ISPI Analysis March 2024

AN OVERVIEW OF THE IMPACT OF GEN*AI* AND DEEPFAKES ON GLOBAL ELECTORAL PROCESSES

Enzo Maria Le Fevre Cervini and **Maria Victoria Carro**

ISPI
ITALIAN INSTITUTE
FOR INTERNATIONAL
POLITICAL STUDIES

An Overview of the Impact of GenAI and Deepfakes on Global Electoral Processes
March 2024
by Enzo Maria Le Fevre Cervini and Maria Victoria Carro

The views expressed by the author are strictly personal and do not necessarily reflect those of the European Commission

ISPI
via Clerici, 5
20121 Milano

AN OVERVIEW OF THE IMPACT OF GEN*AI* AND DEEPFAKES ON GLOBAL ELECTORAL PROCESSES

Enzo Maria Le Fevre Cervini and **Maria Victoria Carro**

ISPI

Enzo Maria Le Fevre Cervini, Ph.D., is Project Leader in the European Commission's Directorate for Informatics (DG DIGIT) and Adjunct Professor of Digital Governance at the Università degli Studi della Tuscia. Between 2017 and 2020 served as Senior Expert at the Agency for Digital Italy – Presidency of the Council of Ministers (AGID) and Coordinator of the Thematic Group on Emerging Technologies (AI & Blockchain) of the OECD Working Party of Senior Digital Government Officials (E-Leaders). He is also a Senior Research Fellow on international practices of digital governance at the Laboratory on Innovation and Artificial Intelligence of the University of Buenos Aires, and Board Member of the Fondazione Adriano Olivetti.

Maria Victoria Carro, is a PhD candidate of University of Genoa and Tarello Institute of Philosophy of Law. Research Director at Laboratory on Innovation and Artificial Intelligence of the University of Buenos Aires (UBA IALAB). She is a lawyer from the University of Buenos Aires.

INTRODUCTION

Recent viral photos depicted Pope Francis clad in a white padded jacket adorned with the logo of the Argentine National Team, while former President Donald Trump was seen being detained and escorted by two police officers.[1] These images gained widespread attention not only due to the significance and uniqueness of the events involving two globally recognized figures but also due to the astonishment and outrage felt by users upon discovering they had been deceived by artificial intelligence.

Generative Artificial Intelligence's (GenAI) capacity to produce highly realistic images, videos, and text poses a significant challenge, as it can deceive viewers and consumers into accepting artificially generated content as authentic and genuine. This raises concerns about the dissemination of false information, disinformation, and its implications for public trust and democratic processes. Additionally, this phenomenon prompts critical ethical and legal inquiries, including issues surrounding the attribution of authority and accountability for the generated content.

In 2024, nearly half of the world's population will determine the future of their countries in the biggest election year in history. During the initial days of June, citizens across the European Union will exercise their democratic right by electing representatives to serve as deputies in the European Parliament. Additionally, Croatia, Lithuania, and Belgium are slated for general elections later this year, while Austria prepares for legislative elections.

On the other hand, electoral processes will also be underway in Russia, Indonesia, India, Taiwan, the United Kingdom, and even in totalitarian Iran. Notably, also the United States, will face elections later this year.[2]

Coincidentally, 2024 also marks a significant year in the development of AI, with multimodal generative models like ChatGPT, Gemini, Sora, and Midjourney achieving new heights in both technical capability and public recognition and adoption. While in general the advancement of AI should be seen as an incredible source of opportunities, this technology could lead as well to unprecedented challenges in our societies. Technology in any case should not be viewed in isolation as a standalone threat; rather, it should be recognized as a catalyst amplifying existing threats. Here's why:

- Accessibility: With increasingly sophisticated generative AI tools freely available to the public, individuals lacking technical expertise can now effortlessly produce deepfakes, widening the pool of potential creators.
- Velocity: GenAI enables the rapid creation of deepfakes, rendering them harder to identify and eradicate before they spread.
- Scalability: means deepfakes can be generated on a vast scale, amplifying their societal impact. Furthermore, social networks facilitate their dissemination and virality, circumventing traditional mass media channels.

This article delves into the impact of generative AI on recent and future political elections. We'll examine how deepfakes and other AI-generated content are used, along with their potential to sway voters. We'll also analyze the strategies various stakeholders are deploying to counter this growing phenomenon.energy sector, evaluates them based on four strategic criteria, and ranks them into three tiers representing short-, medium-, and long-term opportunities.

Driven by the three key trends of decarbonization, decentralization, and digitization, the energy sector faces numerous challenges. Stakeholders along energy value chains have often struggled to adapt to technological and economic shifts, but doing so will be crucial for addressing climate change. As technologies mature and their costs decline, the energy transition will accelerate, resulting in further development and deployment of energy technologies and infrastructure. A path toward a more sustainable, efficient, cost-effective, secure, and reliable energy future will be possible only with a strategic approach and the alignment of the public and private sectors.

.

THE INFLUENCE OF DEEPFAKES ON GLOBAL ELECTORAL PROCESSES

Deepfakes, a term derived from "deep learning" and "fake," also known as synthetic media, are a type of AI-generated content. They can be audio, images, or videos, and are designed to realistically imitate a person's appearance and behavior, often with the intent to deceive viewers.

