DEEPFAKED PUBLIC FIGURES

AN A.I. ELECTION RISK CHECKLIST

BY: ASPEN DIGITAL'S
A.I. ELECTIONS ADVISORY COUNCIL

Altered video, images, and audio can confuse voters about what a public figure has done or said. Low public trust in news media and in civic institutions can make people more vulnerable to deception.

AI tools can be used to depict public figures doing or saying something they did not, to mimic their voice, or to modify authentic content for nefarious purposes.

MISUSE EXAMPLES

- Fake celebrity endorsement of a minor candidate.
- Fake audio of a secretary of state appearing to plot to alter vote totals.
- Fake images show campaign officials appearing to bribe election workers.
- Snippets from a real speech are altered to make the candidate appear unwell.
- Short-form videos on social media falsely depict a celebrity promoting an election boycott.

MITIGATIONS

NEWS MEDIA, ADVOCATES, AND CIVIL SOCIETY

- Empower communities through proactive messaging that prepares the public for risks and promotes good information hygiene: *AI scammers won't change your mind or keep you home this election. If it's not from a real news site or election office, it's not worth your time.*
- Avoid "*spot-the-deepfake*" narratives that rely on individual skill in the face of rapidly improving AI, as that may promote unhealthy suspicion or overwhelm people.
- Identify groups that may be uniquely targeted (e.g., elderly voters may answer a robocall at greater rates than young voters).
- Favor crisis communication strategies that equip existing community leaders with information and messaging.
- Establish relationships with election officials and digital forensic experts to ensure false election information can be quickly corrected.

Interested in learning about other AI election risks? Visit www.aspendigital.org/ai-election-risks for more risk checklists.

ELECTION ADMINISTRATORS

- Adopt internal communication and verification protocols to combat sophisticated phishing, such as the use of a supervisor's voice or likeness (e.g., using a state elections director's voice to direct different procedures via phone call to a county elections administrator).
- Establish a communications playbook for scenarios where AI tools are used to depict election officials or poll workers for nefarious purposes (e.g., a county election director is depicted accepting an envelope of cash).
- Communicate to local news outlets that bad actors may attempt to use the image or voice of election officials to sow confusion on and after Election Day.

SOCIAL MEDIA

- Consider elevating election news content from historically trustworthy accounts (e.g., local election officials and press outlets) throughout the election period to combat information distortion by bad actors, fake accounts, or other sources.
- Make appropriate investment in content moderation and integrity teams across trust and safety functions and encourage first-party use of new AI tools for integrity purposes (e.g., improving classifier performance using multilingual models).
- Monitor content across other platforms (including niche forums like 4chan, 8kun and others) to surface harmful fakes and evolving terminology that bad actors may migrate to major platforms. Consider using AI-enabled, narrative-level summation tools to monitor real time narrative trends.
- Evaluate whether present integrity systems adequately mitigate AI-enabled shifts from viral content to many variations on core content; from language-limited reach to multilingual distribution; and from foreign troll farms to onshored automated systems.
- Establish data-backed interventions that inform users that content is likely synthetic (e.g., labels). Consider preparing crisis protocols that elevate civic content with high signals of authenticity during a time of heightened concern (e.g., leveraging C2PA metadata and other internal signals).
- Share signals of inauthentic behavior and influence operations across platforms and relevant government actors, where permitted (e.g., hashtags being used to spread fake images, threat insights, and behavioral signals).

MESSAGING APPS & SERVICES

- Adopt crisis playbook for civic deepfake scenarios.
- Establish news partnerships for quality content in-app.

A.I. LABS, DEVELOPERS, & COMPANIES

- Develop and implement integrity mitigations that combat deceptive appropriation of public figures' likeness and voice (e.g., rejecting requests to modify images or videos of public figures).
- Establish dedicated communication channels for trusted partners to flag problematic content that may have originated with the model.
- Establish channels with content distribution platforms to notify the AI company/lab when harmful fakes appear to have originated with its model.
- Make appropriate investment in content moderation and integrity teams across trust and safety functions.