Final Report
Role Of Deepfakes in Spreading in Political Misinformation

Team JARID
Johnny Sun, Ajay Byadgi, Rittivuth Chea,
Ihunna Onyekachiuzoamaka, Dan Aube, Neha Keshan

December 2025

Overview

AI has been rapidly improving over the past 50 years, which has introduced powerful tools that are capable of generating highly realistic audio, images, and videos known as deepfakes. While collaborating with Societal Surveyors (Group 3), we were able to get a lot of insight on biometrics and how these deepfakes are created. To make the deepfakes, raw data is used to train the model. Specifically for deepfake audio clips, the AI uses biometrics like voice, pitch, tone, inflection, pacing, and breathing noise to recreate highly realistic fake voices. The model then builds a digital "voice fingerprint" that captures phonetic habits. Although it seems complex, it is rather simple for the general public to make these clips as many models come pretrained and users just need to fine tune the model to create deepfakes of the person they are trying to copy.

As these technologies have been improving and are becoming more accessible, more of the general public has access to deepfakes, accelerating the spread of misinformation. Sora 2 is a great example of this, as many people were using Sora 2 to create funny and realistic AI generated videos to post on the internet, many times including political figures, all for likes and attention. The rapid growth in deepfake technology has resulted in a growing inability for the general public to distinguish between authentic vs synthetic content in the media. A study was done in the US that showed that less than 25% of the general public could detect "good quality" deepfakes [1].

This misinformation can be found all over the internet, on social media platforms with over millions of users, and even private messaging between friends and family. These deepfakes can spread incredibly quickly, reaching millions of people before platforms or fact checkers have any chance to respond. In fact, back in 2023, Keir Starmer, the Prime Minister of the UK, found himself in a predicament when a deepfake audio clip of him went viral, making it seem as though he was verbally abusing his staff members. Although this audio clip was debunked rather quickly, it still reached around 1.5 million views on X in the first 24 hours, showing just how fast misinformation can spread if it shines a negative light on a political member [2].

Many platforms use algorithms to capture user attention, meaning if a user is liking videos on a specific topic, the algorithm continues to give them more videos on that topic. These are known as echo chambers, and the issue with these is how easily they can constantly show the user false information, making the repeated false information feel true. The spread of misinformation has a lot of impact on politics, distorting public perception, weakening democracy, deepening division, and eroding social unity. These deepfakes can give people an incorrect view on a political candidate, swinging the polls from one simple video. Back in 2022 in Shreveport, Louisiana, Adrian Perkins claims he lost his election as mayor due to a deepfake advertisement portraying him as a high school student being chastised by the principal [3]. With Perkins' few resources and the limited news coverage at the time, the public was not able to discern fact from fiction. The spread of misinformation has become a widely believed issue, with over 95% of Americans believing it is an issue [4].

Something needs to be done to tackle these issues with low media literacy as well as the impacts of deepfakes and misinformation on the general public. That is why we have come up with two suggested solutions that directly deal with the issue.

Here is the link to the GitHub: https://notrittivuth.github.io/Data-and-Society/

Solution 1: Education of Deepfakes

It is now evident that tackling the underlying cause of user vulnerability is crucial given the scope of these effects and the speed at which deepfakes are increasingly influencing public opinion. The best course of action is to improve the public's capacity to assess information before interacting with it, rather than reacting only after false information has already proliferated. Enhancing media literacy and public education is a key long-term solution to the proliferation of false information and deepfakes. People are finding it more difficult to discern what is real online as AI-generated content gets more lifelike. Less than 40% of adults really fact-check the stuff they encounter on social media before sharing it, according to a recent Poynter media literacy poll [5]. This implies that the majority of people are susceptible to false information just because they lack the abilities or routines necessary to double-check the material they consume online.