In the falsification of images, figures such as Donald Trump, Joe Biden and Pope Francis have been protagonists. Trump has been digitally depicted both in altercations with law enforcement officers, purportedly being arrested,[3] and engaging with black voters in efforts to secure Republican support.[4] Biden was shown wearing military gear in the Situation Room, sitting at a desk with advisors.[5] As a result, the image generative AI platform Midjourney is contemplating a ban on the creation of images featuring political content.[6]

What is more, an image depicting Pope Francis donning an elegant white jacket by the Balenciaga brand garnered rapid viral attention upon its release in March 2023[7]. Even Taylor Swift has been a victim of controversy when fake pornographic images of the singer were spread by X.[8]

Regarding audio falsifications, as example we have identified three recent electoral processes in different countries in which they have been used: the United States (2024), Slovakia (2023) and India (2023).

In January, New Hampshire voters were startled by a call claiming to be from President Biden himself, urging Democrats to refrain from voting in the upcoming primary elections. The voice asserted, "Your vote matters in November, not this Tuesday".[9] However, it soon became evident that the voice was not Biden's, but rather an artificial creation.

On the one hand, suspicions pointed to ElevenLabs, a company that promptly faced repercussions banning the account involved. On the other hand, investigations revealed the involvement of a Texas telemarketing firm, now under scrutiny by New Hampshire's attorney general for alleged voter suppression tactics. Moreover, the Federal Communications Commission swiftly deemed robocalls utilizing AI-generated voices unlawful, signaling potential purposes and legal actions against offenders under federal telecommunications regulations.

Secondly, just two days prior to Slovakia's election, a troubling audio recording surfaced on Facebook featuring purported voices of Michal Šimečka, leader of Slovakia's liberal Progressive party, and Monika Tódová from the newspaper Denník N.[10] In the recording, they appeared to be discussing strategies to manipulate the election, including the unethical practice of purchasing votes from the marginalized Roma minority.

Šimečka and Denník N promptly refuted the authenticity of the audio, branding it as a forgery. The AFP news agency's fact-checking division noted indications of AI manipulation in the recording. However, its release during the "quiet period" before elections, a time when media and politicians are expected to refrain from influencing voters, made it challenging to debunk widely under

Slovakia's electoral regulations. Furthermore, exploiting a loophole in Meta's manipulated media policy, which primarily addresses deepfake videos, the audio escaped immediate scrutiny.

This underscores a critical aspect of the detrimental impact posed by deepfakes: heightened concern particularly arises in the immediate lead-up to an election, a time when journalists and campaigns have minimal opportunity to thoroughly verify facts or claims, amplifying the potential for misinformation to discredit the electoral process.

Subsequent investigations revealed that the forgery had been facilitated by technology from Eleven Labs, a company that offers an AI tool enabling instantaneous conversion of text into natural voice speech. Consequently, Facebook introduced a label accompanying the post, cautioning viewers about potential misinformation.[11] Despite this, users were still permitted to access the video at their discretion.

In the aftermath, it was announced that the pro-NATO party, Progressive Slovakia, had been defeated by SMER, which advocated for the withdrawal of military support from neighboring Ukraine. Given the timing of the deepfake's release, there exists a possibility that the manipulation indeed influenced the election outcome.

Finally, in a state in southern India, the leader of a political party, which holds power at the national level, made public audio recordings of a legislator from the opposing party alleging illicit accumulation of $3.6 billion by members of his own party.[12] The individual whose voice was recorded vehemently denied the authenticity of the recordings, labeling them as "fabricated" and "machine-generated." Subsequently, another clip surfaced, also refuted by the legislator as false.

To ascertain the validity of the recordings, three independent experts in deepfake technology were enlisted to conduct an analysis. Their findings indicated that the second clip was genuine but raised doubts regarding the authenticity of the first clip, suggesting potential manipulation.

In this scenario, it is striking that the use of deepfakes occurs in the opposite direction: a politician exploits the issue of artificial generated content to discredit inconvenient information. If widespread awareness of the gravity of this phenomenon takes hold, fear and public perception could be manipulated to undermine trust in all forms of information.

The concept of the "liar's dividend",[13] illustrates this dynamic. It suggests that as the public becomes increasingly aware of the possibility of convincingly forged video and audio, some individuals may seek to evade accountability for their actions by dismissing genuine audio and video evidence as deepfakes.[14]

This tactic not only undermines institutions and corrupts democratic principles but also creates an atmosphere of uncertainty regarding the reliability of information. Indeed, by branding authentic media as fake, malicious actors can deflect blame onto their political adversaries. This phenomenon gains potency as familiarity with deepfakes grows. Consequently, media consumers may become predisposed to dismiss legitimate campaign messages.

Hence, it is paramount that the public maintains confidence in its ability to discern between authentic and manipulated media. Without this trust, the foundations of informed decision-making and democratic discourse are jeopardized. In a subsequent section, we will delve into the strategies being devised to uphold the ability to differentiate between authentic and manipulated media.

Thus far, our attention has been directed towards audio manipulations, but the most impressive cases of deepfakes are those that use videos, where not only a person's voice is replaced, but also his/her face and expressions in an almost indistinguishable way. In these second type of situations we have identified four other cases in different countries: the United Kingdom (2023), Turkey (2023), South Korea (2024) and Indonesia (2024).

