Enhancing Synthetic Media Detection and Mitigation to Thwart Election Interference

By: Jacob Locklear, Ritika Dixit, and Brian Ngac

Artificial intelligence (AI) is machine intelligence capable of mimicking human functions. It is relatively youthful but has grown in sophistication since the second half of the twentieth century. Initially, AI was limited to specific use cases, such as playing chess or spotting trends in data. In recent years, it has grown into a vast sector that encompasses diverse machine learning (ML) techniques and advanced algorithmic strategies. For example, artificial narrow intelligence (ANI) focuses on specialized tasks, such as voice and image recognition, and lacks the ability to perform additional tasks without the intelligence being retrained.[1]

The capabilities of ANI have historically focused on synthetic media generation, specifically the development of deepfakes. Deepfakes are artificially created graphics that appear to depict real individuals in computerized forms. Initially, they were used in oil and gas, entertainment, and art-related works, but their shift to political and social uses have become apparent in recent years. Concerns have grown regarding the influence of deepfakes on people's perceptions of political figures and events.[2]

In recent years, foreign access intelligence and deepfake tools have proven useful to intruders for other purposes, such as performing influence operations and meddling in US elections.[3] Such operatives seek to manipulate public opinion and undermine democratic systems by creating false narratives and using social media. Using deepfake media, the threat of meddling in the 2016 election became very powerful. Officials today expect AI systems will create false media even faster and with better accuracy. This issue calls for a structured approach to safeguard electoral processes. The proposed Deepfake Analysis and Detection (DAD) framework provides a clear and effective course of action to detect and mitigate deepfake media, with a focus on online platforms.

Mitigation Strategies

As synthetic media and deepfake technology continue to grow, maintaining credibility in digital content is imperative. Without effective detection strategies, synthetic media can be utilized to mislead the general population. To address these challenges, it is essential to examine three current key mitigation strategies.

Blockchain Technology

To combat the creation of deepfakes, blockchain is used to identify the origin of information and content. Blockchain uses a digital ledger system to keep track of information and consistently verify that same information when used.[1] In addition, the decentralized nature of blockchain helps validate and prevent the manipulation of digital information. With this, content on the blockchain can be hashed for integrity. Through this cryptographic process, a digital fingerprint unique to specific content can be created to ensure the authenticity of information. This attribute helps identify misinformation because if there is interference with content or the metadata on the blockchain, the hash value will change.[2] Considering deepfake technology includes video, image, or audio, blockchain manipulation serves a beneficial counter to this growing issue.

Digital Watermarking

Like blockchain, watermarking can help authenticate content. Watermarking includes embedding data in digital content that computers can detect without being noticeable to the consumers of that content.[6] If the embedded data or features are missing from digital content, then potential deepfakes can be detected. **Figure 1** shows common types of watermarking implemented on digital content for authenticity verification.

Figure 1—Types of Watermarks[7]

Type of Watermark	How It Works
Pixel-based	Watermark is embedded by modifying the individual pixels of an image or file
Vector-based	Watermark is turned into a scalable vector graphic without loss of quality
Frequency domain	Watermark is hidden in the color patterns of an image or video, making it hard to see and even harder to remove

Federal and State Deepfake Laws

Currently, federal regulation in the United States to combat the use of deepfakes does not exist. At the US-state level, several states have enacted laws to regulate the usage of deepfakes—many of which relate to the creation of deepfake sexual content. Recently, efforts to combat deepfake content in elections have also commenced:

- In 2019, Texas SB 751 made it criminally punishable to create a deepfake video with the intent to injure a candidate or influence the result of an election.[8]
- In March 2024, both Indiana HB1133 and Oregon SB 1571 began compelling election campaigns to provide a disclaimer if digitally altered media is used in campaign communications. The law also allows affected parties of digitally altered media that did not possess a disclaimer to pursue litigation against the creator.[9]
- As of September 2024, California AB 2655 requires large online platforms to block the posting of materially deceptive content related to California elections during specific periods before and after an election. The bill also requires the platforms to implement tools to report and label deceptive content as inauthentic, fake, or false.[10]

Deepfake Usage in Recent US Elections

In May 2019, a digitally altered video circulated on social media platforms portraying former US House of Representatives Speaker Nancy Pelosi as inebriated. The slowed and distorted video of Pelosi's speech garnered millions of views and fostered a negative public perception, which reinforced the narratives promoted by her political opponents.[11]

In June 2023, a super Political Action Committee (PAC) for Ron DeSantis posted AI deepfakes on Twitter showing US President Donald Trump and Dr. Anthony Fauci locked in an embrace intended to suggest in an unflattering manner that Trump was overly friendly with Fauci throughout the COVID-19 pandemic.[12] While these pictures were created purely for humorous purposes, the realism of these images raised concerns about the application of deepfakes to change the opinions of the electorate. Other critics maintained that such content, even if parody, would still be a source of depolarization in contexts such as political campaigning.

In January 2024, a political consultant used AI-generated voice cloning technology and caller ID spoofing to send a deepfake phone call to US State of New Hampshire residents, discouraging them from voting in the upcoming primary election.[13] The phone call mimicked the voice of former US President Joe Biden and included common phrases that he uses.[14] This incident indicates how deepfakes can be used for voter suppression.

Also in January 2024, a false audio clip generated by AI technology circulated on social media claiming that Biden threatened to send F-15s to Texas with the aim of "quelling" opposition near the border. The White House quickly refuted the fake audio. It was difficult for people to dismiss the deepfake due to its believable voice and other convincing aspects.[15]

Comparative Analysis

With the recent use of deepfake media to influence US elections, there is major concern about its impact on public perception. This analysis aims to compare the previously noted instances of political deepfake usage.
The Pelosi, Biden phone call, and F-15 deepfakes highlight the use of audio deepfakes to foster a negative public perception of the politician involved, while the Trump–Fauci deepfake highlights the use of deepfake imagery. The Pelosi deepfake aimed to reduce the former House Speaker's credibility and influence voter opinions ahead of the 2020 election. In contrast, the deepfake phone call from Biden was a more direct attempt to suppress democratic voters in the primary election. Similarly, the F-15 deepfake was used to spread misinformation on social media platforms.

The Pelosi, Trump–Fauci, and F-15 deepfakes were all spread on social media platforms. More specifically, the Pelosi deepfake was spread on Facebook and X (formerly Twitter). This highlights the importance of further initiatives by social media platforms to combat the spread of disinformation from deepfake media.[16]

The use of audio deepfakes is presented to be more impactful on public perception, as audio deepfakes are less detectable and therefore a greater threat to influencing the general population. Whether created for trickery or humor, deepfakes mix the real and fake, undermining confidence in government and media credibility. This confusion can escalate in political systems, as deepfakes often cater to specific groups, reinforcing their biases. Voters become ill-informed, leading to disorder in political discussion and eroding faith in the democratic system.

Foreign Adversaries and Their Use of Synthetic Media

As generative AI advances, US elections face an increased risk of influence by foreign adversaries. The following analysis focuses on key groups within China, Iran, and Russia. The three countries were selected based on their active engagements and technological capabilities in influencing US elections. Their interests in swaying public opinion and undermining the democratic process make them significant threats to US election security.

Russian Adversarial Threat to the United States

According to the Microsoft Threat Analysis Center,[17] leading up to the 2024 presidential election, Russian actors (some noted in **figure 2**) aligned with the Kremlin are being monitored for spreading disinformation online. They are believed to be associated with the Internet Research Agency, a Russian company involved in spreading disinformation for political gain.

Figure 2—Russian Foreign Adversaries to the United States

Foreign Adversaries	Influence Operations	Influence Techniques	Goals
Storm-1516	Spread of anti-Ukraine narratives through online channels	AI-generated content	Influence US policy, incite domestic controversy, divert public attention
Storm-1516	False depiction of Ukrainian support for Joe Biden [18]	Deepfake video, AI-generated voice	Influence voter opinion
Storm-1679	Fake billboard containing false assertions about US presidential candidate Kamala Harris's policies[19]	Digitally altered video	Discredit political candidate, influence voter opinion

Chinese Adversarial Threat to the United States

Chinese cyber actors and hackers (some noted in **figure 3**), affiliated with the Chinese Communist Party (CCP) and the People's Liberation Army (PLA), are increasingly using tools to interfere with US elections. According to US intelligence officials, Chinese advocates use AI to enhance disinformation campaigns.[20] The PLA has developed an Integrated Network Electronic Warfare (INEW) strategy, combining cyberattacks with other forms of information warfare. State-affiliated hackers in China have employed AI to automate the creation of disinformation, making it easier to spread misleading content.[21]

Figure 3—Chinese Foreign Adversaries to the United States[22]

