Comments Regarding Petition to Clarify Rulemaking Regarding the Regulation of "Deepfakes" in Campaign Communications

Overview

Dear Federal Election Commission,

We agree with and support Public Citizen's request that the Commission amend its regulation on fraudulent misrepresentation of campaign authority to make clear that its scope applies to the deliberately deceptive use of deepfakes in official elections campaign communications. This will mean, in practical terms, that deepfakes must be disclosed as such, in context, in order to comply with FEC guidelines.

In our comments, we address the reasons we believe Artificial Intelligence technology used in campaign communications warrants inclusion under the Commission's regulations for fraudulent misrepresentation. We provide recommendations of factors to consider in amending this regulation, drawing from our professional and technical expertise as current and past workers at social media and technology companies that actively contend with many of the risks outlined by Public Citizen. We welcome this opportunity and look forward to future engagements on this critical topic.

Additionally, the document contains an appendix, providing recommendations for public/private/civic sector engagement to address the broader problem of manipulated media in political communications outside of the FEC's immediate scope.

We submit these comments in our capacity as members of the Integrity Institute[1], a think tank of professionals who work in Trust & Safety. We are engineers, product managers, researchers, analysts, data scientists, operations specialists, policy experts and more, with decades of combined experience across numerous platforms. Our mission[2] is to advance the theory and practice of protecting the social internet, powered by our community of integrity professionals. We understand the systemic causes of problems on the social internet and how to mitigate them. We bring this experience and expertise directly to the people theorizing, building, and governing the social internet. We produced a two-part Election Integrity Best Practices guide on how to support healthy elections on online platforms[3][4].

Background/Problem statement

Definitions

Artificial Intelligence is a technology that's used extensively in a wide range of tools, such as photo enhancement, personalized playlists, and the detection of email spam. Generative AI is a type of AI that can be used to create new or heavily modified content (text and multimedia). Generative AI technology is rapidly becoming more available and accessible to a broad audience through existing interfaces like browsers, media editing tools, and virtual assistants. Generative AI enables the production of text and multimedia content at far greater speed, scale, and sophistication than other methods.

Deepfakes are "synthetic media"[5] that have been created or substantially manipulated using Generative AI to persuasively depict falsehoods as having actually occurred. Deepfakes can be produced in image, audio, and video formats. They're developed by training AI algorithms on reference content until they can produce media that's nearly indistinguishable from real life. For example, two days before Slovakia's national election in September 2023, a damaging audio deepfake of a party leader discussing how to rig the election was circulated on social media.

Although this letter is concerned with the capacity of deepfakes to undermine democratic integrity, it should be noted that deepfakes have also been used for hoaxes such as a fake explosion at the Pentagon (which caused the stock market to dip),[6] as well as for expressly criminal purposes such as posing as ostensibly kidnapped family members,[7] and misrepresentation by criminals in remote job interviews.[8]

[1] https://integrityinstitute.org
[2] https://integrityinstitute.org/our-mission
[3] Election Integrity Best Practices Guide Part 1: https://integrityinstitute.org/news/institute-news/elections-deck-v1
[4] Election Integrity Best Practices Guide Part 2: https://integrityinstitute.org/news/institute-news/elections-pt-2
[5] https://en.wikipedia.org/wiki/Synthetic_media
[6] https://www.cnn.com/2023/05/22/tech/twitter-fake-image-pentagon-explosion/index.html
[7] https://abcnews.go.com/Technology/experts-warn-rise-scammers-ai-mimic-voices-loved/story?id=100769857
[8] https://www.darkreading.com/attacks-breaches/criminals-deepfake-video-interview-remote-work

Generative AI tools like ChatGPT have achieved an incredible velocity of user adoption, ranking as a product with the fastest growing user base, and birthing an ecosystem of developer tools around the large language models underlying it. The fast-paced growth has presented a challenge for AI vendors wary of seeking to limit potential abuses of the technology by adversarial actors.