People fall for fake or AI-generated content for a variety of reasons, according to research. The American Psychological Association claims that while interacting with posts, users frequently rely on emotion, identification, and group norms, which increases the likelihood that they would believe and spread false material that ”feels right” rather than factual information [6]. According to the BBC, fraudulent posts have a psychological benefit over accurate reporting since they are purposefully made to be dramatic, memorable, or emotionally stirring [7]. Yale Insights also demonstrates how algorithmic design on social media platforms worsens this issue: posts that are contentious or emotionally charged attract engagement and are therefore pushed to users more frequently, even when the material is untrue [8]. According to Pew Research, only 23% of American people are confident in their capacity to determine if material found online is accurate or false, which reflects a general lack of confidence in digital evaluation abilities [9].

These shortcomings are directly addressed by media literacy instruction. People's capacity to assess information responsibly is enhanced when they are taught how to recognize manipulated media, check publication dates, examine emotionally charged posts, and identify reliable sources. According to Somoray and Miller's experimental research, giving users explicit detection techniques greatly enhances their capacity to identify deepfakes, demonstrating that education dramatically boosts resistance to false information [10]. People can acquire useful verification methods, such reverse image searches or identifying indicators of digital tampering, through awareness campaigns, open workshops, and easily accessible online guidelines.

Encouraging people to rely on reliable sources, recognize bias, and double-check statements across different outlets is a crucial component of this strategy. Good journalism helps readers discern between fact-based news and deceptive content and provide the factual basis required to combat deepfakes that are making the rounds on the internet. Collaboration with Group 10, who stressed the significance of incorporating deepfake and disinformation education in school systems, strengthens this strategy. Students would learn early on how deepfakes are made, how algorithms affect what they see, and why fact-checking is important if media literacy were incorporated into current digital citizenship or social studies courses. Early education reduces the possibility that future generations would fall for or disseminate false information by preparing young people to use digital spaces appropriately. Education continues to be the most scalable, durable, and successful long-term defense despite obstacles like quickly developing AI and varying degrees of public interest.

All things considered, a stronger, more informed population that is better able to recognize, challenge, and fend off deepfakes and false information is produced by combining media literacy training, early education, awareness campaigns, and support for high-quality journalism. People become more adept at spotting manipulation, less likely to disseminate false information, and better equipped to navigate a digital world where AI-generated media is becoming more prevalent when these skills are developed early and reinforced in schools, communities, and online platforms.

Explanation on Solution 1

One of the most realistic long-term solutions is to educate the public. Most people still cannot reliably identify manipulated videos, and many do not regularly verify what they see online. Improving general media literacy would help create a more aware population and reduce the impact of misleading political content. The biggest benefit of this approach is that educated citizens are less likely to share harmful or inaccurate information. When people understand how deepfakes work and what warning signs to look for, they become much harder to mislead.

However, public education has clear limitations. Not everyone cares about misinformation, and many people will ignore educational materials entirely. There is also only so much that outreach campaigns, public service announcements, or optional training can accomplish. These efforts often rely on individuals choosing to participate, and many will not. To make this solution more effective, it would need some type of incentive or built-in benefit. Workshops could be connected to digital safety certification, or platforms could reward users for completing short training modules. Even small incentives can increase engagement and help overcome the problem of public disinterest.

A more structured version of this solution involves incorporating media literacy into schools. Teaching students about deepfakes at a young age would give them an advantage as they grow up in an environment where synthetic media is common. Introducing these concepts during middle school or high school could help students form good habits early, including checking sources and thinking critically about political content. This creates a generation that is better prepared for future digital challenges.

There are drawbacks to this approach as well. Creating new curriculum materials costs money, and schools would need to invest in training, lesson plans, and updated technology. It is reasonable to ask whether the expenses are worth it, but the long-term benefits likely outweigh the costs. Students who learn digital literacy become smarter online citizens and are less vulnerable to manipulation during elections. Another concern is that students might not take the class seriously or might not engage with the material. One way to address that issue is to include media literacy topics in standardized testing, which ensures that students treat it as an important academic subject rather than an optional add-on.

Overall, educating both the general public and younger students is a realistic and durable solution to the spread of political deepfakes. It is not perfect, and it requires investment and careful implementation, but it strengthens society's ability to resist misinformation in a way that technical or regulatory fixes cannot fully achieve.

Solution 2: Lawmaking to fight AI misinformation

A second major solution to limiting political deepfake misinformation involves creating stronger laws and clearer regulations that target both the AI-content-generating users and platforms where it spreads. With our goal of preventing political deepfakes from circulating without clear identification in mind, our approach focuses on transparency and platform responsibility, ensuring that AI-generated political content is easily distinguishable and managed before it can distort public opinion.

The first part of this solution is to require AI generators to watermark all synthetic content. At the moment, AI tools can create incredibly convincing audio, images, and videos with little to no built-in markers indicating they are artificial. However, traditional watermarks are not sufficient enough to handle the issue alone, since simple editing tools can easily remove them. Legislation should mandate multi-layering fingerprinting, which adds persistent digital markers across audio and visual content. These fingerprints survive common edits, making it harder to hide the origin of synthetic political deepfakes.

The second part of this solution focuses on social media platforms where misinformation spreads fastest. Platforms should be legally required to detect AI-generated content and visibly label it. These labels would warn users that the content may be inaccurate or manipulated. In addition, platforms should implement moderation tools specifically aimed at limiting the reach of misinformation. This includes algorithmic de-ranking, context flags, visibility warnings, etc. To prevent impersonation or fabricated speech, platforms should also adopt ID verification for political figures.

Our collaboration with Group 5 provided us with deeper insight into how AI is currently regulated across the EU and US, clarifying where existing policies succeed and where they fall short. The EU AI Act uses a safety focused model that includes mandatory labeling and developer obligations, showing us how enforceable rules can strengthen transparency. The US Executive Order promotes watermarking and election protections, but lacks strict requirements. Group 5's analysis helped us recognize that combining our ideas with the two frameworks gives our proposal a clearer, and more realistic direction.

Together, by enforcing durable identifiers, requiring platform-level labeling, and establishing clear accountability for companies, these legal measures offer a structured path toward reducing the impact of AI-generated misinformation on political communication.

Explanation on Solution 2

As stated before, new litigation can be used to take control of AI misinformation and limit its negative impact on the internet, and on politics in particular. There are multiple ways of attacking this problem, some originating at the root of it all: the AI content generators. What initially makes sense to ensure AI generated photos and videos are not to be confused with real, authentically filmed content is to watermark AI generated content. Sora 2 AI is a great example of this, with its clearly seen watermark that labels it as AI generated with the "Sora" and its logo. As such, a law could be passed to require AI content generators to put a clearly visible watermark on any content generated by them. However, there is a particularly troubling issue with this: watermarks can be easily removed. For photos, it's as simple as one swipe of the spot healing brush tool in Photoshop, and the watermark is gone. For videos, it might be more difficult, say, the prior-mentioned done once for each frame of the video, or simply using one of many online tools for removing watermarks from videos. Luckily, there is a more permanent way to label AI content beyond watermarking: multi-layer fingerprinting. This fix involves embedding meaningful noise into the video, audio, and metadata so as to create a sort of digital 'fingerprint' that will survive attempts to alter the video such as via changes in brightness, color, volume, and other typical alterations [11], while revealing clear evidence of tampering [12] when the video is severely manipulated in an attempt to remove the fingerprint. The fingerprint could be detected by social media platforms, as could evidence of tampering with such a fingerprint. Therefore, we would propose that a law be passed requiring AI photo and video generators to both watermark and fingerprint all generated content.

Another strategy for curbing the spread of AI misinformation is through the social media platforms themselves. Assuming we first have the cooperation of AI generators regarding the digital fingerprinting, the mere detection of most AI content would be trivial. As such, I would propose a law requiring social media platforms to visibly and obviously display a flag labelling any AI generated content, including a warning about the potential inaccuracy of AI generated information. ChatGPT offers a good example of this, with its disclaimer "ChatGPT can make mistakes. Check important info" [1]. Some platforms are already deploying some form of labelling or flagging of AI generated content on their sites, such as TikTok and YouTube [13]. Beyond just this however, we can employ moderation strategies to discourage the spread of AI misinformation specifically. We would suggest passing a law requiring social media platforms to moderate AI in all of the following ways: misinformation warning flags, an algorithmic deranking system for posts containing misinformation, punishment for accounts that are repetitive offenders of misinformation, and, to combat AI misinformation in politics, ID verification for political figures. For example, if a post is detected to contain misinformation, a flag will display warning viewers of such. An algorithm will sandbag the post's popularity, discouraging people from making misinformation posts for ragebait interaction or likewise. Accounts that attempt to farm views by spreading misinformation will be punished with temporary or permanent mutes, bans, or likewise to prevent them from becoming a problem. Lastly, political figures and their personas will be protected by ID verification, so impersonations, deepfakes, and fake accounts will be quickly detected and dealt with.

Group 5 offers us some insight on how AI is regulated in the real world. In the USA, the government employs an "innovation-first" approach, letting companies have free reign over their creation and usage of AI, and relying on existing legal frameworks (i.e. the judicial system) to take on specific cases of potential misuse, ruling on said cases under laws not specific to AI [14]. The EU applies a "safety-first" approach, creating strict, binding rules as soon as they can, aimed at preventing harm. The cornerstone is the EU AI Act, which categorizes AI systems by risk level and imposes mandatory requirements on high-risk systems. This includes regulating transparency, data quality, human oversight, documentation, cybersecurity, monitorization of deployment [15]. Although our own set of legal propositions lack the comprehensiveness of these real-world examples, our strategy can be more closely compared to that of the EU.

Group Contribution

- Ajay: For the midtech report, did the issues handled section. For the final report presentation, had a slide on misinformation impact on politics and society as well as half a slide about solution of watermarking and embedding as well. For 4.3 I worked on the explanation on solution 1 which included explaining specifics of the solution as well as positives and drawbacks.

- Dan: For the midtech report, wrote the abstract and researched, created slides for, and found sources on public awareness of misinformation and media literacy, added a section about collaborating with group 3 in the doc and slideshow. For the final report presentation, created the title slide, a slide on the public awareness of misinformation, a slide on our collaboration with group 3, and a slide on media literacy. For 4.3, worked on the explanation for lawmaking to fight AI generated misinformation, including requiring AI generators to watermark content, requiring social media platforms to moderate AI content, and our collaboration with group 5 on US vs EU policies on AI.

- Ihunna: For the midterm report, I contributed by summarizing the current updates on the project and outlining our progress at that stage. For the final presentation, I created the accountability slide and the section explaining why misinformation spreads, supported by credible research. For section 4.3 of the final report, I wrote Solution 1, which focused on education about deepfakes, media literacy statistics, spreading awareness, promoting good journalism, and incorporating insights from Group 10 about teaching these skills in schools.

- Johnny: Worked on Hurdles facing, found examples of political deepfake misinformation and found articles for it. Also worked on the current plan for the remaining week. Proofread midtech report. Created slides on final presentation and mid tech report. Did research for specific examples like Adrian Perkins and Keir Starmer. Also created outline for assignment 4.3, helping group members choose to work on something they wanted to. Distributed work evenly to help make sure everything was done on time. Finished working on the overview for the final, proofread every part and made the overleaf file. Moved everything from the doc to this overleaf file for submission, as well as ensured everyone was on top of their work.

- Ritti: On the midtech report, I worked on our expected results, outlining how transparency, labeling, platform accountability, and policy frameworks could limit the influence of political deepfakes. I also created the Overleaf project, organized document structure into a single file for submission. For the final presentation, I focused on explaining what AI deepfakes are and how they work and presented our collaboration with group 5, summarizing how the EU AI Act and US Executive Order informed our proposed solution. For this written report, I contributed the overview of solution 2, outlining how lawmaking can reduce AI misinformation through required labeling, multilayer fingerprinting, platform moderation, and clearer accountability for AI companies and social media companies.

Acknowledgments

This project was completed as part of a Data and Society course. We thank our teaching assistant, Srihari Sridharan, for guidance and feedback throughout the semester, and our classmates for discussions that improved our framing and revisions.

We acknowledge insight from peer groups whose focus areas directly informed parts of our report.

- Teaching Assistant, Srihari Sridharan: Provided guidance, clarification, and feedback throughout the semester, helping refine our problem framing, solution scope, and revisions as the project evolved.

- Group 3, Societal Surveyors(Steven Cadogan, Daniel Krupp, Liana Lloyd-Wood, Joseph Celtruda, David Li): Their work informed our explanation of deepfake construction through biometric signals in audio, including voice features and the idea of a digital voice fingerprint.

- Group 5, Algo Avengers(William Shin, Matthew Lausev, David Smith, William Sweet, Nevin Joshy): Their policy comparison informed our lawmaking solution, including labeling requirements, watermarking, and stronger platform accountability with EU AI Act versus US Executive Order context.

- Group 10, Red Flags(Isadora Lee, Timothy Liakh, Doan Nguyen, Hazel Yu, Michael Lam): Their emphasis on public perception and safety strengthened our education solution, especially the case for teaching media literacy and deepfake awareness in school systems.

References

[1] OpenAI. (2025) ChatGPT. [Online]. Available: https://chatgpt.com/

[2] J. Sturcke. (2024) Uk's keir starmer and labour party hit by 'deepfake' ai politics risk. [Online]. Available: https://www.politico.eu/article/uk-keir-starmer-labour-party-deepfake-ai-politics-elections/

[3] R. Salama. (2024, Nov.) Election 2024: Ai is a gift and a curse for down-ballot campaigns. [Online]. Available: https://apnews.com/article/artificial-intelligence-local-races-deepfakes-2024-1d5080a5c916d5ff10eadd1d81f43dfd

[4] The Associated Press-NORC Center for Public Affairs Research. (2021, Oct.) The american public views the spread of misinformation as a major problem. [Online]. Available: https://apnorc.org/wp-content/uploads/2021/10/misinformation_Formatted_v2-002.pdf

[5] Poynter. (2023) Survey identifies media literacy skills gap amidst rise in ai-generated content. [Online]. Available: https://www.poynter.org/fact-checking/media-literacy/2023/adults-worry-misleading-ai-images-lack-media-literacy

[6] American Psychological Association. (2023, Nov.) How and why does misinformation spread? [Online]. Available: https://www.apa.org/topics/journalism-facts/how-why-misinformation-spreads

[7] BBC Bitesize. How false information spreads. [Online]. Available: https://www.bbc.co.uk/bitesize/articles/zcr8r2p

[8] S. Allen. (2023, Mar.) How social media rewards misinformation. [Online]. Available: https://insights.som.yale.edu/insights/how-social-media-rewards-misinformation

[9] Pew Research Center. (2022) Americans' Confidence in Identifying Misinformation. [Online]. Available: https://www.pewresearch.org/internet/2022/

[10] K. Somoray and D. M. Miller, "Providing detection strategies to improve human detection of deepfakes: An experimental study," *Computers in Human Behavior*, vol. 149, p. 107917, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0747563223002686

[11] Coalition for Content Provenance and Authenticity (C2PA). (2025) C2PA Specification 2.2. [Online]. Available: https://spec.c2pa.org/specifications/specifications/2.2/specs/C2PA_Specification.html

[12] ——. (2025) C2PA 2.2 Explainer. [Online]. Available: https://spec.c2pa.org/specifications/specifications/2.2/explainer/_attachments/Explainer.pdf

[13] Kinesso. (2024) How social media is labelling ai-generated content. [Online]. Available: https://kinesso.co.uk/insights/how-social-media-is-labelling-ai-generated-content/

[14] C. Rosello and A. Frank. (2024) Regulating general-purpose ai: Areas of convergence and divergence across the eu and the us. [Online]. Available: https://www.brookings.edu/articles/regulating-general-purpose-ai-areas-of-convergence-and-divergence-across-the-eu-and-the-us

[15] European Parliament. (2023) EU AI Act: First regulation on artificial intelligence. [Online]. Available: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence