How to Cite:
Ghediri, K. (2024). Countering the negative impacts of deepfake technology: Approaches for effective combat. *International Journal of Economic Perspectives*, *18*(12), 2871–2890. Retrieved from https://ijeponline.org/index.php/journal/article/view/818

Countering the negative impacts of deepfake technology: Approaches for effective combat

Dr. Karima Ghediri
Lecturer at ENSJSI, Algeria

***Abstract*---**The rapid rise of deepfake technology poses significant challenges at individual, societal, and national levels. Although there are positive applications, malicious uses have largely overshadowed them, making it crucial to examine available methods for addressing this growing threat. This paper reviews three primary approaches to mitigate the risks of deepfake technology: technical detection methods, legal and regulatory frameworks, and media literacy initiatives. While no single solution fully addresses the challenges of deepfake misuse, a comprehensive understanding of these combined strategies can provide a strong foundation for reducing the harmful impacts of this technology.

***Keywords*---**Deepfake, Fake Videos, Visual Manipulation, Deepfake Detection, Media Literacy, Legal Responsibility.

I. Introduction:

Artificial intelligence is one of the most prominent technological advancements of recent decades. Among its notable emerging applications is deepfake technology, which leverages deep learning to simulate and generate highly realistic visual and audio content, making it increasingly challenging to distinguish between fabricated and authentic videos.

As the utilization of this technology expands, it has become essential to assess its impact across various social, cultural, political, and legal domains. Addressing the complexities of deepfake technology requires a comprehensive approach that considers its positive potential and associated risks. This involves understanding the evolution of deepfake technology, its diverse applications, and the challenges and dangers it entails, along with potential measures to mitigate its negative effects. Therfore examining the development of this technology, the accompanying challenges and the effective measures to reduce its harmful impact have become imperative.

© 2024 by The Author(s). ISSN: 1307-1637 International journal of economic perspectives is licensed under a Creative Commons Attribution 4.0 International License.
***Corresponding author**: Ghediri, K., Email: Ghediri.karima@ensjsi.dz*
Submitted: 09 October 2024, Revised: 18 November 2024, Accepted: 24 December 2024

II. The emergence of deepfake technology and its implications

- **The Origins and Rise of Deepfake Technology**

While deepfake technology represents a recent advancement in digital manipulation, the roots of image and video manipulation extend far back in media history. The first known instance of photo manipulation dates to 1860 when an image of the southern politician John Calhoun was skillfully altered by replacing his face with that of President Abraham Lincoln to make Lincoln appear more authoritative (Masood & Nawaz, M, 2021).This form of manipulation was achieved through rudimentary techniques that relied on manual photo editing. Over time, image manipulation methods evolved, particularly with the advent of software such as Photoshop and After Effects, which enabled digital editing of photos and videos. With the rise of artificial intelligence, this capability to manipulate took on a new dimension, allowing fake content to appear more realistic than ever before.

The term "deepfake" first appeared in 2017 when a user posted digitally altered adult videos on Reddit forums, where they digitally manipulated faces, replacing actors' faces with those of celebrities. (Kietzmann & W. