Deepfakes in Political Manipulation: Evaluating Risks Under the AI Act

Mst Rafia Islam[1,2*] **and Azmine Toushik Wasi**[1]
[1]Computational Intelligence and Operations Laboratory (CIOL)
[2]Independent University, Bangladesh
Correspondence to: islam.mst.rafia@gmail.com

Abstract

Deepfake technology poses a major threat to democratic processes, particularly elections. In the USA, deepfakes were expected to influence the 2024 presidential election, while in India they were used during the 2024 general elections to sway voters. European countries like Poland and Slovakia also experienced deepfake-driven campaigns undermining political narratives. These cases highlight how low-cost AI tools can industrialize disinformation at an unprecedented scale. This paper examines political deepfakes under the AI Act's systemic risk framework, evaluating intent, scale, and impact, and proposes AI detection systems and regulatory measures to enhance transparency and fairness. By combining case studies with regulatory theory, it emphasizes the urgent need for legal, technological, and social countermeasures.

1 Introduction

In an era where digital technology shapes political engagement, AI has become both a tool and a weapon in elections [3]. Deepfakes, AI-generated media that convincingly imitate real people, raise concerns about eroding trust in political communication [22]. They threaten the epistemic foundation of democracy by making it harder for citizens to distinguish authentic information from manipulation [24]. While useful in entertainment, education, and accessibility, deepfakes in political campaigns can manipulate voter perceptions, spread disinformation, and disrupt democratic institutions. This *liar's dividend* further allows malicious actors to dismiss authentic evidence as fabricated, undermining institutional trust [18].

The 2024 elections in multiple democracies illustrated how deepfakes could be weaponized [9]. In the United States, deepfake-generated robocalls falsely imitated the voices of political candidates, discouraging voters from participating in the electoral process. In India, deepfake videos created deceptive endorsements, altering the way political speeches were perceived [9]. Meanwhile, in European nations such as Poland and Slovakia, deepfake-driven misinformation campaigns sought to undermine political leaders and shift public opinion [19]. These cases also highlight temporal tactics, late-night or eve-of-vote releases, designed to maximize exposure before corrections ("exposure-before-correction" effect) [27]. Unlike traditional propaganda, deepfakes exploit the psychological weight of audiovisual content. Research shows that humans are more likely to trust video evidence than text, creating a "truth bias" that amplifies the persuasive power of manipulated media [24]. This epistemic vulnerability makes deepfakes a systemic risk for democracy. Controlled experiments confirm that audiovisual misinformation has a stronger and longer-lasting effect on attitudes than text-based misinformation, even after viewers are told the material was fake [5].

Given the magnitude of these threats, the European Union has taken steps toward regulating AI-based risks through the AI Act, a landmark legislation designed to address AI applications that pose systemic risks [23]. The Act categorizes AI technologies based on their potential harm, with deepfakes likely falling under high-risk or systemic-risk classifications due to their impact on electoral

Workshop on Regulatable ML at the 39th Conference on Neural Information Processing Systems (NeurIPS 2025).

integrity. This paper explores how deepfakes in political manipulation should be evaluated under the AI Act, examining their intent, scale, and consequences. Furthermore, it discusses policy interventions, detection mechanisms, and regulatory measures that could help mitigate the risks posed by AI-generated disinformation.

2 Rise of Deepfakes in Political Manipulation

Technical drivers. Modern deepfakes are driven by key technical shifts. Generative adversarial networks and diffusion models enable photorealistic frames with temporal consistency [2]. High-quality voice cloning and prosody transfer can now be trained on short samples, lowering data thresholds and accelerating attacks [14]. Modular toolchains allow lip-sync, face reenactment, and audio synthesis to be combined into end-to-end pipelines usable by non-experts [15], compressing the time from concept to release and increasing attack volume and variety [22]. Zero-shot TTS systems generate convincing voices from under 10 seconds of audio, easily scraped from public speeches or interviews [25]. We outline a taxonomy of political deepfake attacks. Face-forward clips mimic direct-to-camera addresses using micro-expression modeling [1]. Crowd-context clips simulate off-axis angles and background noise to mask artifacts [2]. Audio-first attacks exploit telephony and voice notes, where compression hides synthesis traces [7]. Cross-modal attacks combine text prompts with video templates, producing multiple region-specific variants from a single script [14]. Each type stresses different detection components, explaining uneven field performance [15]. Cross-modal pipelines appeared in India's 2024 elections, with identical scripts repurposed in Hindi, Bengali, and Tamil using cloned candidate voices [9].