Risks posed by artificial intelligence

Although AI isn't the first and won't be the last technology used to produce deceptive media, we believe it deserves special attention because it's a powerful tool that can augment existing practices and elevate existing risks. We believe these tools pose added risks by enabling bad-faith actors to more easily and efficiently create more realistic-seeming deceptive media across multiple mediums. This could result in a faster pace and larger volume of such content flooding information ecosystems such as on social media, which platforms and the general public are not well-equipped to handle.

However, as discussed in the appendix, political advertising is not the only avenue through which deceptive media is distributed, standard disclosure requirements can help educate the public on the use of AI in paid advertising and other public communications.[9] It also lays a framework to build on in the short term. Learnings can be drawn from political ads and applied to unpaid content.

Threats to Democratic Integrity

We believe the risks of AI-manipulated and AI-generated deceptive media warrants additional scrutiny in the context of content related to elections. These risks have risen as confidence in the integrity of election processes and institutions has declined in recent years.[10] Many Americans do not trust established news media, and in fact believe that news institutions have the intent to deceive them[11]. The mere potential for deepfakes to proliferate would further accelerate declining trust in information—for instance, any claims by political leaders that undesirable authentic content is a deepfake could contribute to further erosion in trust of information and democratic institutions.

Our position

As stated, the use of deepfakes to influence elections by persuasively presenting fiction as truth, is inherently and demonstrably a fraudulent act. We urge the FEC to amend its regulation on

[9] Public communications as defined by the FEC: https://www.fec.gov/press/resources-journalists/public-communications

[10] https://www.pewresearch.org/politics/2022/10/31/views-of-election-administration-and-confidence-in-vote-counts

[11] https://knightfoundation.org/reports/american-views-2023-part-2/

"fraudulent misrepresentation" at 11 CFR 110.16(a),[12] to clarify that its regulatory oversight applies to materially deceptive "public communications"[13] produced using AI technologies. Within the parameters of existing regulations, we believe this is an important step towards mitigating the outsize threat of deepfakes on the integrity of elections.

(1) As we interpret the parameters articulated in 11 CFR 110.16(a) and then-Commissioner Goodman's 2018 discussion of the Fraudulent Misrepresentation Doctrine:[14] media would not be in compliance if it believably portrays, without disclaimers, another candidate, saying or doing something they never actually did, would be in violation.

(2) Additionally, we believe that by definition, the proposed clarification wouldn't apply to AI-produced synthetic or heavily manipulated content that is very clearly satire, parody, or otherwise artistic in nature, to a reasonable person regardless of the person's awareness of current events, candidates, or communication styles or political beliefs of particular candidates.

Therefore, we support the recommendation that the FEC require clear and conspicuous disclosure by creators of political communications when AI is used to manipulate and generate content that fits the criteria above. There are some practices that may apply to how disclosures should be implemented, which we discuss below. We recognize the challenges in monitoring and enforcing such a requirement in real time, and believe therefore this is ultimately a shared responsibility of a broad ecosystem of stakeholders including any platforms that host such content (including social media, messaging, phone calling), civil society, and academia.

Digital and social media platforms are in a much better position to support compliance for this requirement in political advertisements on their platforms compared to non-advertising public communications. There is an existing checkpoint[15] during which the platforms review advertisements both automatically and manually before they are shown to users, for a number of legal and platform requirements.

Recommended specifications for disclosing deepfakes

In order to facilitate consistency across platforms and mediums, as well as with existing political advertising disclosure requirements, we recommend incorporating the following specifications for disclosures of the use of AI to generate manipulated media.