Foreign Adversaries		Influence Operations	Influence Techniques	Goals
Storm-1376		Utilization of bot farms to post inflammatory comments about highly polarizing topics	AI-generated content, spamouflage	Influence voter opinion, incite domestic controversy
APT41 and APT10		Cyberespionage and spread of disinformation	AI-generated content	Interrupt US political process

Iranian Adversarial Threat to the United States

Iran-linked actors and hackers are ramping up their efforts to mislead voters and foster a divisive political environment in the United States. However, US audiences are not the only demographic at risk of election interference. Campaign and government stakeholders are also subject to Iran-linked efforts to impact US elections. Although the specific motives behind the attempted Mint Sandstorm hacks (noted in figure 4 along with other activities) are unclear because the threat was mitigated, the repercussions of the incident could have been detrimental to US election integrity and security.

Figure 4—Iranian Foreign Adversaries to the United States[23]

Foreign Adversaries	Influence Operations	Influence Techniques	Goals
Storm-2035	Creation of websites releasing controversial political content	AI-generated content plagiarized from authentic sources	Divide voters, incite domestic controversy
Sefid Flood	Preparation for possible operations	Media fabrication, impersonation, intimidation, doxing of political figures and groups	Weaken US election integrity, incite political violence
Mint Sandstorm (group linked to IRGC)	Compromise of the email account of a former presidential campaign's advisor and attempted to access a former candidate's email	Password spraying, spear fishing	Undermine US election security

Looking forward, there is an opportunity for domestic organizations to help counteract influence operations that spread deepfake and synthetic media. The artificial intelligence company, OpenAI, has disrupted more than 20 attempts of deceptive content creation using their AI models since the beginning of 2024.[24] OpenAI recognizes the importance of building safeguards that identify and prevent the use of their AI services in generating fake content that can be used to influence people on internet platforms.[25] This serves as a precedent for further initiatives from organizations with AI capabilities to mitigate threats using their detection tools.

Call to Action

Given the ongoing threat of election inference by cyber adversaries, mitigating the impact of synthetic media use is crucial. Ensuring the integrity of political content is essential to strengthening the democratic process.

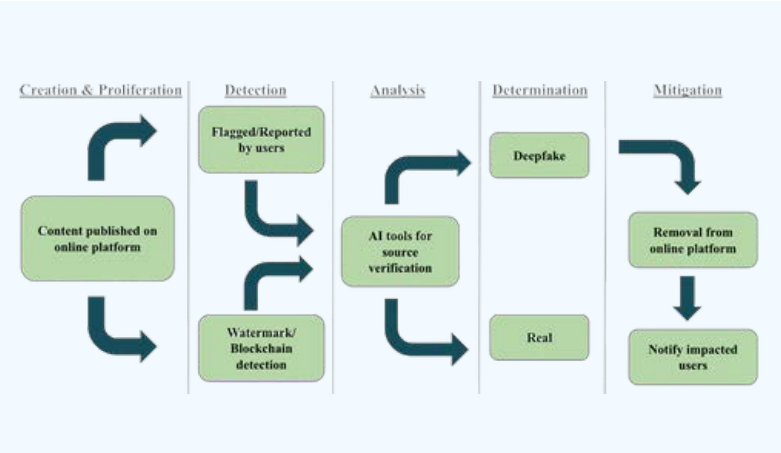
The most common social media platforms containing published deepfake content are X, Facebook, and TikTok. Deepfake content can reach many users in a short amount of time. Furthermore, synthetic content has a higher tendency to spread more rapidly than authentic content. On X, false news stories are 70% more likely to be reposted than true ones and have the ability to reach 1,500 people six times faster.[26] Therefore, there must be a stricter system that social media platforms use to stop the proliferation of deepfake content.

X uses their own technologies to identify synthetic and manipulated media through in-app reporting. Additionally, X partners with third party organizations to detect manipulated media. Meta (formerly Facebook) labels posts as "AI info" on their platform that they detect were generated with AI.[27] The Chinese-owned company, TikTok, allows users to flag posts that are believed to contain AI-generated or deepfake content.[28] Similar initiatives should be more commonplace surrounding political content. With this, platforms can take action to mitigate the spread of synthetic media.

Deepfake Analysis and Detection Framework

After a thorough analysis of current deepfake identification and mitigation operations, the authors propose the Deepfake Analysis and Detection (DAD) framework, which is applicable to any online platform that allows users to publish content. The goal of the DAD framework is to allow for effective identification of synthetic media using AI tools and mitigate the impact on public perception. **Figure 5** lays out the DAD framework and the five phases involved in the process.