In the first place, a Facebook video surfaced, seemingly depicting BBC news presenter Sarah Campbell discussing UK Prime Minister Rishi Sunak's endorsement of a new investment venture by Tesla CEO Elon Musk.[15] In the footage, Campbell asserts that Musk's initiative aims to combat global poverty, with Great Britain being its inaugural target. She then attempts to elicit Sunak's response. The clip transitions to Sunak addressing the purported project from a podium, where he remarks, "I and my entourage have been earning through this project for several months".[16]

However, investigation revealed this to be an AI-generated deepfake, and Musk had not introduced any such investment projects for British citizens. The original videos of both the journalist and the prime minister discussing unrelated topics served as the foundation for manipulation and synthesis into the deepfake.

It's estimated that the video potentially reached over 400,000 users and generated more than 100 advertisements, resulting in a disbursement by individuals exceeding £12,929 across 23 countries, including the US, Turkey, Malaysia, and the Philippines.[17]

The second scenario took place in Turkey. In the midst of a contentious election season, incumbent President Erdoğan found himself in a runoff against his primary rival, Kemal Kılıçdaroğlu. Throughout the campaign, Erdoğan's team circulated a manipulated video portraying candidate Kılıçdaroğlu purportedly participating in demonstrations alongside militants from the Kurdistan Workers' Party (PKK), an organization designated as a terrorist entity by both Turkey and the United States.[18]

Despite the video's obvious falsification, it gained significant traction among the public and permeated search engine results for the opposition candidate. On the day in question, Google's algorithm prominently suggested fake news stories when users queried Kılıçdaroğlu, contributing to the dissemination of misinformation during a critical juncture in the electoral process.

South Korea's National Election Commission revealed that it identified 129 illicit posts employing AI-generated deepfakes—fake videos and audio—in the lead-up to the April general election.[19] Following 19 days of intensive monitoring, the NEC uncovered a pattern: the bulk of these deepfakes had been disseminated across social media channels. These deceptive videos primarily targeted opposition candidates, with manipulations ranging from distorting segments of their speeches to outright fabrication.

Finally, in Indonesia earlier this year, the electoral landscape witnessed a flood of deepfake videos. While some of these videos originating from Indonesia are relatively easy to debunk, others present a more nuanced challenge. For instance, a video that circulated widely in January depicted Suharto, the late former president of Indonesia, purportedly endorsing his former political party, Golkar[20]. Meanwhile, there are subtler instances, such as one featuring Joko Widodo, the outgoing president, supposedly speaking in Mandarin. Additionally, there have been videos portraying two candidates for the 2024 election conversing in Arabic.

In each of these instances, we've encountered prominent cases of deception. Given the national recognition of these personalities, mechanisms have swiftly been set in motion by both organizations and the media to authenticate the veracity of such information. However, in local or regional elections, the impact of AI-generated fakes could potentially be more pronounced compared to national elections. Candidates in such elections typically enjoy less prominence and may not receive the same level of scrutiny from the media or the public. Consequently, misinformation and fake content propagated through AI could significantly sway voter opinion and outcomes.

Unlike national elections, where there is often intense media coverage and robust fact-checking mechanisms in place, local elections may lack the same level of scrutiny and resources. Local media outlets might have limited capacity to thoroughly investigate every piece of information, leaving room for false narratives to proliferate unchecked. Additionally, local communities may have fewer resources allocated for monitoring and combating misinformation, exacerbating the vulnerability to AI-generated fakes.

Furthermore, in the context of local elections, there may be a scarcity of expertise available to discern between genuine information and fabricated content. Unlike national elections, where seasoned journalists and fact-checkers may be more prevalent, local communities may lack the specialized skills and resources necessary to effectively combat the spread of misinformation.

As a result, AI-generated fakes could pose a significant threat to the integrity of local and regional elections, undermining trust in the democratic process and potentially altering electoral outcomes. It underscores the urgent need for enhanced awareness, collaboration between stakeholders, and the development of robust countermeasures to safeguard the integrity of elections at all levels.

For instance, in Australia, Queensland is gearing up for its state and local elections in 2024. In 2020, a fabricated video purportedly depicting Prime Minister Annastacia Palaszczuk delivering a press conference surfaced.[21] While the video was clearly fake at the time, technological advancements have since made such fabrications more sophisticated. However, the State has adopted specific legislation that can discourage the creation of this kind of content. Section 185 of the Electoral Act 1992 makes it an offense to publish false statements about a candidate or to mislead voters during the electoral period for an election by printing, publishing, distributing or transmitting.[22] This includes publication on the Internet, even if the Internet site on which the publication is made is located outside of Queensland.

However, generative AI also holds promise in enabling the creation and distribution of authentic information. Local campaigns lacking substantial funding can leverage this technology to swiftly craft messages tailored to a community's unique vernacular. New York City Mayor Eric Adams, for instance, has employed artificial intelligence software to translate his voice into languages like

Yiddish, Spanish, and Mandarin.[23] As a result, participating in this form of mass customization has become incredibly cost-effective.