[12] https://www.ecfr.gov/current/title-11/chapter-I/subchapter-A/part-110/section-110.16

[13] https://www.fec.gov/press/resources-journalists/public-communications

[14] https://www.fec.gov/resources/cms-content/documents/Commissioner_Lee_E._Goodman_Policy_Statement_-_Fraudulent_Misrepresentation.pdf

[15] E.g., Facebook verifications for political ads https://www.facebook.com/business/help/2992964394067299?id=288762101909005

Platforms and Mediums

- Any public communication (as defined by the FEC) that supports AI-manipulated or generated media, including audio, video, image, and print.
- Any outlet or platform that distributes political advertisements or medium that is directly distributed to voters, including phone lines (e.g., robocalls), television, radio, internet services and social media, messaging platforms, and print media. Communications that are not considered to be public communications by the FEC's definition but currently require "paid for by" disclaimers, such as electronic mail and websites of political committees that are available to the general public, are also in scope.

Placement and format

- In all instances of messages posted online, the disclaimers should be contained within the media file rather than just appearing alongside it on a page so that the disclaimer is still included if people reshare or clip individual media files.
- If size or other constraints of a message make it impossible to provide a disclaimer that meets the "clear and conspicuous" standard as defined by the FEC, the message is not permitted.
- Video
    - The video should include visual disclaimers, at minimum lasting four seconds or more, at both the beginning and end. A visual disclaimer should also accompany the duration of any display of fictitious content.
- Images
    - "Adapted" disclaimers, as defined by the FEC[16], still need to display pertinent and comprehensible information without users needing to take additional steps
- Audio
    - The disclaimers should be placed at the beginning and end.

Disclaimer content and messaging

This section contains *suggested* disclaimer wording. This wording is aimed at achieving consistency across platforms and mediums. While different disclaimers are suggested based on format, these should not be technology-prescriptive. We recommend that future modifications are informed by academic and user research on the effectiveness of messages and related tactics. It is also for this reason that transparency and reporting requirements are critical.

- Format, e.g., text size and color: Same as current advertising requirements.
- Language: Any disclaimer should be included in the majority language used in the ad, as well as English.
- We suggested the following content and messages for disclosures on each medium:
    - Images:

[16] https://www.fec.gov/help-candidates-and-committees/advertising-and-disclaimers/#special-rules-for-internet-public-communications

        - *“This [ad/promoted content] contains materially altered or fabricated [events/actions/statements].”*
    - Video
        - Beginning: “*This [ad/promoted content] contains materially altered or fabricated [events/actions/statements].”*
        - End: “*This [ad/promoted content] contained depictions of materially altered or fabricated [events/actions/statements].”*
    - Audio message
        - Beginning: “*This advertisement contains materially altered audio.”*
        - End: “*This advertisement contained materially altered audio.*”

High-Level Implications for Advertising Platforms

While the FEC will be the ultimate authority as to whether or not an ad violated these regulations, we think platforms should require advertisers to disclose when AI is used in political advertisements. Additionally, we think platforms should build the tools to help advertisers be in compliance. The platforms' responsibilities would include:

1. Inclusion of the recommended AI disclosure requirements into their Advertising Policies.
2. Reasonable effort to ensure that disclaimers are visible. For example, disclaimers should remain visible irrespective of which client (Web, iOS, or Android) the content is being consumed on.
3. Platforms that provide ads transparency libraries should voluntarily include relevant metadata to enable researchers to identify ads that have disclosed the use of AI in creation.
4. Platforms should publicly release data on how much manipulated media is being uploaded to their services and how many people are viewing it.

Advertising platforms may face the following challenges in supporting such a disclosure requirement:

1. Tracking & monitoring for compliance: In reviewing ads to ensure they’ve included the disclosures, even as part of reviewing for adherence to their own Advertising Policies, Advertising platforms will face a challenge of detecting that generative AI has been used. Automated detection is technically difficult, and even human reviewers may find it difficult to distinguish content that has been manipulated. However, the best point of intervention is at the point that an advertisement is being set up in an advertising platform’s system, before it obtains widespread reach that becomes even more challenging to track and monitor.
2. Feasibility of responding to recommendations in a timely manner: Some platforms that may host altered content and/or advertisements may not have the technical or monetary resources to set up the infrastructure needed to classify, detect, arbitrate, enforce, and effectively address deepfakes in a highly responsive manner before they have been broadly exposed to the general public.

3. Cross-platform consistency: Implementation of recommendations could be inconsistent across platforms. On one hand, high-level recommendations provide needed guidance and can be applied throughout the sector; on the other hand, interpretation and implementation quality of the guidance could vary drastically depending on each company's resourcing and capabilities.
4. Understanding the impact of disclosures: We recommend disclosure as the best available mechanism with existing precedence to help the general public understand and contextualize the media they consume. However, we believe it is important to consistently research and understand the actual effectiveness of disclosures in informing the public and building trust in information; and for FEC requirements to adjust based on the evidence.

Ecosystem Guidance

Although additional engagement and oversight by the FEC is apt and warranted, the most impactful election deepfakes to date were produced and distributed surreptitiously via social media rather than paid advertisements. The issue of deepfakes is multifaceted, the product of numerous technical and non-technical factors; a blended, cross-sector, and cross-government approach is needed. This section provides a subset of recommendations in development towards that end.

Legislative agenda

As a general rule, we believe that public policy should not get far ahead of emerging technologies or markets, instead taking the long view and acting deliberately over time as needed. However, as the misuse of Generative AI is a profoundly distinct and high risk matter, our vantage point might best be described as cautiously proactive.

At present, all applicable Federal[17] and state[18] legislation regarding deepfakes is either limited or in nascent stages. There has also not been a case of enforcement of any of the state laws to our knowledge. We recommend adoption of the following principles:

- Focus on whether media content is materially deceptive (we define this below) and for purposes of influencing election outcomes; apply regulations irrespective of the technology used, delivery systems (e.g., posting a deepfake rather than using it in ads), and parties involved.
    - Technology neutrality. Although deepfakes fundamentally surpass all other machine-based techniques for rapidly producing credible synthetic or manipulated media at scale, other techniques continue to be problematic. One example is simple slowing down of content, as was the case in a manipulated video of Nancy Pelosi, which made her appear to be drunk or mentally

[17] Crowdsourced Federal Legislative Proposals Pertaining to Generative AI (118th): https://docs.google.com/document/d/1A1bJ1mkIfE3eZuSbDmz3HGVtOvQDegHl53q3ArO7m44/edit

[18] State AI legislation tracker: https://www.ncsl.org/technology-and-communication/artificial-intelligence-2023-legislation

impaired.[19] Additionally, there will likely be other technology in the future that can be subverted for illicit use. A technology-agnostic approach will help "future proof"[20] regulations.

    - Delivery system neutrality. As stated, so far the more damaging election deepfakes were not formally affiliated with a candidate or campaign. Regulations should be applied regardless of whether a deepfake was shared on social media or used in television ads.
    - Affiliation neutrality. Similarly, regulations should be applied regardless of the affiliation of the people involved, again focusing instead on whether manipulated media is materially deceptive. Regulations should be crafted to account not just for traditional campaign personnel but also foreign influence operations, solo actors, groups that have unofficial relationships with campaigns, etc.
- California's election code for machine-manipulated media[21] provides a good model for technology-, distribution- and actor-agnostic approaches, including using a "reasonable person" approach:
    - *(1) The image or audio or video recording would falsely appear to a reasonable person to be authentic. (2)The image or audio or video recording would cause a reasonable person to have a fundamentally different understanding or impression of the expressive content of the image or audio or video recording than that person would have if the person were hearing or seeing the unaltered, original version of the image or audio or video recording.*
- Avoid mandates for the broad use of labels and warnings. Such practices quickly lead to diminishing returns, as users are more likely to disregard them as noise. Although the labeling of Generative AI-produced content is better than no labeling of any kind, ideally, only materially deceptive media would be labeled. Barring that, labeling requirements should be written such that materially deceptive media is distinct from benign uses of Generative AI, such as for evidently creative[22], artistic, or parodic works.
- Ensure legislation addresses the issue of materially deceptive synthetic media that doesn't directly target a candidate's opposition but still is used to influence the outcome of an election. For example, election deepfakes could be used to suggest that an incendiary event, such as violent actions by particular demographics or authorities, has taken place. Such deepfakes could spark unrest and/or undermine an opponent's support structure.
- Future considerations:
    - All relevant legislation sets a specific timeline of 30 - 60 days leading up to elections for when the laws against election deepfakes and other manipulated media should be applied. We support this approach (in keeping with our cautiously proactive bent). However, as there are always campaigns in between

[19] Reuters, "Fact check: "Drunk" Nancy Pelosi video is manipulated," August 3, 2020 https://www.reuters.com/article/uk-factcheck-nancypelosi-manipulated/fact-check-drunk-nancy-pelosi-video-is-manipulated-idUSKCN24Z2BI

[20] https://en.wikipedia.org/wiki/Future-proof

[21] California Election Code § 20010 on casetext: https://bit.ly/3rWxfeH

[22] Balenciaga parody deepfake, https://www.youtube.com/watch?v=iE39q-IKOzA

elections to influence public policy and political decisions. We expect there will be growth in the misuse of Generative AI for such purposes.

- Related legislation
  - Bots do not have free speech rights. We recommend prohibitions against the use of bots to pose as people or impersonate real people online for purposes of influencing elections. California[23] has such legislation on the books. It's also notable that the prohibition is year-round and applies to certain business practices as well.
  - Similarly, computer crime laws in some states, such as Rhode Island,[24] prohibit online impersonation. Such laws could potentially be amended to clarify that bots and non-consensual deepfakes (which would address other much-needed protections) are also in scope.

Reporting recommendations

- We recommend the development of a public repository of ads; this is needed due to the newness of the technology and to evaluate the impact of the rules.
- Such external mechanisms could be updated to allow external contributors to support "active" defense measures against deepfakes; we recommend
  - That there should be a real-time update mechanism for this database so that emerging trends can readily be identified.
  - External sharing includes metadata and records of political advertisements for auditors and researchers in general, including post-hoc access available to identify whether the original media in the advertisement is AI-generated,
  - Support be provided to evaluate the patterns in advertisements shared by accounts producing electoral advertisements online, within reasonable privacy-preserving mechanisms similar to past sharing of privacy-preserving platform data at scale by Meta, Twitter, and others.
  - That advertisements verifiably detected by external researchers as deepfakes should have pathways for sharing such that they are actionable upon by platforms within a reasonable window of time.

Thank you

We appreciate the opportunity to assist the Federal Election Commission in its engagement on this timely and important topic. We thank you for your consideration of our recommendations and would welcome further public engagement on these issues.

Sincerely,

[23] California Bot Disclosure law:
https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1001

[24] Rhode Island Computer Crime code, §11-52-7.1
http://webserver.rilin.state.ri.us/Statutes/TITLE11/11-52/11-52-7.1.htm

Eric Davis
Integrity Institute Member
Advisor, Trust & Safety, Security, Privacy
Menlo Park, CA

Lucía Gamboa
Integrity Institute Member
Advisor, Public Policy, Trust & Safety
Austin, TX

Swapneel Mehta
Integrity Institute Member
Postdoctoral Associate, Boston University and MIT
Cambridge, MA

Diane Chang
Integrity Institute Member
Entrepreneur-in-Residence, Brown Institute for Media Innovation, Columbia University
New York, NY

Amari Cowan
Integrity Institute Member
Oakland, CA

Nichole Sessego
Integrity Institute Member
Phoenix, AZ

David Evan Harris
Integrity Institute Visiting Fellow
Chancellor’s Public Scholar, University of California, Berkeley
San Francisco, CA