

How does the proliferation of AI-generated deepfakes erode institutional trust and destabilize democratic governance by altering the evidentiary basis of political accountability?

August 2, 2026 | SnugLab Research

Executive Summary

The proliferation of AI-generated deepfakes fundamentally erodes institutional trust and destabilizes democratic governance by causing a foundational epistemic rupture in the evidentiary basis of political accountability. This rupture is driven by the "liar's dividend," which allows public figures to dismiss authentic evidence as synthetic, and a pervasive "skepticism tax" that fosters informational nihilism among citizens, making shared verification increasingly difficult and leading to governance paralysis. While various technical and regulatory mitigation efforts are underway, their efficacy is limited by the rapid advancement of deepfake technology, user fatigue, and legal challenges, often triggering a perpetual generation-detection race rather than restoring cognitive trust.

Key Findings

Deepfakes Cause a Foundational Epistemic Rupture in Evidentiary Standards

The 'evidentiary basis of political accountability' must be broadly defined as institutionally mediated evidence, rather than strictly as directly verifiable audiovisual proof, to understand the systemic impact of deepfakes [3, 7, 8, 10]. Deepfakes exploit the "liar's dividend," enabling public figures to dismiss authentic but damaging video or audio evidence by simply claiming it is AI-generated [7, 8, 9]. This dynamic shifts the burden of truth toward institutional mediation, such as media fact-checking and legal standards of proof, which are themselves under systemic pressure [3, 8]. The proliferation of synthetic media creates "synthetic realities" where content and identity are jointly manufactured, leading to a foundational epistemic rupture rather than merely adding manageable noise [10]. This rupture is characterized by escalating verification costs, making political disagreement epistemically irresolvable and potentially leading to governance paralysis [10].

Pervasive Synthetic Media Erodes Trust and Amplifies Polarization

Exposure to synthetic media imposes a cognitive "skepticism tax" on news consumption, fostering informational nihilism where citizens lose confidence in both true and false content [9]. This generalized suspicion disproportionately impacts politically moderate individuals, leading to "reality fatigue" and disengagement from civic life [9, 10].

Simultaneously, deepfakes amplify pre-existing partisan polarization by nudging different political groups into incompatible interpretive frames, each supported by seemingly credible synthetic evidence [10, 17]. This shifts public confidence away from broad institutions toward closed interpersonal networks and partisan communities [9, 11]. Both the skepticism tax and amplified polarization undermine the shared reality necessary for political accountability [7, 8, 9, 10].

The "Liar's Dividend" Decouples Actions from Accountability

The strategic weaponization of the "liar's dividend" fundamentally alters the evidentiary basis of political accountability by making strategic denial more efficient than institutional verification [10]. Public figures can exploit the mere possibility of deepfakes to dismiss authentic but damaging evidence, increasing the payoff for delay and denial even when allegations are true [7, 8, 9]. This systematically outpaces verification costs, forcing citizens and institutions to navigate a "synthetic reality" where shared evidence is scarce and any claim can be plausibly contested [10]. This dynamic erodes institutional trust, fuels polarization, and creates conditions for governance paralysis, particularly during crises [1, 10]. Examples include Elon Musk's legal team suggesting past statements could have been deepfakes [3, 8] and an Indian politician dismissing authentic embarrassing audio as AI-generated [8]. Deepfake content was also deployed in Utah elections prior to October 2024 to cast doubt on political realities [4].

Algorithms Convert Deepfakes into Self-Reinforcing "Synthetic Realities"

Platform algorithms actively transform isolated deepfake artifacts into self-reinforcing "synthetic realities" by prioritizing emotionally engaging, novel, and identity-affirming content over accurate reporting [7, 13]. This algorithmic curation creates personalized information environments where synthetic content, digital identities, and social interactions are jointly manufactured and mutually reinforcing [10]. This shifts the epistemic burden onto citizens, imposing a continuous "skepticism tax" that forces

individuals to constantly question what they see and hear [9]. The resulting informational nihilism and reality fatigue cause citizens to disengage from civic and informational life, making it significantly more expensive and difficult to verify political actions at scale [9, 10].

Mitigation Efforts Face Significant Challenges

While various policies and frameworks have been implemented, their documented efficacy in stabilizing democratic governance and restoring evidentiary credibility remains limited.