Figure 5—The Deepfake Analysis and Detection (DAD) Framework



The differential factor in the DAD framework is the ability to mitigate the impact of each incident. An agile response from online platforms is essential in minimizing potential impacts on public perception. There is value in examining a walkthrough of the DAD Framework and use case scenarios using the F-15 deepfake example discussed previously.

F-15 Use Case:

Creation and Proliferation

Content is published on an online platform, and content spreads rapidly, impacting public perception on a large scale.

The audio clip is posted on TikTok, reaching a wide audience, and spreading disinformation.

Detection

Suspicious content is flagged either by platform users or the detection of watermarking and blockchain technology (implemented by the online platform) for further inspection.
Users are unsure about the authenticity of the audio clip and flag the suspicious content for further analysis by TikTok. If the audio clip contains digital watermarking, TikTok uses these findings in their analysis. The combination of both user alerts and notification from the watermarking/blockchain technologies would prioritize the need for analysis on the suspicious content.

Analysis

Platforms leverage AI tools to investigate the source of content and assess credibility.
TikTok uses AI tools to scan the audio clip for traits consistent with deepfakes. The audio clip is fact checked and the source is investigated.

Determination

Following analysis, a decision is made on content authenticity. If the content is deemed authentic, it may remain on the online platform and will be recorded in a database to prevent repetition of the DAD framework.
The audio clip is deemed an AI-generated deepfake, necessitating further action.

Mitigation

If content is inauthentic, the platform can potentially remove the content. In addition, any user who viewed the content is promptly notified through the online platform to mitigate the impact on public perception.
TikTok removes the audio clip. Every user that encountered the audio clip is notified through TikTok with the date and time of the encounter and the results of the analysis.

Conclusion

Synthetic media has been an acute and escalating threat to the integrity of electoral processes in recent years. Entities are systematically using AI to interfere in elections, requiring coordinated national efforts to combat these activities. The proposed DAD framework facilitates a straightforward approach to deepfake detection and mitigation, specifically for online platforms. Through the DAD framework, a prompt and clear response limits the false public perception that deepfakes can create. With the unified efforts between the private and public sectors, stronger safeguards can be built that protect elections in an increasingly digital world.

References

[1]Delipetrev, B.; Tsinaraki, C.; et al.; Historical Evolution of Artificial Intelligence, EUR 30221 EN, Publications Office of the European Union, Luxembourg, 2020, https://publications.jrc.ec.europa.eu/repository/handle/JRC120469

[2]Delipetrev; Historical Evolution

[3]Smith, B. (2025b, January 29). Securing us elections from nation-state adversaries. Microsoft On the Issues. https://blogs.microsoft.com/on-the-issues/2024/09/18/securing-us-elections-from-nation-state-adversaries/

[4]Harrison, K.; Leopold, A.; "How Blockchain Can Help Combat Disinformation," Harvard Business Review, 6 September 2024, https://hbr.org/2021/07/how-blockchain-can-help-combat-disinformation

[5]Ho, C.; "AI and Blockchain Can Mitigate Fraud Risk Caused by Deepfakes," Forbes, 29 August 2024, www.forbes.com/sites/digital-assets/2024/07/06/ai-and-blockchain-synergies-mitigate-risk-of-deepfakes-in-kyc/

[6]U.S. Government Accountability Office, "Science & Tech Spotlight: Combating Deepfakes," USA, 11 March 2024, www.gao.gov/products/gao-24-107292

[7]Box, "What Is a Digital Watermark?," www.box.com/resources/what-is-a-digital-watermark

[8]Graham, M.; "Deepfakes: Federal and State Regulation Aims to Curb a Growing Threat," Thomson Reuters Institute, 27 June 2024, www.thomsonreuters.com/en-us/posts/government/deepfakes-federal-state-regulation/

[9]Graham, M.; "Deepfakes: Federal and State Regulation Aims to Curb a Growing Threat," Thomson Reuters Institute, 27 June 2024, www.thomsonreuters.com/en-us/posts/government/deepfakes-federal-state-regulation/

[10]California Legislative Information, AB-2655 Defending Democracy from Deepfake Deception Act of 2024, USA, leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240AB2655

