How does the classification of AI-generated content as propaganda reshape institutional accountability and public trust in information ecosystems?

September 1, 2026 | SnugLab Research

Executive Summary

The classification of AI-generated content as propaganda fundamentally reshapes institutional accountability by compelling a structural overhaul of oversight mechanisms, while simultaneously eroding public trust through the "liar's dividend." The sheer volume and sophistication of AI-generated synthetic sentiment overwhelm traditional verification thresholds, forcing institutions to implement new provenance tracking and automated auditing to distinguish genuine public input from fabricated content [7]. Concurrently, the widespread public distrust in AI-generated information allows officials to exploit the "liar's dividend," dismissing authentic criticism as "AI noise" and further undermining confidence in information ecosystems [11, 12].

Key Findings

AI Propaganda Functions as Both Coordinated Messaging and Decentralized Noise

AI-generated content classified as propaganda encompasses both intentionally coordinated messaging and decentralized, algorithmically amplified synthetic signals. State operatives, political campaigns, and malicious actors actively deploy AI to create targeted deepfakes and robocalls designed to engineer mass perception [5, 9, 14]. Simultaneously, generative AI democratizes content creation, enabling a wider range of actors to produce vast quantities of synthetic media that algorithms amplify virally, blurring the line between organic public discourse and manufactured manipulation [1, 4]. This dual nature complicates institutional efforts to discern genuine constituent sentiment from fabricated input at scale, leading to epistemic uncertainty and accelerating public skepticism [1, 5, 7, 12].

Classification Drives Structural Overhauls and Epistemic Evasion

The classification of AI propaganda is driving both structural overhauls in institutional oversight and providing officials with a pretext for epistemic evasion via the "liar's dividend." The overwhelming scale and speed of AI-generated synthetic media necessitate new regulatory and technical frameworks. The EU AI Act, which entered into force on August 1, 2024, and became applicable on August 2, 2026, mandates transparency and disclosure for generative AI providers [8, 13]. In the United States, proposed legislation, such as the REAL Political Ads Act, seeks to legally mandate disclaimers on AI-generated political content [14]. The ability of AI to flood regulatory comment sections with millions of synthetic inputs compels agencies to adopt automated auditing and machine-readable provenance tracking to verify public input [7]. Major technology firms are also deploying technical solutions like embedded watermarks and metadata indicators to establish content origin [8, 14].

Conversely, this classification offers an epistemic shield for officials. Public trust in AI-generated content is low, with only 26% of respondents trusting it and 68% considering it untrustworthy [11]. This widespread skepticism allows politicians to dismiss authentic evidence or damaging criticism as "AI noise" or "deepfakes" [12]. Experimental studies confirm that false claims of misinformation significantly increase politician support, particularly against text-based scandals [12]. Institutions have been shown to respond to AI-generated content at rates statistically indistinguishable from human-written input, suggesting that the classification can enable evasion without necessarily requiring immediate structural changes [7, 12].

AI-Generated Sentiment Decouples Institutional Responsiveness from Public Will

AI-generated synthetic sentiment is mechanistically decoupling institutional responsiveness from genuine public will because existing filtering mechanisms are struggling to preserve meaningful democratic accountability. The sheer volume and sophistication of AI output enable actors to generate false constituent sentiment at scale, overwhelming traditional verification thresholds [7]. Generative AI can produce vast quantities of coherent, human-like content rapidly and cheaply, blurring the line between authentic and inauthentic input [1, 3, 7]. A 2020 field experiment demonstrated that state legislators responded to AI-generated advocacy letters at rates only 1.9 percentage points lower than human-written letters (15.4% vs 17.3%) [7]. Regulatory bodies face the prospect of limitless unique comments flooding e-rulemaking platforms, making it difficult to ascertain genuine public input [7]. Advanced AI easily overcomes older detection

methods, such as identifying duplicate text, which previously caught bot-driven astroturfing like the 2017 FCC net neutrality comment flood where 94% of comments were not unique [7].

Existing filtering mechanisms, such as labeling and watermarking, are struggling to preserve accountability. While AIGC labels help users distinguish synthetic from human content, a 2024 study found they minimally affect perceived accuracy, message credibility, or sharing intention [2]. Paradoxically, labeling can slightly enhance sharing intention and perceived accuracy for misinformation in certain contexts [2]. Furthermore, these initiatives are often voluntary, inconspicuous, easily bypassed, or rely on self-disclosure [14].