Socio-technical enablers. Three socio-technical dynamics heighten political risk. *Algorithmic amplification* favors novelty and emotional salience, letting high-arousal fakes outpace verification [3]. *Micro-targeting* tailors content to demographics with local dialects and issues, increasing plausibility and reducing cross-audience correction [15]. *Coordination* has improved via off-the-shelf scheduling and bot orchestration, enabling synchronized releases around debates or rulings [15]. Twitter/X data show false news spreads faster and deeper than true news, amplifying synthetic media's impact [26]. Detection advances exist but are fragile. Frequency-domain artifacts and reconstruction-error fingerprints improve held-out performance [7], and cross-modal consistency checks add resilience. Yet adversaries adapt through re-encoding, noise injection, or fine-tuning generators on detector losses [22], creating an arms race where benchmark gains often fail in adversarial contexts. Policy must therefore complement detection with provenance and accountability layers [24].

3 Case Studies: Deepfakes in the 2024 Elections

Deepfake Robocalls and Misleading Campaigns in United States The United States case shows how synthetic audio exploits legacy infrastructure. Robocalls are cheap, hard to pre-screen, and geo-targetable [21]. Combined with voice cloning, a single message can reach tens of thousands of voters within hours, leveraging the authenticity of voice and the urgency of get-out-the-vote periods [21]. In January 2024, New Hampshire voters received a cloned Biden robocall instructing abstention; regulators later ruled AI robocalls illegal under the FCC's TCPA [21].

Video deepfakes further strained platforms and fact-checkers, from subtle edits reframing events to fully synthetic statements [5]. Even when flagged, such content often achieved millions of impressions, and experimental studies show audiovisual misinformation can create lasting belief effects, raising social costs of delayed moderation [5]. This example highlights two risks: voter suppression or mobilization via false instructions [3], and reputational damage that shifts agendas and consumes debunking resources [21]. Scholars now recommend treating deepfake robocalls as both disinformation and unlawful abuse for faster legal intervention [16].

$\Rightarrow$ *Operational lessons*. Authorities should treat synthetic robocalls as disinformation and telephony abuse, implying rapid call-trace protocols with carriers [21], authenticated caller ID for political messaging [16], and public advisories communicating specific manipulation tactics [21].

Deepfake Videos in Multilingual Election Campaigns in India India's linguistic diversity created ideal conditions for localized manipulation. Fabricated endorsements and translated speeches appeared in regional dialects that conferred identity proximity and trust [1]. Messaging platforms increased velocity because closed groups limit cross-community correction and reduce the probability of early fact-checker visibility [9].

This Indian case illustrates how cultural matching strengthens perceived authenticity. Accurate lip-sync, regionally appropriate idioms, and local visual backdrops create coherence that resists quick

debunking [9]. Because many constituencies consume video on low-end devices with aggressive compression, artifacts may be invisible on such mobile streams [15]. These constraints matter for detector deployment and for media-literacy design [15]. Fact-checking organizations such as AltNews reported a surge in regional deepfakes during the 2024 election cycle, underscoring the difficulty of multilingual detection at scale.
$\nRightarrow$ *Operational lessons*. Parties and election commissions should pre-record and pre-register canonical speeches in multiple dialects [7], publish signed hashes through official channels [15], and coordinate with messaging platforms to provide rapid provenance checks to high-reach WhatsApp community admins during the election silence period [6].