[11]O'Sullivan, D.; "Doctored Videos Shared to Make Pelosi Sound Drunk Viewed Millions of Times on Social Media," CNN, 24 May 2019, www.cnn.com/2019/05/23/politics/doctored-video-pelosi/index.html

[12]Nehamas, N.; "DeSantis Campaign Uses Apparently Fake Images to Attack Trump on Twitter," The New York Times, 8 June 2023, www.nytimes.com/2023/06/08/us/politics/desantis-deepfakes-trump-fauci.html

[13]Bracken, M.; "FCC Hits Operative Behind New Hampshire Robocall With $6 Million Fine," CyberScoop, 26 September 2024, cyberscoop.com/fcc-fine-joe-biden-deepfake-new-hampshire-robocall-steve-kramer/

[14]Swenson, A.; Weissert, W.; "New Hampshire Investigating Fake Biden Robocall Meant to Discourage Voters Ahead of Primary," AP News, 22 January 2024, apnews.com/article/new-hampshire-primary-biden-ai-deepfake-robocall-f3469ceb6dd613079092287994663db5

[15]Atherton, D.; "Incident 703: Deepfake Audio Sparks False Claims of Biden Threatening Texas with f-15s," AI Incident Database RSS, incidentdatabase.ai/cite/703/

[16]Mervosh, S. (2019, May 24). Distorted videos of Nancy Pelosi spread on Facebook and Twitter, helped by Trump. The New York Times. https://www.nytimes.com/2019/05/24/us/politics/pelosi-doctored-video.html

[17]Microsoft On the Issues, "Russian US Election Interference Targets Support for Ukraine After Slow Start," 18 November 2024, blogs.microsoft.com/on-the-issues/2024/04/17/russia-us-election-interference-deepfakes-ai/

[18]Atherton, D.; "Incident 727: Synthetic Voice 'Olesya' by Storm-1516 Falsely Accuses Ukraine in U.S. Election Disinformation Campaign," AI Incident Database, incidentdatabase.ai/cite/727/

[20]Collier, K.; "Russia, Iran, China Using AI in Election Interference Efforts, U.S. Intelligence Officials Say," NBC News, 24 September 2024, https://www.nbcnews.com/tech/security/russia-iran-china-are-using-ai-election-interference-efforts-us-intell-rcna172476

[21]Collier, K.; "Russia, Iran, China Using AI in Election Interference Efforts, U.S. Intelligence Officials Say," NBC News, 24 September 2024, https://www.nbcnews.com/tech/security/russia-iran-china-are-using-ai-election-interference-efforts-us-intell-rcna172476

[22]Microsoft On the Issues, "China Tests US Voter Fault Lines and Ramps AI Content to Boost its Geopolitical Interests," 4 April 2024, https://blogs.microsoft.com/on-the-issues/2024/04/04/china-ai-influence-elections-mtac-cybersecurity/

[23]Microsoft On the Issues, "Iran Targeting 2024 US Election," 8 August 2024, blogs.microsoft.com/on-the-issues/2024/08/08/iran-targeting-2024-us-election/

[24]OpenAI. (2024, October 9). An update on disrupting deceptive uses of ai. https://openai.com/global-affairs/an-update-on-disrupting-deceptive-uses-of-ai

[25]Nimmo, B.; Flossman, M.; Influence and Cyber Operations: An Update, OpenAI, October 2024, cdn.openai.com/threat-intelligence-reports/influence-and-cyber-operations-an-update_October-2024.pdf

[26]Vosoughi, S.; Roy, D.; et al.; The Spread of True and False News Online, 2018, ide.mit.edu/wp-content/uploads/2018/12/2017-IDE-Research-Brief-False-News.pdf

[27]Bickert, M. (2024, September 12). Our approach to labeling AI-generated content and Manipulated Media. Meta. https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/

[28]Colman, B.; "Why Tiktok's New Deepfake Policy Falls Short," Reality Defender, 8 May 2023, www.realitydefender.com/blog/tiktok-deepfake-detection-measures

Jacob Locklear
Is a business systems analyst for CACI International Inc.

Ritika Dixit
Is a business analyst for the Commonwealth of Kentucky.

Brian Ngac
Is an instructional faculty member and Dean's Teaching Fellow at George Mason University's Costello College of Business (Fairfax County, Virginia, USA). He developed the Professional Readiness Experiential Program (PREP) where undergraduate students focus on business process improvement efforts with industry participants on real projects. If your organization is interested in participating as a sponsor of PREP, please contact bngac@gmu.edu.