Mandatory Labels Fail to Rebuild Trust, Inconsistent Enforcement Erodes Authority

Mandatory provenance labels and watermarks for AI-generated content have not effectively rebuilt public trust and their inconsistent enforcement further erodes institutional authority. While labels help users distinguish AI-generated content from human-created content [2], a 2024 study found they minimally affect perceived accuracy, message credibility, or sharing intention [2]. In some contexts, labeling can even slightly enhance the perceived accuracy and sharing intention of misinformation [2]. Inconsistent enforcement, where measures are often voluntary, inconspicuous, easily removed, or rely on self-disclosure, fuels skepticism rather than resolving ambiguity [14]. The "implied truth effect" suggests that the absence of a warning, due to inconsistent application, can inadvertently imply veracity, leaving the public uncertain about what to trust [2].

This dynamic creates a paradox where verified synthetic content is perceived as equally credible (or less credible) than unverified human reporting. Public skepticism toward AI-generated media is deep, with only 26% trusting information produced by AI [11]. This uncertainty allows officials to exploit the "liar's dividend," using public skepticism of AI to dismiss genuine criticism as synthetic noise or to falsely claim authentic damaging evidence is a deepfake [9, 12]. Consequently, inconsistent labeling contributes to a broader crisis where institutional authority is undermined by blanket skepticism toward all information [7, 10].

AI Propaganda Accelerates Disinformation and Necessitates Epistemic Reforms

AI-generated propaganda represents both an accelerated iteration of historical disinformation cycles and a fundamental shift requiring new epistemic reforms to maintain institutional trust. Historically, bot-driven astroturfing campaigns were detectable by repetitive patterns; for example, 94% of the 21.7 million comments to the FCC in 2017 were not unique [7]. Advanced generative AI overcomes this by producing vast quantities of coherent, human-like content at low cost and high speed, making synthetic outputs indistinguishable from authentic ones [1, 7].

This capability has permanently lowered the cognitive threshold for public skepticism. Even before widespread generative AI, humans struggled to distinguish synthetic media from human-created content [2, 5]. Today's deepfakes are considered more persuasive and likely to go viral than text-based disinformation [3]. Public skepticism has surged, with 76% of Americans concerned about AI producing false information, and only 26% trusting information produced by AI [6, 11].

Despite this, traditional journalism institutions still maintain a significant trust advantage, with 78.2% of respondents trusting traditional media compared to 82.5% distrusting AI-created news [1]. However, the "liar's dividend" complicates institutional accountability by allowing public figures to dismiss damaging but authentic evidence as "fake news" or deepfakes [9, 12]. While established newsrooms use AI tools sparingly with stringent checks, malicious actors can easily create fake websites mimicking legitimate ones, threatening the broader information ecosystem [4].

Because traditional trust buffers are no longer sufficient to verify authenticity at scale, institutions must adopt fundamental epistemic reforms. These include implementing machine-readable provenance standards, deploying automated verification protocols, and investing in sustained digital literacy campaigns [1, 7, 8, 13]. Without these structural changes, AI-fueled deception risks permanently undermining democratic governance and societal trust [9, 13, 14].

Government and Regulatory Bodies Face Significant Decoupling

The government (legislative and executive branches), political campaigning, and regulatory bodies are experiencing the most significant decoupling of institutional responsiveness from genuine public will due to AI synthetic sentiment flooding. In political campaigning and legislative representation, AI allows actors to manufacture false "constituent sentiment" at scale [7]. The 2020 field experiment showing state legislators responded to AI-generated advocacy letters at rates only 1.9 percentage points lower

than human-written letters highlights this decoupling [7]. Similarly, regulatory bodies face a decoupling of policy-making from authentic public input as AI floods e-rulemaking platforms with limitless unique comments, easily bypassing older detection methods [7]. The Ukrainian Center for Countering Disinformation (CCD) provides a concrete example of a national government institution formally addressing AI-generated content as propaganda, having uncovered an AI-constructed information campaign against President Volodymyr Zelenskyy involving bogus films and articles [15].

Implications

The classification of AI-generated content as propaganda necessitates a dual approach for institutions: implementing robust technical and regulatory frameworks for accountability while simultaneously addressing the erosion of public trust. For governments and regulatory bodies, this means prioritizing the development and mandatory adoption of machine-readable provenance standards and automated auditing systems to verify the authenticity of public input and prevent the manipulation of democratic processes [7, 8, 13]. For media organizations and the public, it implies a need for sustained digital literacy campaigns to equip citizens with critical thinking skills to navigate an increasingly complex information environment [1, 7, 8, 13]. Without these reforms, the "liar's dividend" will continue to undermine institutional authority, fostering political nihilism and making it increasingly difficult for genuine public will to influence governance [12].