Another crucial aspect to consider is that, while the majority of instances highlighted pertain to the manipulation of men, the chilling effect of AI-generated images and videos used to target women in politics is becoming increasingly prevalent. The ramifications of online harassment against women and other marginalized groups, including the creation and circulation of sexually explicit content through artificial means, will undoubtedly stifle the political aspirations and involvement of women and girls. This will inevitably reduce their visibility and influence in political spheres and public discourse[24].

POLITICAL BIASES IN LARGE LANGUAGE MODELS (LLMS)

There has been a lot of research trying to measure the presence of political biases in large language models like ChatGPT.

Firstly, a study from Stanford University can be mentioned. This research highlights a troubling misalignment between the perspectives embedded within these models and the viewpoints held across diverse demographic segments in the United States[25]. These disparities serve as poignant indicators of biases present within language models, particularly concerning contentious topics, which may deviate from the prevailing public sentiment.

For instance, recent studies reveal that newer models exhibit over 99 percent approval for President Joe Biden, despite public opinion polls painting a more nuanced picture. Moreover, the research identifies certain demographic groups, such as individuals aged 65 and older, Mormons, and widows and widowers, as underrepresented in the data. The researchers advocate for improved credibility in language models, emphasizing the need for these systems to better capture the intricate nuances and diverse perspectives of public opinion.

This recent conclusion aligns with recent analyses released by UNESCO. These studies highlight two significant findings: firstly, the exacerbation of biases against women by LLMs,[26], and secondly, the necessity for artificial intelligence systems to integrate the perspectives and needs of indigenous communities.[27] It is imperative to envision an AI ecosystem that is participatory, multicultural, and egalitarian in nature.

Notably, models trained solely on internet data often exhibit biases favoring less educated, low-income, or conservative viewpoints. Conversely, newer models, refined through curated human feedback, tend to lean towards more liberal, higher-educated, and higher-income audiences.

It is crucial to acknowledge the methods employed by large companies like OpenAI to ensure model alignment with human values and preferences, including public opinion. While reinforcement learning from human feedback is utilized for this purpose, it's important to recognize that the subjectivities of the labeling group may not adequately represent society at large.

Secondly, in August 2023, another inquiry unveiled significant and systematic political leanings within ChatGPT.[28] The findings revealed a notable bias favoring the Democratic Party in the United States, Lula's political faction in Brazil, and the UK's Labor Party. In essence, these researchers concluded that ChatGPT demonstrates a clear inclination towards the left side of the political spectrum.

All this evidence shows that users should be cautioned about these inherent biases in language models. However, striving for complete alignment with the opinions of the general public poses its own set of challenges and complexities, warranting careful consideration in the pursuit of model improvement.

STRATEGIES AGAINST DEEPFAKE THREATS FROM DIFFERENT STAKEHOLDERS

Various stakeholders employ distinct strategies to mitigate the detrimental impact of deepfakes. Firstly, companies developing generative AI tools can incorporate watermarks to identify artificially generated content. Secondly, social media platforms, owned by different companies, play a crucial role in combating the spread of deepfakes by swiftly removing false content to prevent its virality. Thirdly, governments enact legislation to hold individuals or entities accountable for creating artificial content and to establish lawful requirements governing its generation. Finally, media and journalists are the ones in the best position to verify the facts. Each of these strategies will be addressed below, examining their advantages and challenges.

Content moderation. The amount of damage generated by deepfakes is heavily influenced by the platforms where they are disseminated, primarily owned by major companies. These social networks enforce their own content moderation policies and employ various strategies to combat the spread of deepfakes. Typically, they utilize AI systems to automatically identify and remove content that violates their guidelines. However, in cases where the violation is ambiguous, human moderators are relied upon to make informed decisions based on contextual analysis.

These policies are not static; rather, they are subject to unilateral modification by the platform-owning companies as deemed necessary. For instance, major platforms such as Facebook, YouTube, and TikTok outline criteria for removing manipulated content, restricting action to instances of "egregious harm" or content intended to deceive people regarding electoral processes. This trend reflects a broader relaxation of moderation standards, exemplified by the rescinding of 17 policies related to hate speech, harassment, and misinformation across these platforms in the past year.[29]

In response to AI-generated content, platforms like Facebook, YouTube, and TikTok plan to implement labeling systems to indicate such content's artificial origin. Conversely, X (Twitter) will label content identified as "misleading media" based on recent policy updates.[30] However, these labels have not yet been rolled out and will require users' adaptation over time. It's important to note that AI-generated content may evade detection by moderation teams, potentially remaining untagged or unremoved, thus providing false reassurance to users. Additionally, aside from X's policy, these labels do not explicitly indicate whether the content is harmful or false; rather, they simply denote its AI-generated nature for the user to interpret.

Specific Deepfakes Legislation. Regulating deepfakes poses complex challenges. Firstly, regulation must address what specific conduct should be regulated, as deepfakes can range from manipulating content to spread misinformation to the unauthorized use of individuals' images. Additionally, regulation must balance the protection of freedom of expression with the need to prevent harm to reputation, security, and public trust.