- **Regulatory Frameworks:** The EU AI Act, which becomes applicable on August 2, 2026, mandates machine-readable marking of AI-generated deepfakes [20, 22]. The EU Digital Services Act (DSA), fully applicable from February 2024, mandates transparency and deepfake labeling for Very Large Online Platforms [5, 13, 14, 15]. However, enforcing these regulations is difficult due to the rapid evolution of deepfake technology, which makes reliable detection technically challenging and costly [22, 24].
- **Platform Policies:** Meta requires political advertisements and organic AI-generated content to be labeled [26, 27]. Public opinion research across 13 countries indicates that 82% of respondents favor warning labels for AI-generated content depicting people saying things they did not say [26]. However, labeling news headlines as AI-generated can paradoxically diminish belief in both true and false headlines, eroding general confidence in media evidence [21].
- **Provenance Standards:** Cryptographic provenance standards like the Coalition for Content Provenance and Authenticity (C2PA) embed tamperproof metadata to prove content authenticity [19, 23]. While promising, widespread adoption is a "long road," and sophisticated actors can still exploit these techniques [25].
- **Legal and Technical Hurdles:** Mandatory disclosure requirements and technical compliance mechanisms often prioritize bureaucratic verification over genuine cognitive trust [9]. Legislative efforts, such as California's AB 2655 mandating labels for synthetic media in elections, have been overturned in federal court as First Amendment violations [25]. This creates a perpetual "race between generation and detection" that cannot be fully delegated to automated systems [15].

Deepfakes Lead to Institutional Paralysis

The transition from epistemic uncertainty to institutional paralysis is evident in

governmental responses. Executive Order 14365, "Ensuring a National Policy Framework for Artificial Intelligence" (December 11, 2025), was issued to address regulatory gridlock caused by conflicting state-level AI laws and created an AI Litigation Task Force to challenge inconsistent state statutes [1, 2, 5, 9, 11, 12]. Similarly, the FCC Ruling on AI Voices in Robocalls (February 2024) directly resulted from a January 2024 deepfake robocall impersonating President Biden [18]. Judicial bodies also struggle, with courts encountering the "deepfake defense" where real audiovisual evidence is challenged as AI-generated, altering traditional evidentiary standards [8]. The U.S. Judicial Conference's Advisory Committee on Evidence Rules declined to amend Rule 901 in May 2025 despite recognizing the problem [6, 16]. These instances demonstrate how deepfakes outpace existing regulatory and legal frameworks, leading to under-resourced and fragmented justice systems [2, 4, 17].

Implications

The proliferation of AI-generated deepfakes has profound implications for democratic governance and political accountability. The shift from directly verifiable audiovisual proof to institutionally mediated evidence means that the burden of truth increasingly falls on institutions already struggling with public trust. This environment empowers political actors to strategically deny authentic but damaging evidence, leading to a breakdown in accountability and potential governance paralysis, especially during crises. The cognitive "skepticism tax" and reality fatigue among citizens threaten civic engagement and the informed public discourse essential for democracy. While technical solutions like provenance standards and regulatory mandates are being implemented, their effectiveness is hampered by the rapid evolution of AI, the inherent limitations of detection, and legal challenges. This suggests a future where the evidentiary basis of political accountability remains contested, requiring continuous adaptation from democratic institutions and a renewed focus on fostering critical media literacy and epistemic vigilance among citizens.

Limitations and Caveats

Direct quantitative economic figures comparing the cost of continuous verification fatigue for citizens against the financial savings from delayed institutional responses in high-profile deepfake controversies are not explicitly available in the provided research. Similarly, specific comparative efficacy rates of cryptographic provenance standards

versus algorithmic curation in reducing the "skepticism tax" among moderate voters aged 35-65 are not quantified. While the research identifies judicial bodies and electoral processes as key institutions facing evidentiary divergence, detailed post-2025 data on this divergence is limited. The impact of deepfakes on government trust can also vary by demographic factors like education level and national context; for example, exposure to deepfakes depicting infrastructure failures heightened distrust among American participants but not Singaporeans, possibly due to Singapore's higher public trust and familiarity with its surroundings [1].