Limitations and Caveats

The available research, while extensive, has limitations. Specific quantitative data on the compliance costs or enforcement penalties associated with the EU AI Act or other mandatory labeling initiatives is not provided. Furthermore, there is a lack of specific named longitudinal studies or comprehensive cross-jurisdictional datasets that fully quantify how the explicit propaganda classification of AI content has quantitatively shifted citizen trust metrics toward information institutions over time [8, 12, 13, 14]. While individual studies measure aspects of trust, a broader, long-term view is needed. Data on the adoption rates and comparative user trust impact of specific mandatory labeling standards (e.g., C2PA, W3C Verifiable Credentials) across major social media platforms and news aggregators is also limited.

Sources

[1] [peer-reviewed] The impact of automated journalism on media bias, accuracy, and public trust: evidence from young Chinese news consumers - Authors: LI, Yan - Journal: Humanities and Social Sciences Communications - https://www.nature.com/articles/s41599-026-06612-6

[2] [peer-reviewed] Impact of Artificial Intelligence-Generated Content Labels On Perceived Accuracy, Message Credibility, and Sharing Intentions for Misinformation: Web-Based, Randomized, Controlled Experiment - Authors: Fan Li; Ya Yang - Journal: JMIR Formative Research - https://pmc.ncbi.nlm.nih.gov/articles/PMC11892328/

[3] [edu] Disinformation Machine How Susceptible Are We Ai Propaganda - hai.stanford.edu - https://hai.stanford.edu/news/disinformation-machine-how-susceptible-are-we-ai-propaganda

[4] [edu] Milton Wolf Seminar Media And Diplomacy 0 - asc.upenn.edu - https://www.asc.upenn.edu/research/centers/milton-wolf-seminar-media-and-diplomacy-0

[5] [edu] The Origin Of Public Concerns Over Ai Supercharging Misinfor - misinforeview.hks.harvard.edu - https://misinforeview.hks.harvard.edu/article/the-origin-of-public-concerns-over-ai-supercharging-misinformation-in-the-2024-u-s-presidential-election/

[6] [edu] What Americans Really Think About Ai Algorithms Public Confi - publicpolicy.cornell.edu - https://publicpolicy.cornell.edu/masters-blog/what-americans-really-think-about-ai-algorithms-public-confidence-and-transparency-in-government/

[7] How Ai Threatens Democracy - journalofdemocracy.org - https://www.journalofdemocracy.org/articles/how-ai-threatens-democracy/

[8] [peer-reviewed] A Regulation-Aware Architecture for Governing AI-Generated Content: Provenance, Labeling, Privacy, Safety, and Institutional Compliance - Authors: GuillÃ©n Yparrea, Nicia; GarcÃ-a LÃ³pez, IvÃ¡n Miguel - Journal: Frontiers in Artificial Intelligence - https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2026.1885306/full

[9] Gauging Ai Threat Free And Fair Elections - brennancenter.org - https://www.brennancenter.org/our-work/analysis-opinion/gauging-ai-threat-free-and-fair-elections

[10] Has Ai Ushered In An Existential Crisis Of Trust In Democrac - odi.org - https://odi.org/en/insights/has-ai-ushered-in-an-existential-crisis-of-trust-in-democracy/

[11] PressRelease.Asp - nab.org - https://www.nab.org/documents/newsroom/pressRelease.asp?id=7344

[12] [peer-reviewed] The Liar's Dividend: Can Politicians Claim Misinformation to Evade Accountability? - Authors: KAYLYN JACKSON SCHIFF; DANIEL S. SCHIFF; NATÃ•LIA S. BUENO - Journal: American Political Science Review - https://www.cambridge.org/core/journals/american-political-science-review/article/liars-dividend-can-politicians-claim-misinformation-to-evade-accountability/687FEE54DBD7ED0C96D72B26606AA073

[13] [peer-reviewed] AI-driven disinformation: policy recommendations for democratic resilience - Authors: Alexander Romanishyn; Olena Malytska; Vitaliy Goncharuk - Journal: Frontiers in Artificial Intelligence - https://pmc.ncbi.nlm.nih.gov/articles/PMC12351547/

[14] [gov] Press Releases Id 653e18a2 5bb8 40e9 A95d 98c5d0941a29 - bennet.senate.gov - https://www.bennet.senate.gov/2023/06/29/press-releases-id-653e18a2-5bb8-40e9-a95d-98c5d0941a29/

[15] [peer-reviewed] Countering AI-powered disinformation through national regulation: learning from the case of Ukraine - Authors: Anatolii Marushchak; Stanislav Petrov; Anayit Khoperiya - Journal: Frontiers in Artificial Intelligence - https://pmc.ncbi.nlm.nih.gov/articles/PMC11747593/