Manipulated content, though often viewed with suspicion, can indeed fulfill legitimate and non-misleading purposes. Examples include its use in satire, commentary, or artistic expression. Such forms of speech carry intrinsic value and, notably in the United States, warrant significant legal safeguarding under the First Amendment. U.S. law specifies that even outright deception with no redeeming artistic or other licit purpose, while typically entitled to less constitutional protection, cannot be prohibited simply for its own sake[31]. The government must still provide an independent justification for any restriction and demonstrate that the restriction in question is appropriately tailored to its stated goal.

Regarding who should be regulated, responsibility falls on a variety of actors, including creators of deepfakes, online platforms that host and disseminate them, and users who share them. Regulations should be designed to be applicable and effective in addressing all stages of the production and dissemination chain of deepfakes while avoiding imposing excessive restrictions on freedom of expression and technological innovation.

Targeting deepfakes is gaining traction across the United States, with 27 states' legislatures taking proactive steps to regulate their use in elections, often garnering bipartisan support.[32] Notably, bipartisan backing has been a consistent feature of these bills in every state where they've been enacted.

These State bills typically fall into two main categories: disclosure requirements and prohibitions.[33] Disclosure requirements mandate the placement of disclaimers on any AI-generated media intended to influence an election within a specified timeframe.

Prohibitive measures, on the other hand, aim to ban deepfakes during a designated period, usually 60 or 90 days preceding an election. Such bans often include exceptions under specific conditions. For instance, a law passed in Michigan last year exempts media from the ban if a disclosure has been shared and the creator of the content is unaware that it "falsely represents" the individuals depicted.

For instance, a proposed bill by Republicans in Kentucky seeks to establish definitions for deepfakes and prohibits their dissemination without consent from the depicted individuals, while affording them the right to pursue legal recourse for redress and damages.[34] Another example is the Minnesota law under which anyone who widely shares a deep fake within 90 days of an election is guilty of a crime if the person, both: knows or should have known that the item was a deep fake made without the consent of the depicted individual; and acts with the intent to harm a candidate's reputation or influence the result of an election.[35]

Other nations around the globe are embracing similar measures. In Brazil, the Superior Electoral Court has set forth regulations governing the use of artificial intelligence in electoral settings, including an outright prohibition, even with authorization, of employing deepfake technology in electoral propaganda for the 2024 elections[36]. Any utilization of artificial intelligence for electoral purposes other than this must be clearly disclosed to the public. Furthermore, the use of chatbots and avatars to facilitate campaign communication has been limited; they are prohibited from simulating interaction with candidates or real individuals. Candidates found employing deepfake technology during their campaigns risk having their candidacies blocked or, if elected, facing termination of their terms.[37]

Content labeling and watermarks. The new European AI Act[38] mandates content labeling and watermarks for combating deepfakes. According to the law, implementers of AI systems generating or manipulating image, audio, or video content—such as deepfakes—must disclose their artificial origin. Additionally, individuals interacting with chatbots must be notified, and system providers are obliged to mark synthetic content in a machine-readable and detectable format. The prescribed solution must be efficient, interoperable, and technically feasible, ensuring robustness and reliability.[39]

Similarly, content labeling stands out as a significant provision in the Executive Order issued by President Biden in the United States. Under this order, the Department of Commerce is tasked with formulating guidelines for labeling AI-generated content, aiding companies in developing classification and watermarking tools.[40] Moreover, major AI firms, including Google and OpenAI, have voluntarily committed to the White House to develop technologies for classifying artificially generated content.

Content classification tools. Content classification tools currently face significant challenges in reliably tagging text or determining whether content has been generated by machines. AI detection tools remain susceptible to deception for several reasons.

Firstly, generative AI systems are trained on human-written text and are engineered to mimic natural language or generate other types of content. The difficulty in distinguishing their outputs from human productions is indicative of their success in achieving this objective. Consequently, achieving 100% certainty in detecting AI-generated text is highly improbable.

Secondly, content classification tools don't yield binary answers; instead, they provide probabilities. These tools typically identify hallmarks of AI-generated text, such as repetition, and calculate the likelihood that the text is machine-generated. However, interpreting these probabilities poses challenges. For instance, if a tool indicates a 51% probability of AI generation, how should this information be interpreted? Is the content deemed a deepfake or not? Communicating these nuances to the audience is crucial yet complex.

The AI writing detection feature serves to alert users to the presence of AI-generated text, prompting further examination. However, it does not determine the appropriateness of AI writing usage or whether it constitutes deception or misconduct.[41]

Moreover, these classification tools often operate as black boxes, lacking transparency in their decision-making processes.

Watermarks. Watermarks are a technique used to embed a signal within a text or image, indicating that it was generated by AI. Traditionally, images have been watermarked through visible overlays or by encoding information in their metadata. However, this approach is susceptible to being compromised; watermarks may be lost when images are cropped, resized, or otherwise edited.[42]

In response to this challenge, Google introduced the SynthID watermarking tool in August 2023, following discussions with the White House.[43] SynthID employs two neural networks: the first subtly modifies pixels in the original image, creating an embedded pattern invisible to the human eye, while the second neural network detects this pattern. This enables users to determine whether an image contains a watermark, is suspected of containing one, or is devoid of any watermark. Importantly, SynthID is engineered to detect watermarks even after an image is edited, cropped, or resized.