Deepfakes and Political Disinformation in Poland and Slovakia (Europe) European campaigns showed careful timing around debates and investigative reporting cycles [10]. Synthetic clips were released shortly before major televised events, which constrained the time window for verification and allowed narratives to set before corrections surfaced [19]. Investigations reported cross-border infrastructure, which complicates jurisdiction and slows takedown requests. This environment reveals the regulatory complexity of attribution. Even when a platform removes the content for deception, the responsible entity may sit outside national jurisdiction or operate through cutouts [19]. That gap makes systemic risk a useful lens, since the harm is generated by coupling synthetic media with cross-border reach and synchronized release, not only by the realism of a single clip [10].
$\nRightarrow$ *Operational lessons*. National regulators should maintain cross-border memoranda on expedited data preservation and disclosure for election-period deception cases [19]. They should also coordinate with broadcasters so that debate organizers can display live verification cues when provenance checks are inconclusive but risk signals are high [23].

4 Evaluating Deepfakes Under the AI Act

The European Union's AI Act adopts a risk-based framework for regulating artificial intelligence, classifying applications by their potential to cause harm [23]. Political deepfakes pose a unique challenge because they intersect with electoral integrity, free speech, disinformation, and national security [13]. Assessment of deepfakes under the AI Act centers on three criteria: intent, scale, and impact [5]:
✓✓ **Intent:** Indicators include deceptive framing ("real" presentation) [13], impersonation of a natural person without disclosure [8], and targeting during restricted election periods [23]. Where satire or artistic use is claimed, disclosure clarity [8], watermark presence [13], and channel context should be evaluated.
$\rightarrowtail$ *Enforcement lever:* transparency obligations for AI-generated content and penalties that escalate when impersonation produces foreseeable voter harm [13].
✓✓ **Scale:** Indicators include bot-amplified reach, synchronized multi-platform release, telephony broadcast volume [23], and micro-targeting depth by demographic or geography [15].
$\rightarrowtail$ *Enforcement lever:* systemic-risk designation when scale indicators exceed thresholds, which triggers risk-management duties [10], auditability, and crisis protocols for providers and deployers.
✓✓ **Impact:** Indicators include voter suppression or mobilization signals [4], poll-measurable attitude shifts [5], localized unrest [14], and measurable reductions in trust toward verified institutions.
$\rightarrowtail$ *Enforcement lever:* sanctions that consider downstream harms [4], the availability of timely corrections, and remedies such as funding for local fact-checking capacity.

Classification under the AI Act. Political deepfakes that involve impersonation for electoral influence should be treated as high-risk at minimum [10]. They should also be classified as systemic when scale and cross-border coordination are present [13]. Because detection is imperfect, compliance cannot depend on ex-post identification alone [23]. Providers and deployers should therefore be obligated to implement provenance and disclosure by default, and platforms should maintain incident response playbooks for election periods. Articles 51–55 of the AI Act impose "systemic risk" obligations on general-purpose AI providers, including adversarial testing and incident reporting — mechanisms directly relevant for deepfake misuse in elections.

5 Regulatory and Technological Solutions

Mitigating deepfake risks requires a multi-layered approach involving technology, law, and society.
1. Detection and forensics.
$\dashrightarrow$ *Multi-view detection.* Combine spatial artifacts [2], temporal coherence checks, and audio-visual alignment tests, and fuse scores with provenance signals when available [22].
$\dashrightarrow$ *Adversarial robustness.* Train detectors on re-encodes, compression, speed-ups, and pitch-shifts

that adversaries use to defeat fingerprints [1].
⇢ *Field deployment.* Push lightweight models to client devices used by journalists and election officials, with a simple traffic-light interface for triage rather than binary truth claims [22].
⇢ *Continuous evaluation.* Maintain red-team programs that simulate release strategies and report time-to-detection and exposure before takedown. Benchmarks such as the DeepFake Detection Challenge (DFDC) and FaceForensics++ show good in-lab accuracy, but detection rates drop sharply in adversarial real-world settings, requiring hybrid provenance + detection approaches [17].
⇢ *Limitation.* Detectors produce probabilistic outputs and are sensitive to domain shift [14]. Policy should treat them as triage tools that inform response, not as sole arbiters of authenticity [24].