Sources

- [1] [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/>
- [2] [edu] Deepfake Warnings For Political Videos Increase Disbelief But Do Not Improve - <https://isps.yale.edu/events/2021/01/deepfake-warnings-for-political-videos-increase-disbelief-but-do-not-improve>
- [3] [edu] View - journals.library.columbia.edu - <https://journals.library.columbia.edu/index.php/stlr/blog/view/669>
- [4] [edu] Research Summary - uvu.edu - <https://www.uvu.edu/news/2024/10/media/research-summary.pdf>
- [5] [edu] Deepfake Scams Poisoned Chatbots - cetas.turing.ac.uk - <https://cetas.turing.ac.uk/publications/deepfake-scams-poisoned-chatbots>
- [6] The Nature And Risks Of Ai Generated Deepfakes In Political Communications - citizen.org - <https://www.citizen.org/article/the-nature-and-risks-of-ai-generated-deepfakes-in-political-communications/>
- [7] 5.04: The Eye Of The Storm Deepfakes AI And The Crisis Of Democratic Trust - socialsci.libretexts.org - https://socialsci.libretexts.org/Courses/University_of_Illinois_Springfield/Digital_Media_Literacy_in_the_Age_of_Artificial_Intelligence%3A_Navigating_Truth_Power_and_Misinformation_v2/05%3A_The_Technology_Trap-_AI_Surveillance_Capitalism_and_the_Architecture_of_Reality/5.04%3A_The_Eye_of_the_Storm-_Deepfakes_AI_and_the_Crisis_of_Democratic_Trust
- [8] Deepfakes Elections And Shrinking Liars Dividend - brennancenter.org - <https://www.brennancenter.org/our-work/research-reports/deepfakes-elections-and-shrinking-liars-dividend>
- [9] [peer-reviewed] Synthetic media, political disinformation, and the erosion of public trust: a systematic review and synthetic epistemic destabilization model - Authors: Alamin, Mohanad; Alrached, Saleh Abied Abied; Abasher, Mathani Hassan Hassan; Mahmoud, Khalid Osman; Altaher, Abdulla Ebrahim; Mohamed, Elsir Ali Saad Ali Saad - Journal: Frontiers in Political Science - <https://www.frontiersin.org/journals/political-science/articles/10.3389/fpos.2026.1811974/full>
- [10] [preprint] HTML - arxiv.org - AUTHORS UNAVAILABLE - <https://arxiv.org/html/2601.00306v1>
- [11] Deepfakes And Crisis Knowing - unesco.org - <https://www.unesco.org/en/articles/deepfakes-and-crisis-knowing>
- [12] [peer-reviewed] Stable - jstor.org - AUTHORS UNAVAILABLE - <https://www.jstor.org/stable/27384397>
- [13] Final Deepfakes Elections Report Civicus DDI Research 2025 - civicus.org - <https://www.civicus.org/documents/ddi/final-deepfakes-elections-report-civicus-ddi-research-2025.pdf>
- [14] How Does AI Impact Trust In Democracy - digitalpublicsquare.org - <https://digitalpublicsquare.org/insights/how-does-ai-impact-trust-in-democracy/>
- [15] Epistemic Vigilance Truth Illusion Misinformation - imaginingthefuture.org - <https://imaginingthefuture.org/reports-and-publications/human-resilience-in-the-age-of-ai/epistemic-vigilance-truth-illusion-misinformation/>

- [16] [peer-reviewed] Article - sciencedirect.com - AUTHORS UNAVAILABLE - <https://www.sciencedirect.com/science/article/pii/S2444569X25001271>
- [17] [peer-reviewed] 398735896 The Impact Of Deepfake Technology On Elections And - researchgate.net - AUTHORS UNAVAILABLE - https://www.researchgate.net/publication/398735896_The_Impact_of_Deepfake_Technology_on_Elections_and_Democratic_Trust
- [18] [edu] Ai Regulation Wrapped - sites.wustl.edu - <https://sites.wustl.edu/westcoastclub/ai-regulation-wrapped/>
- [19] [gov] CSI CONTENT CREDENTIALS.PDF - media.defense.gov - <https://media.defense.gov/2025/Jan/29/2003634788/-1/-1/0/CSI-CONTENT-CREDENTIALS.PDF>
- [20] [peer-reviewed] AI-driven disinformation: policy recommendations for democratic resilience - Authors: Alexander Romanishyn; Olena Malyska; Vitaliy Goncharuk - Journal: Frontiers in Artificial Intelligence - <https://pmc.ncbi.nlm.nih.gov/articles/PMC12351547/>
- [21] [edu] AI Policy Labeling - computing.mit.edu - https://computing.mit.edu/wp-content/uploads/2023/11/AI-Policy_Labeling.pdf
- [22] [preprint] Html - arxiv.org - AUTHORS UNAVAILABLE - <https://arxiv.org/html/2507.08879v1>
- [23] c2pa.org - <https://c2pa.org/>
- [24] [peer-reviewed] Beyond phase-in: assessing impacts on disinformation of the EU Digital Services Act - Authors: Nannini, Luca; Bonel, Eleonora; Bassi, Davide; Maggini, Michele Joshua - Journal: AI and Ethics - <https://link.springer.com/article/10.1007/s43681-024-00467-w>
- [25] Synthetic Media Policy Provenance And Authentication Expert - techpolicy.press - <https://www.techpolicy.press/synthetic-media-policy-provenance-and-authentication-expert-insights-and-questions/>
- [26] Amp - about.fb.com - <https://about.fb.com/news/2024/04/metaspaces-approach-to-labeling-ai-generated-content-and-manipulated-media/amp/>
- [27] Meta Label Ai Generated Deepfake Political Ads 2024 Election - fortune.com - <https://fortune.com/2023/11/08/meta-label-ai-generated-deepfake-political-ads-2024-election/>