However, implementing such watermarks poses challenges, particularly for AI companies integrating them into chatbots. Although OpenAI is developing these systems, they have yet to be incorporated into any products. This delay is because of concerns regarding the appropriateness of watermarking all AI-generated text.[44] Not all uses of AI-generated text are intended to deceive or mislead;[45] for instance, ChatGPT could serve as a helpful tool for composing emails or improving spell-checking in word processors. Automatically flagging all AI-generated text with watermarks could lead to unwarranted suspicions and accusations.

Navigating the decision of when to require watermarks on AI-generated content is complex and requires careful consideration of various factors, including the context and intent of the content's usage.

Data verification organizations. Firstly, The role of media and journalists in combating deep fakes is crucial in the modern landscape of information dissemination. By raising awareness about this technology and its implications, they empower individuals to critically evaluate the authenticity of the content they encounter online. But even more important is the fact that journalists are trained to verify information and fact-check sources. In the era of deep fakes, this skill becomes even more critical. Media organizations can dedicate resources to develop tools and methodologies for detecting and debunking deep fakes, thereby preventing the spread of misinformation.

Journalists often rely on established techniques to discern the authenticity of content. These strategies include consulting native speakers to verify audio, examining metadata,[46] and utilizing online tools like "AI or not" for initial assessment.[47] While these methods offer valuable insights, they serve as starting points for further investigation rather than definitive conclusions.

Media professionals adhere to ethical standards that prioritize accuracy, fairness, and objectivity. In the face of deep fakes, maintaining these principles becomes paramount. Journalists must exercise caution when reporting on potentially manipulated content, ensuring that they do not unwittingly amplify false narratives.

What is more, combatting deep fakes requires collaboration among various stakeholders, including technology companies, academia, and government agencies. Media outlets and journalists can facilitate such partnerships by serving as intermediaries and fostering dialogue between different sectors. By working together, these groups can develop strategies and solutions to mitigate the impact of deep fakes effectively.

Fact-checking entities and certain pro-democracy NGOs have taken steps to assist citizen groups and newsrooms in scrutinizing potentially deceptive viral election material. For instance, a human rights empowerment nonprofit named WITNESS initiated a Deepfakes rapid response pilot project last year, mobilizing a network of approximately 40 commercial and research specialists to scrutinize numerous suspicious clips.[48] Their findings revealed that they had greater success in identifying manipulated video content compared to audio.

CONCLUSION

The proliferation of deepfakes presents a multifaceted challenge to society, with implications spanning from the manipulation of police procedures to the dissemination of misinformation. Through a careful examination of various cases, we have witnessed the potential harm caused by these deceptive technologies, underscoring the urgent need for effective solutions.

Strategies such as implementing watermarks and enacting specific legislation have emerged as promising avenues to combat the damaging effects of deepfakes. By leveraging technological innovations and legal frameworks, we can strive to mitigate the risks posed by deepfakes and safeguard the integrity of our information ecosystems and justice systems alike. It is imperative that we continue to stay vigilant and proactive in our efforts to address this evolving threat, ensuring that our societies remain resilient in the face of technological advancements.

NOTES

1. Published on March 2023, available at: https://twitter.com/TheInfiniteDude/status/1637211568692932608
2. González Gómez Martín, Astor Maggie, "*Who's Running for President in 2024?*" The New York Times, March 2024, available at: https://www.nytimes.com/interactive/2023/us/politics/presidential-candidates-2024.html
3. Rogers Reece, "*Cosa ci insegnano i deepfake sull'arresto di Donald Trump*" The Wired, March 2023, available at: https://www.wired.it/article/donald-trump-immagini-arresto-false-deepfake/
4. Spring Marianna, *"Trump supporters target black voters with faked AI images"* BBC, March 2024, available at: https://www.bbc.com/news/world-us-canada-68440150
5. Sen Sumanti, "*AI-generated photos of Joe Biden in military gear surface online after Jordan drone strike kills US troops*" Hindustan Times, January 2024, available at: https://www.hindustantimes.com/world-news/us-news/aigenerated-photos-of-joe-biden-in-military-gear-surface-online-after-jordan-drone-strike-kills-us-troops-101706678009453.html
6. Crescenzi Chiara, "*Midjourney vieterà le immagini di Trump e Biden generate dall'intelligenza artificiale*" The Wired, February 2024, available at: https://www.wired.it/article/midjourney-immagini-trump-biden-deepfake/#intcid=_wired-it-bottom-recirc_d77d181a-30a9-41b9-ab46-d8f0a688a1ae_entity-topic-similarity-v2
7. "*El creador del 'deepfake' del Papa explica que se encontraba bajo los efectos de unos alucinógenos*" La Vanguardia, March 2023, available at: https://www.lavanguardia.com/tecnologia/20230328/8859266/creador-imagen-viral-papa-francisco-chaqueta-deep-fake-pmv.html
8. Saner Emine, *"Inside the Taylor Swift deepfake scandal: 'It's men telling a powerful woman to get back in her box'"* The Guardian, January 2024, available at: https://www.theguardian.com/technology/2024/jan/31/inside-the-taylor-swift-deepfake-scandal-its-men-telling-a-powerful-woman-to-get-back-in-her-box
9. The information about the deepfake case in New Hampshire was extracted from: Bond Shannon, "*AI fakes raise election risks as lawmakers and tech companies scramble to catch up*" NPR, February 2024, available at: https://www.npr.org/2024/02/08/1229641751/ai-deepfakes-election-risks-lawmakers-tech-companies-artificial-intelligence