2. Watermarking and provenance.
⇢ *C2PA-style provenance.* Encourage signing at capture [7], maintain cryptographic chains through edits [8], and surface provenance badges in UIs without revealing sensitive metadata to adversaries.
⇢ *Labeling.* Require clear textual disclosure when AI generates or materially edits a political communication [8], with penalties for removal or obfuscation.
⇢ *Resilience.* Assume watermarks can be stripped. Pair provenance with rapid public verification channels so that officials and journalists can request hashes or signed originals when a clip trends [8]. Google's SynthID and Meta's AudioSeal exemplify emerging watermarking tools, though research shows such marks can often be removed via model fine-tuning or audio transformations.

3. Legal accountability.
⇢ *Impersonation harms.* Criminalize knowing distribution of deceptive impersonations intended to affect voting or public order, with aggravated penalties for coordinated operations [8]. The proposed No Fakes Act in the United States is a notable example, aiming to prohibit unauthorized digital replicas of individuals' voices or likenesses when used for deceptive or exploitative purposes. Such provisions could serve as a model for election-related protections.
⇢ *Civil remedies.* Provide fast-track takedown and injunctive relief for candidates and journalists who are targeted, and permit statutory damages that scale with reach [23]. The TAKE IT DOWN Act, though primarily targeting non-consensual intimate imagery, establishes a precedent for obligating platforms to rapidly remove harmful synthetic content, a principle that could be extended to political deepfakes during elections.
⇢ *Telephony-specific rules.* Require authenticated caller identification for political robocalls [10], with carrier obligations to block repeat violators. The FCC's 2024 clarification that AI-generated robocalls are illegal under the TCPA provides a template for medium-specific governance.

4. Platform responsibility.
⇢ *Pre-election posture.* Freeze certain recommendation tests [3], stand up 24-hour policy teams, and publish high-risk manipulation categories with reporting channels for election bodies and newsrooms.
⇢ *Content flows.* Down-rank items with low-provenance confidence during peak election windows [15], and attach friction for rapidly spreading political clips that fail lineage checks.
⇢ *Third-party oversight.* Commission independent audits of detection coverage and false-positive rates on political samples, and publish incident statistics after the cycle. Under the Digital Services Act, Very Large Online Platforms (VLOPs) already face systemic-risk duties for election integrity, which could complement AI Act Article 50 disclosure rules.

5. Public awareness and resilience.
⇢ *Prebunking.* Run short, platform-specific explainers on common deepfake tactics before election silence periods [4].
⇢ *Community pathways.* Equip local civil society and newsroom partners with simple authenticity request tools and verified contact points at platforms.
⇢ *Measurement.* Track changes in trust and correction uptake to refine outreach and to avoid over-seeding skepticism that would harm legitimate journalism [5]. Field experiments by Google Jigsaw show prebunking videos improve user resistance to misinformation during elections [11].

6 Concluding Remarks

Deepfake threatens electoral integrity, as seen in the elections in the US, India, and Europe [19], requiring urgent intervention. The EU's AI Act offers a framework for risk assessment and enforcement [23]. Safeguarding democracy demands AI detection, legal accountability, platform responsibility, and public education [12, 24]. Resilience should be measured by reduced exposure-before-correction and betetr public trust, aligning with AI Act goals and democratic health [20].

References

[1] Samer H Al-Khazraji, Hassan Hadi Saleh, Adil I Khalid, and Israa Adnan Mishkhal. Impact of deepfake technology on social media: Detection, misinformation and societal implications. *The Eurasia Proceedings of Science Technology Engineering and Mathematics*, 23:429–441, 2023.

[2] Upasana Bisht and Pooja. Evolving deepfake technologies: Advancements, detection techniques, and societal impact. 1(2):38–43, February 2025.

[3] Herbert Chang, Benjamin Shaman, Yung-Chun Chen, Mingyue Zha, Sean Noh, Chiyu Wei, Tracy Weener, and Maya Magee. Generative memesis: Ai mediates political information in the 2024 united states presidential election. 2025.

[4] Nicholas Diakopoulos and Deborah Johnson. Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media Soc.*, 23(7):2072–2098, July 2021.

[5] Tom Dobber, Nadia Metoui, Damian Trilling, Natali Helberger, and Claes de Vreese. Do (microtargeted) deepfakes have real effects on political attitudes? *Int. J. Press Polit.*, 26(1):69–91, January 2021.

