generated content labelling in Indonesia

Layer	Responsible actor	Core obligation	Regulatory basis
Source marking	AI service provider	Embed machine-readable provenance or watermarking where technically feasible; document limitations.	AI ethics, NIST provenance, C2PA, future AI rule
Visible disclosure	Publisher, advertiser, or deployer	Show a clear label when AI content represents persons, institutions, public events, public services, or offers in electronic transactions.	ITE Law, consumer protection logic, AI Ethics Circular
Label preservation	Electronic system provider	Do not remove embedded provenance; display labels on upload, repost, recommendation, and paid promotion interfaces.	PP 71/2019 and PSE private-scope regulation
High-risk escalation	Platform and sector regulator	Apply stronger review for political advertising, financial services, health, disaster information, and identity-sensitive content.	PDP Law, OJK AI guidance, sectoral risk governance
Due process	Regulator and platform	Provide notice, appeal, correction, and record-keeping before sanctions, except in urgent fraud or public-safety cases.	Rule of law, human-rights safeguards, proportionality

The model should also include sanctions that match the nature of the breach. Failure to label should first trigger warning, correction, notice to users, temporary restriction of paid amplification, or administrative fines. Criminal liability should remain reserved for conduct that independently satisfies fraud, unlawful access, identity misuse, personal-data offences, or other criminal elements. This structure prevents labelling law from becoming a general censorship instrument.

The financial sector illustrates why sectoral coordination is necessary. OJK's responsible and trustworthy AI guidance for fintech emphasises beneficial use, fairness and accountability, transparency and explicability, robustness, and security [26]. Those principles support a stricter labelling rule for AI-generated investment advice, loan offers, customer-service bots, and voice or video communications that may induce consumers to transfer money or disclose credentials. The same logic applies to public services and election-related content, where mistaken trust can create collective harm.

The proposed rule should be accompanied by public literacy and procurement standards. Public institutions should label their own AI-generated public communications and require vendors to preserve provenance in government-facing systems. Private platforms should publish transparency reports on labelled content, removed labels, appeal outcomes, and detected impersonation campaigns. These duties would align Indonesia with a broader movement toward responsible AI governance while respecting the principle that human actors remain accountable for AI-assisted communication [27], [28].

CONCLUSION

Mandatory labelling of AI-generated content is urgent for Indonesia because existing cyber, data-protection, and electronic-system rules mainly respond after harmful content has circulated, while generative AI creates deception risks at the moment content is encountered. A labelling duty would not solve disinformation or digital fraud by itself, but it would supply a preventive layer that helps users assess source, helps platforms preserve provenance, helps regulators audit compliance, and helps courts distinguish lawful synthetic expression from deceptive impersonation. Indonesia should therefore adopt a layered and risk-based labelling regime that combines machine-readable provenance, visible user-facing labels, platform preservation duties, heightened obligations for transaction and public-interest contexts, and due-process safeguards. Framed in this way, labelling is not anti-innovation; it is a proportional transparency obligation needed to protect public trust in Indonesia's digital ecosystem.

REFERENCES

[1] R. Chesney and D. K. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," California Law Review, vol. 107, pp. 1753-1819, 2019. Available: https://www.californialawreview.org/print/deep-fakes-a-looming-challenge-for-privacy-democracy-and-national-security

[2] U.S. Government Accountability Office, "Science & Tech Spotlight: Deepfakes," GAO-20-379SP, Feb. 2020. Available: https://www.gao.gov/assets/gao-20-379sp.pdf

[3] B. Paris and J. Donovan, "Deepfakes and cheap fakes: The manipulation of audio and visual evidence," Data & Society, 2019. Available: https://datasociety.net/library/deepfakes-and-cheap-fakes/

[4] Republic of Indonesia, Law No. 1 of 2024 on the Second Amendment to Law No. 11 of 2008 concerning Electronic Information and Transactions. Available: https://peraturan.bpk.go.id/details/274494/uu-no-1-tahun-2024

[5] Republic of Indonesia, Law No. 27 of 2022 concerning Personal Data Protection. Available: https://peraturan.bpk.go.id/Details/229798/uu-no-27-tahun-2022

[6] Republic of Indonesia, Government Regulation No. 71 of 2019 concerning the Operation of Electronic Systems and Transactions. Available: https://peraturan.bpk.go.id/Details/122030/pp-no-71-tahun-2019

[7] Ministry of Communication and Informatics, Regulation No. 5 of 2020 concerning Private-Scope Electronic System Providers. Available: https://peraturan.bpk.go.id/Details/203049/permenkominfo-no-5-tahun-2020