10. The information about the deepfake case in Slovakia was extracted from: Meaker Morgan, "*Slovakia's Election Deepfakes Show AI Is a Danger to Democracy*" The Wired, October 2023, available at: https://www.wired.co.uk/article/slovakia-election-deepfakes
11. The audio can be found on Telegram. It is available at: https://t.me/harabin/10147
12. The information about the deepfake case in India was extracted from: Christopher Nilesh, "*An Indian politician says scandalous audio clips are AI deepfakes. We had them tested*", Rest of World, available at: https://restofworld.org/2023/indian-politician-leaked-audio-ai-deepfake/?utm_source=Rest+of+World+Newsletter&utm_campaign=c2481b9591-EMAIL_CAMPAIGN_2023_07_10_05_51&utm_medium=email&utm_term=0_-c2481b9591-%5BLIST_EMAIL_ID%5D
13. The concept of the "liar's dividend," was coined by jurists Robert Chesney and Danielle Keats Citron in a 2018 California Review article. Chesney Robert, Keats Citron Danielle, "*Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*" California Law Review, 2019, available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3213954
14. Mirza Rehan, "*How AI deepfakes threaten the 2024 elections*" The Journalist´s Resource, February 2024, available at: https://journalistsresource.org/home/how-ai-deepfakes-threaten-the-2024-elections/
15. The deepfake video is available at: https://www.facebook.com/permalink.php?story_fbid=pfbid0bFUh6GUqzQuUGxrtARkXVZftBMujouzYonzMzFLjPQNUFjVsiKWgMaMh7soyFbXbl&id=100069495133064
16. The information about the deepfake case in United Kingdom was extracted from: Kulkarni Ankita, "*Fact Check: This Video Of Rishi Sunak Promoting Musk's Project Is A Deepfake*" ABP, December 2023, available at: https://news.abplive.com/fact-check/deepfake-video-of-uk-pm-rishi-sunak-promoting-elon-musk-investment-project-1651247
17. Quinn Ben, "*Slew of deepfake video adverts of Sunak on Facebook raises alarm over AI risk to election*" The Guardian, January 2024, available at: https://www.theguardian.com/technology/2024/jan/12/deepfake-video-adverts-sunak-facebook-alarm-ai-risk-election
18. The information about the deepfake case in Turkey was extracted from: Ioannou Demetrios, "*Deepfakes, Cheapfakes, and Twitter Censorship Mar Turkey's Elections*" The Wired, May 2023,available at: https://www.wired.com/story/deepfakes-cheapfakes-and-twitter-censorship-mar-turkeys-elections/
19. The information about the deepfake case in South Korea was extracted from: "*Editorial: South Korea detects 129 cases of deepfake fraud before the end of election nominations*" The Chouson Daily, February 2024, available at: https://www.chosun.com/english/opinion-en/2024/02/20/ZCL4Z7T3MJFMFO7YLEWZ6VBVK4/
20. The information about the deepfake case in Indonesia was extracted from: Ware Gemma, "*Deepfakes and disinformation swirl ahead of Indonesian election – podcast*" The Conversation, February 2024, available at: https://theconversation.com/deepfakes-and-disinformation-swirl-ahead-of-indonesian-election-podcast-223119
21. "*Disinformation and elections in the age of artificial intelligence*" Queensland Government, February 2024, available at: https://www.forgov.qld.gov.au/information-and-communication-technology/queensland-government-digital-futures-and-foresight/signals/disinformation-and-elections-in-the-age-of-artificial-intelligence