[6] Max-Paul Förster, Luca Deck, Raimund Weidlich, and Niklas Kühl. A multi-level strategy for deepfake content moderation under EU regulation. 2025.

[7] Masabah Bint E Islam, Muhammad Haseeb, Hina Batool, Nasir Ahtasham, and Zia Muhammad. AI threats to politics, elections, and democracy: A blockchain-based deepfake authenticity verification framework. *Blockchains*, 2(4):458–481, November 2024.

[8] Catherine Jasserand. Deceptive deepfakes: Is the law coping with AI-altered representations of ourselves? In *2024 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–4. IEEE, September 2024.

[9] Shweta Kashyap. Deepfake cases in the 2024 lok sabha election: Impact and implications for democracy. *International Journal For Multidisciplinary Research*, 6(5), September 2024.

[10] Mateusz Łabuz. Regulating deep fakes in the artificial intelligence act. *Applied Cybersecurity & Internet Governance*, 2(1), December 2023.

[11] Stephan Lewandowsky, Ullrich K H Ecker, John Cook, Sander van der Linden, Jon Roozenbeek, and Naomi Oreskes. Misinformation and the epistemic integrity of democracy. *Curr. Opin. Psychol.*, 54(101711):101711, December 2023.

[12] Hanzhe Li, Jiaran Zhou, Yuezun Li, Baoyuan Wu, Bin Li, and Junyu Dong. FreqBlender: Enhancing DeepFake detection by blending frequency knowledge. 2024.

[13] Kristof Meding and Christoph Sorge. What constitutes a deep fake? the blurry line between legitimate processing and manipulation under the EU AI act. 2024.

[14] Mina Momeni. Artificial intelligence and political deepfakes: Shaping citizen perceptions through misinformation. *J. Creat. Commun.*, October 2024.

[15] Samson Olufemi Olanipekun. Computational propaganda and misinformation: AI technologies as tools of media manipulation. *World J. Adv. Res. Rev.*, 25(1):911–923, January 2025.

[16] Andrew Ray. Disinformation, deepfakes and democracies: The need for legislative reform. *Univ. N. S. W. Law J.*, 44(3), September 2021.

[17] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. 2019.

[18] Kaylyn Jackson Schiff, Daniel S. Schiff, and Natalia Bueno. The liar's dividend: The impact of deepfakes and fake news on trust in political discourse. SocArXiv x43ph, Center for Open Science, Aug 2023.

[19] Aditya Kumar Shukla and Shraddha Tripathi. AI-generated misinformation in the election year 2024: measures of european union. *Front. Polit. Sci.*, 6, August 2024.

[20] Gregory Smith, Karlyn D. Stanley, Krystyna Marcinek, Paul Cormarie, and Salil Gunashekar. General-purpose artificial intelligence (gpai) models and gpai models with systemic risk, aug 2024. Published August 8, 2024.

[21] Sam Stockwell, Megan Hughes, Phil Swatton, Albert Zhang, Jonathan Hall, and Kieran. Ai-enabled influence operations: Safeguarding future elections. November 2024. CETaS Research Reports.

[22] Lingala Thirupathi, Alumalla Shivani Reddy, Visnu Vardhan, Tapasvi Panchagnula, Nata Sheker Reddy Miriyala, Saritha Gattoju, and T. V. Rajini Kanth. *Deepfakes: Understanding, Detection, and Mitigation in Cyber-Politics and Cyber-Economics*, page 351–374. IGI Global, January 2025.

[23] Divine-Favour Ukoh and Michael Adetunji. AI act: The EU regulation. *SSRN Electron. J.*, 2025.

[24] Cristian Vaccari and Andrew Chadwick. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Soc. Media Soc.*, 6(1):205630512090340, January 2020.

[25] Vilija Vainaite. Electoral processes in EU member states and deepfake-based disinformation: How do the responses differ? 2025.

[26] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, March 2018.

[27] Tianjiao Wang and Wenting Yu. The elephant in the room: Prior exposure to misinformation and correction effect. *SAGE Open*, 14(4), October 2024.