

How does the deployment of AI filters to obscure state propaganda reshape public accountability and alter the power structures governing information control?

August 1, 2026 | SnugLab Research

Executive Summary

The deployment of AI filters to obscure state propaganda fundamentally reshapes public accountability by degrading the shared evidentiary baseline, making it increasingly difficult for citizens to independently verify information. This erosion of trust, often termed the "liar's dividend," allows both state and corporate actors to evade scrutiny, as the very foundation of verifiable truth is undermined [3, 6, 17]. Simultaneously, these filters alter power structures by concentrating information control in large technology corporations, which leverage proprietary code, economic dominance, and self-regulation to act as indispensable gatekeepers, often subordinating state power to corporate interests [12, 13]. While legislative efforts like the EU AI Act aim for transparency, compliance complexity and proprietary data advantages tend to entrench these firms, rather than redistribute power to citizens [1, 13]. The evidence suggests both the epistemic crisis and institutional consolidation operate simultaneously and reinforce each other, rather than one being primary, leading to a system where technocratic efficiency legitimizes opaque control over democratic legibility [10, 13].

Key Findings

The Erosion of Public Accountability Through the "Liar's Dividend"

The deployment of AI filters significantly reshapes public accountability by creating an epistemic crisis where the ability to distinguish authentic evidence from synthetic propaganda is severely degraded. This phenomenon, known as the "liar's dividend," arises when public awareness of fabricated AI content fosters blanket skepticism, allowing malign actors to dismiss authentic evidence as fake [3, 6, 17]. Consequently, even when citizens attempt to verify information, the proliferation of synthetic media and LLM-generated authoritative-seeming sources deepens informational uncertainty [3]. This erosion of trust is widespread, with 76% of Americans concerned about AI producing false information, and 77% distrusting both businesses and government agencies to use

AI responsibly [4].

The impact on accountability is measurable:

- During Argentina's 2023 presidential elections, politicians dismissed scandalous audio recordings as potentially fabricated using artificial intelligence [19].
- In legal contexts, a "deepfake defense" has been successfully employed to challenge authentic audiovisual evidence [8].
- Survey experiments with over 15,000 American adults indicate that false claims of misinformation increase politician support across partisan subgroups after a scandal, outperforming silence or apologies [2, 6]. A March 2026 study confirmed these false deepfake claims yield higher perceived leadership abilities, even when responding to video evidence, mediated by participants misidentifying authentic footage as synthetic [21].
- An experimental study found that exposing American participants to deepfakes of localized infrastructure failures significantly increased distrust in government, with mean distrust scores rising from 3.19 in the control group to 3.48 in the deepfake condition [20].
- Public trust in the federal government remains low, with only 22% of Americans expressing trust as of April 2024 [23], and just 31% reporting moderate trust in media in a 2024 Gallup poll [22].

This degradation of the shared evidentiary baseline undermines traditional accountability mechanisms, as the very foundation for verification is compromised [6].

Institutional Consolidation and Corporate Control Over Information

The deployment of AI filters fundamentally alters power structures by concentrating information control in large technology corporations, which act as indispensable gatekeepers. These firms, including Alphabet, Apple, Amazon, Meta/Facebook, and Microsoft, dominate algorithmic activity by owning the means of algorithmic production, training, and deployment [12, 13]. Their economic power, derived from user data extraction and funding of machine learning software libraries and proprietary hardware, allows them to shape regulatory standards and often subordinate state power to corporate interests through lobbying for grants or relaxed regulations [13].

Key aspects of this institutional shift include:

- **Proprietary Code and Opaque Moderation:** Content moderation practices are largely shaped by Silicon Valley elites through self-regulation, prioritizing engagement

metrics over democratic transparency and exposing individual rights to corporate discretion [1, 7]. AI-driven moderation decisions frequently lack context sensitivity and provide minimal explanations for content removal, leaving citizens unable to challenge or understand the algorithmic decisions affecting their information access [7].

- **Outsourcing Censorship:** This institutional pathway lowers the transaction costs of state-corporate compliance, allowing governments to outsource censorship and information control to private platforms that operate with minimal public oversight [12, 18].
- **Algorithmic Power as a Dimension of Influence:** "Algorithmic power" has emerged as a distinct dimension of international influence, shaping political outcomes through mediated information channels, independent of traditional military or economic metrics [10].

Technocratic Efficiency and the Shift to Invisible Architectures

Current deployments of AI filters prioritize technocratic efficiency, which legitimizes opaque power structures and shifts accountability from visible state actors to invisible technical architectures that lack democratic recourse [10, 13, 16]. AI systems are valued for their speed, scale, and proactive capabilities, enabling more precise censorship that prevents content from being posted at all [5, 16, 18].

This shift manifests as:

- **Proactive Censorship:** Predictive filters analyze trends to preemptively throttle engagement or block discussions before they gain momentum, moving censorship from reactive to proactive [16]. This obscures the line between regulation and repression through subtle algorithmic practices like shadow banning and content demotion, undermining transparency [15, 18].
- **Contextual Blindness and User Circumvention:** While aiming for efficiency, AI censorship struggles with evolving tactics such as homophones, emojis, creative misspellings, satire, and memes, as netizens continuously adapt to bypass filters [16]. Human moderation remains crucial in hybrid models due to AI's lack of nuanced cultural context understanding [16].
- **Exportation of Censorship Architectures:** China, for example, exports its AI censorship model to other authoritarian regimes (e.g., Iran, Venezuela, Ethiopia), fostering a global trend of AI-assisted authoritarianism [16]. This involves coordinated integration of government-owned AI firms with private platforms, where regulatory frameworks mandate content suppression and promotion of state-approved narratives

[16]. A 2026 study found that Chinese state-linked content appeared frequently in major open-source training datasets, causing commercial models prompted in Chinese to give answers more favorable to China 75.3% of the time compared to English prompts [2].

Transparency Mandates Entrench Gatekeepers Rather Than Redistribute Power

Despite legislative efforts like the EU AI Act, which became enforceable in 2026 and mandates transparency for generative AI and disclosure of AI-generated content [1, 6, 8, 11, 13], these requirements tend to entrench large tech firms as indispensable gatekeepers rather than redistribute power to citizens. Compliance complexity and proprietary data advantages allow these firms to satisfy legal thresholds while functionally controlling information ecosystems [12, 13].

- **Proprietary Data Advantages:** Major tech firms maintain their gatekeeper status through proprietary data, owning the means of algorithmic production, and controlling essential hardware and software libraries for large-scale AI training [13].
- **Self-Regulation and Bias:** An over-reliance on self-regulation by private companies has exposed individual rights and concentrated information control within Silicon Valley elites [1]. Meta's AI moderation practices, for instance, face biases and underinvestment in non-English markets, particularly affecting the Global South [7].
- **Limited Audit Trails:** While California's AI Transparency Act requires generative AI developers to enable content provenance and provide disclosure options [8], vague disclosures risk blurring the distinction between truthful and false content. Public understanding depends on discerning specific elements of AI generation rather than receiving comprehensive audit trails [9].
- **Government Oversight Transfer:** The U.S. General Services Administration's proposed procurement clause, "Basic Safeguarding of Artificial Intelligence Systems" (GSAR 552.239-7001), published on March 6, 2026, aims to transfer oversight *from* corporate architectural control *to* public bodies by overriding vendor discretionary policies and safety guardrails with government-imposed standards [13, 14]. However, the broader trend remains institutional consolidation within private entities.

Algorithmically Insulated Ecosystems Fracture Legitimacy

Algorithmically insulated information ecosystems causally fracture shared realities, which directly weakens domestic institutional legitimacy required for governance accountability.

Social media algorithms, optimized for engagement, systematically amplify sensational and AI-generated disinformation over factual content [6]. This structural design, coupled with the "liar's dividend," degrades the shared evidentiary baseline necessary for public discourse and trust [3, 6, 9, 17].

The consequences for institutional legitimacy include:

- **Erosion of Trust in Government:** A Gallup-Bentley University survey found that 77% of Americans distrust both businesses and government agencies to use AI responsibly, and 62% have little or no confidence in federal agencies to regulate AI effectively [4].
- **Fractured Realities:** The algorithmic prioritization of engagement and corporate monetization strategies leads to fragmented information environments, undermining the shared factual baselines and public trust necessary for citizens to hold domestic institutions accountable [3, 6, 13].
- **Impact on Political Discourse:** During the 2022 invasion of Ukraine, AI-driven bot accounts comprised 60% to 80% of posts using specific pro-Russian hashtags [3, 9]. This shift included a 4.7 percentage point increase in the likelihood of users prioritizing US Republican policy issues and a 7.4 percentage point decrease in the likelihood of viewing Ukrainian President Volodymyr Zelenskyy positively [2, 6, 7]. The algorithm systematically altered visibility, increasing the share of right-wing content by 2.9 percentage points and reducing the visibility of external links related to the Russia-Ukraine war by up to a factor of eight [2, 4].

Implications

The deployment of AI filters to obscure state propaganda has profound implications for democratic societies and global information governance. For citizens, the primary implication is a diminished capacity for informed decision-making and holding power accountable, as the very concept of verifiable truth becomes increasingly elusive due to the "liar's dividend" [3, 6, 17]. This necessitates a fundamental re-evaluation of how individuals engage with and trust digital information.

For governments, the implications are twofold: while AI filters offer new tools for information control and narrative management, particularly for authoritarian regimes exporting censorship models [16], they also face a growing crisis of domestic institutional legitimacy as shared realities fracture and public trust erodes [4, 21]. The reliance on

private technology firms for content moderation means that states effectively outsource aspects of their sovereignty and information control to corporate entities driven by economic interests rather than democratic principles [12, 13].

For technology companies, the implications involve navigating increasing regulatory scrutiny, such as the EU AI Act [1, 8], while simultaneously leveraging their proprietary data and algorithmic control to maintain their gatekeeper status [13]. Their business models, optimized for engagement, inadvertently contribute to the spread of disinformation and the erosion of trust, creating a tension between profit motives and societal well-being [6]. The current trajectory suggests a future where information control is increasingly centralized within opaque, technocratic architectures, challenging traditional notions of democratic oversight and public accountability.

Limitations and Caveats

The available research provides a strong conceptual framework and qualitative evidence for the reshaping of public accountability and power structures. However, direct quantitative data for "country=global; entity=AI filters; period=none" is limited. Specific measurable declines in legitimacy indices for named domestic institutions (e.g., electoral commissions, central banks) following AI filter deployment are not provided, nor are statistical correlations with increased policy implementation friction. While the EU AI Act mandates transparency, detailed audit trail requirements specifically for state-aligned content beyond general disclosure are not fully elaborated in the findings. Furthermore, specific AI-driven censorship architectures from companies like Cloudflare, Akamai, or Palantir, and their comparative cost structures for state adoption versus private licensing, are not detailed. The report relies on a source pool where one source ([13]) is cited frequently, though within the prescribed limits, suggesting a need for broader corroboration in some areas.

Sources

[1] [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/>

[2] [edu] Governments May Shape What Ai Chatbots Say By Shaping The We - today.ucsd.edu - <https://today.ucsd.edu/story/governments-may-shape-what-ai-chatbots-say-by-shaping-the-web-they-learn-from>

[3] [edu] Propaganda Foreign Interference And Generative Ai - brookings.edu - <https://www.brookings.edu/articles/propaganda-foreign-interference-and-generative-ai/>

- [4] [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/>
- [5] Repressive Power Artificial Intelligence - freedomhouse.org - <https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence>
- [6] How Ai Weaponises Confusion Fuel Harm Block Accountability - thenewhumanitarian.org - <https://www.thenewhumanitarian.org/opinion/2026/03/26/how-ai-weaponises-confusion-fuel-harm-block-accountability>
- [7] [peer-reviewed] Meta's AI moderation and free speech: Ongoing challenges in the Global South - Authors: Soorya Balendra - Journal: Cambridge Forum on AI: Law and Governance - <https://www.cambridge.org/core/journals/cambridge-forum-on-ai-law-and-governance/article/metas-ai-moderation-and-free-speech-ongoing-challenges-in-the-global-south/2DB952F896DB5744A43CD3E6C1A6DCB4>
- [8] Beyond Section 230 Principles For Ai Governance - harvardlawreview.org - <https://harvardlawreview.org/print/vol-138/beyond-section-230-principles-for-ai-governance/>
- [9] Comment On Disclosure And Transparency Of Artificial Intelli - epic.org - <https://epic.org/documents/comment-on-disclosure-and-transparency-of-artificial-intelligence-generated-content-in-political-advertisements-fcc-request-for-comment/>
- [10] 15.SC SE 16 Ready Final 1 - sci-cult.com - <https://sci-cult.com/wp-content/uploads/2026/03/15.SC-SE-16-Ready-final-1.pdf>
- [11] [gov] IPOL STU(2020)657101 EN - europarl.europa.eu - [https://www.europarl.europa.eu/RegData/etudes/STUD/2020/657101/IPOL_STU\(2020\)657101_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2020/657101/IPOL_STU(2020)657101_EN.pdf)
- [12] Latzer Just 2020 Governance By And Of Algorithms - dev.mediachange.ch - https://dev.mediachange.ch/media/pdf/publications/Latzer_Just_2020_Governance-by-and-of-Algorithms.pdf
- [13] Balch Injustice - andrewbalch.com - https://www.andrewbalch.com/files/Balch_Injustice.pdf
- [14] Policy Orientation Ai Ia.Aspx - international.gc.ca - https://www.international.gc.ca/world-monde/issues_development-enjeux_developpement/human_rights-droits_homme/policy-orientation-ai-ia.aspx?lang=eng
- [15] [peer-reviewed] 397183718 Power Dynamics Of Information Control Censorship A - researchgate.net - AUTHORS UNAVAILABLE - https://www.researchgate.net/publication/397183718_Power_dynamics_of_information_control_censorship_and_digital_transparency_shaping_human_rights_discourse_in_the_information_society
- [16] Download - pioneerpublisher.com - <https://www.pioneerpublisher.com/jrssh/article/download/1223/1121>
- [17] Amp - technologyreview.com - <https://www.technologyreview.com/2023/10/04/1080801/generative-ai-boosting-disinformation-and-propaganda-freedom-house/amp/>
- [18] Power Dynamics Information Control Censorship And Digital Tr - gscarr.gsconlinepress.com - <https://gscarr.gsconlinepress.com/content/power-dynamics-information-control-censorship-and-digital-transparency-shaping-human-rights>
- [19] [peer-reviewed] Article - sciencedirect.com - AUTHORS UNAVAILABLE - <https://www.sciencedirect.com/science/article/pii/S0267364923001036>
- [20] X43ph V1 - osf.io - https://osf.io/preprints/socarxiv/x43ph_v1
- [21] [peer-reviewed] False failures, real distrust: the impact of an infrastructure failure deepfake on government trust - Authors: Saifuddin Ahmed; Muhammad Masood; Adeline Wei Ting Bee; Kei Ichikawa - Journal: Frontiers in Psychology - <https://pmc.ncbi.nlm.nih.gov/articles/PMC12141277/>
- [22] [gov] HHRG 116 IG00 Transcript 20190613 - docs.house.gov - <https://docs.house.gov/meetings/IG/IG00/20190613/109620/HHRG-116-IG00-Transcript-20190613.pdf>
- [23] [edu] Invisible Battle Field Systemic Warfare And The Future Of St - academia.edu - https://www.academia.edu/145343117/Invisible_Battle_Field_Systemic_Warfare_and_the_Future_of_State_Power