22. "*Disinformation and elections in the age of artificial intelligence*" Queensland Government, February 2024, available at: https://www.forgov.qld.gov.au/information-and-communication-technology/queensland-government-digital-futures-and-foresight/signals/disinformation-and-elections-in-the-age-of-artificial-intelligence
23. Stern Jacob, "*AI in Politics Is So Much Bigger Than Deepfakes*" The Atlantic, January 2024, available at: https://www.theatlantic.com/technology/archive/2024/01/ai-elections-deepfakes-biden-robocall/677308/
24. Shukla Vandinika, "*Deepfakes and Elections: The Risk to Women's Political Participation*" Tech Policy Press, February 2024, available at: https://www.techpolicy.press/deepfakes-and-elections-the-risk-to-womens-political-participation/
25. The paper can be found: Santurkar et. al, "*Whose Opinions Do Language Models Reflect?"* March 2023, arXiv:2303.17548, available at: https://arxiv.org/pdf/2303.17548.pdf. The results of the research was extracted from: Myers Andrew, "*Assessing Political Bias in Language Models*" Stanford University Human-Centered Artificial Intelligence, March 2023, available at: https://hai.stanford.edu/news/assessing-political-bias-language-models
26. Van Niekerk et. al, "*Challenging systematic prejudices: an investigation into bias against women and girls in large language models*" UNESCO, 2024, available at: https://unesdoc.unesco.org/ark:/48223/pf0000388971
27. González Zepeda Luz Elena, Martínez Pinto Cristina Elena, "*Inteligencia artificial centrada en los pueblos indígenas: perspectivas desde América Latina y el Caribe*" UNESCO 2023, available at: https://unesdoc.unesco.org/ark:/48223/pf0000387814. For a review of the report: Le Fevre Cervini Enzo Maria "*Shaping the Future: Indigenous Voices Reshaping Artificial Intelligence in Latin America*" Linkedin, November 2023, available at: https://www.linkedin.com/pulse/shaping-future-indigenous-voices-reshaping-artificial-enzo-maria-dyuie/?trackingId=YmeewKCdTsmn%2F3a9VH8WhQ%3D%3D
28. Motoki Fabio, Pinho Neto Valdemar, Rodrigues Victor, "*More human than human: measuring ChatGPT political bias*" Public Choice, 2023, available at: https://link.springer.com/article/10.1007/s11127-023-01097-2
29. Mirza Rehan, "*How AI deepfakes threaten the 2024 elections*" The Journalist´s Resource, February 2024, available at: https://journalistsresource.org/home/how-ai-deepfakes-threaten-the-2024-elections/
30. The Information about the social networks policies was extracted from: Mirza Rehan, "*How AI deepfakes threaten the 2024 elections*" The Journalist´s Resource, February 2024, available at: https://journalistsresource.org/home/how-ai-deepfakes-threaten-the-2024-elections/
31. Weiner Daniel, Norden Lawrence, "*Regulating AI Deepfakes and Synthetic Media in the Political Arena*" Brennan Center for Justice, available at: https://www.brennancenter.org/our-work/research-reports/regulating-ai-deepfakes-and-synthetic-media-political-arena
32. All state laws are available at: https://www.citizen.org/article/tracker-legislation-on-deepfakes-in-elections/
33. Edelman Adam, "*States turn their attention to regulating AI and deepfakes as 2024 kicks off*" NBC News, https://www.nbcnews.com/politics/states-turn-attention-regulating-ai-deepfakes-2024-rcna135122
34. Edelman Adam, "*States turn their attention to regulating AI and deepfakes as 2024 kicks off*" NBC News, https://www.nbcnews.com/politics/states-turn-attention-regulating-ai-deepfakes-2024-rcna135122

35. "*Minnesota Law Regulating Election-Related Deep Fakes"* Cook County Minnesota, available at: https://co.cook.mn.us/news_detail_T6_R410.php
36. "*Brazilian Electoral Court Regulates Artificial Intelligence in Elections and Prohibits Deepfake Use by Campaigns*" UOL, February 2024, available at: https://www1.folha.uol.com.br/internacional/en/brazil/2024/02/brazilian-electoral-court-regulates-artificial-intelligence-in-elections-and-prohibits-deepfake-use-by-campaigns.shtml
37. "*Brazil seeks to curb AI deepfakes as key elections loom"* RTL Today, March 2024, available at: https://today.rtl.lu/news/business-and-tech/a/2175777.html
38. More information available at: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai
39. All provisions mentioned in this paragraph are found in article 52 of the European AI Act.
40. Ryan-Mosley, "*EE UU vs Europa: Biden toma la delantera en la carrera por regular la IA*" MIT Technology Review, October 2023, available at: https://www.technologyreview.es/s/15896/ee-uu-vs-europa-biden-toma-la-delantera-en-la-carrera-por-regular-la-ia
41. Williams Rhiannon "*AI-text detection tools are really easy to fool*" MIT Technology Review, July 2023, available at: https://www.technologyreview.com/2023/07/07/1075982/ai-text-detection-tools-are-really-easy-to-fool/
42. Heikkilä Melissa, "*Google lanza una marca de agua para detectar imágenes generadas por IA"* MIT Technology Review, August 2023, available at: https://www.technologyreview.es/s/15649/google-lanza-una-marca-de-agua-para-detectar-imagenes-generadas-por-ia
43. The information of this tool was extracted from: Heikkilä Melissa, "*Google lanza una marca de agua para detectar imágenes generadas por IA"* MIT Technology Review, August 2023, available at: https://www.technologyreview.es/s/15649/google-lanza-una-marca-de-agua-para-detectar-imagenes-generadas-por-ia
44. Kirchner Jan Hendrik, Ahmad Lama, Aaronson Scott, Leike Jan, "*New AI classifier for indicating AI-written text*" OpenAI Blog, January 2023, available at: https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text
45. Heikkilä Melissa, "*Why detecting AI-generated text is so difficult (and what to do about it)"* MIT Technology Review, February 2023, available at: https://www.technologyreview.com/2023/02/07/1067928/why-detecting-ai-generated-text-is-so-difficult-and-what-to-do-about-it/
46. A list of these techniques can be found in: Rowan Philip, "*How to Identify and Investigate AI Audio Deepfakes, a Major 2024 Election Threat*" Global Investigative Journalism Network, February 2024, available at: https://gijn.org/resource/tipsheet-investigating-ai-audio-deepfakes/
47. The tool "AI or not" is available at: https://www.aiornot.com
48. Rowan Philip, "*How to Identify and Investigate AI Audio Deepfakes, a Major 2024 Election Threat*" Global Investigative Journalism Network, February 2024, available at: https://gijn.org/resource/tipsheet-investigating-ai-audio-deepfakes/