[8] Ministry of Communication and Digital Affairs, "Resmi Terbitkan SE Menkominfo Jadi Pedoman bagi PSE Publik dan Privat," Press Release No. 582/HM/KOMINFO/12/2023. Available: https://www.komdigi.go.id/berita/siaran-pers/detail/siaran-pers-no-582-hm-kominfo-12-2023-tentang-resmi-terbitkan-se-menkominfo-jadi-pedoman-bagi-pse-publik-dan-privat

[9] KORIKA, "Strategi Nasional Kecerdasan Artifisial Indonesia 2020-2045." Available: https://korika.id/en/document/strategi-nasional-kecerdasan-artifisial-indonesia-2020-2045/

[10] OECD, "OECD AI Principles: Transparency and explainability." Available: https://www.oecd.org/en/topics/sub-issues/ai-principles.html

[11] UNESCO, "Recommendation on the Ethics of Artificial Intelligence," 2021. Available: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

[12] National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, Jan. 2023. Available: https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10

[13] National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile," NIST AI 600-1, 2024. Available: https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence

[14] European Parliament and Council, Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence, 13 June 2024. Available: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

[15] Coalition for Content Provenance and Authenticity, "C2PA: Verifying media content sources." Available: https://c2pa.org/
[16] C. Vaccari and A. Chadwick, "Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news," Social Media + Society, vol. 6, no. 1, 2020. doi: 10.1177/2056305120903408.
[17] R. Rini, "Deepfakes and the epistemic backstop," The Philosophers' Imprint, vol. 20, no. 24, 2020. Available: https://philpapers.org/archive/RINDAT.pdf
[18] N. Diakopoulos and D. Johnson, "Anticipating and addressing the ethical implications of deepfakes in the context of elections," New Media & Society, vol. 23, no. 7, pp. 2072-2098, 2021. doi: 10.1177/1461444820925811.
[19] M. Hameleers, T. G. L. A. van der Meer, and T. Dobber, "You won't believe what they just said! The effects of political deepfakes embedded as vox populi on social media," Social Media + Society, vol. 8, no. 3, 2022. doi: 10.1177/20563051221116346.
[20] J. T. Hancock and J. N. Bailenson, "The social impact of deepfakes," Cyberpsychology, Behavior, and Social Networking, vol. 24, no. 3, pp. 149-152, 2021. doi: 10.1089/cyber.2021.29208.jth.
[21] C. Wardle and H. Derakhshan, "Information disorder: Toward an interdisciplinary framework for research and policy making," Council of Europe, 2017. Available: https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html
[22] Council of Europe, "Council of Europe opens first ever global treaty on AI for signature," Sept. 5, 2024. Available: https://www.coe.int/en/web/portal/-/council-of-europe-opens-first-ever-global-treaty-on-ai-for-signature
[23] Library of Congress, "China: Provisions on Deep Synthesis Technology Enter into Effect," Apr. 25, 2023. Available: https://www.loc.gov/item/global-legal-monitor/2023-04-25/china-provisions-on-deep-synthesis-technology-enter-into-effect/
[24] China Law Translate, "Measures for Labeling of AI-Generated Synthetic Content." Available: https://www.chinalawtranslate.com/en/ai-labeling/
[25] European Commission, "Code of Practice on Transparency of AI-Generated Content," 2026. Available: https://digital-strategy.ec.europa.eu/en/policies/code-practice-ai-generated-content
[26] Otoritas Jasa Keuangan, "Panduan Kode Etik Kecerdasan Buatan (AI) yang Bertanggung Jawab dan Terpercaya di Industri Teknologi Finansial," 2023. Available: https://www.ojk.go.id/id/berita-dan-kegiatan/publikasi/Documents/Pages/Panduan-Kode-Etik-Kecerdasan-Buatan-AI-yang-Bertanggung-Jawab-dan-Terpercaya-di-Industri-Teknologi-Finansial/OJK_Panduan%20Kode%20Etik%20Kecerdasan%20Buatan%20AI%20Yang%20Bertanggungjawab%20dan%20Terpercaya%20di%20Industri%20Teknologi%20Finansial.pdf
[27] L. Floridi et al., "AI4People: An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations," Minds and Machines, vol. 28, no. 4, pp. 689-707, 2018. doi: 10.1007/s11023-018-9482-5.
[28] V. Dignum, Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Cham, Switzerland: Springer, 2019. doi: 10.1007/978-3-030-30371-6.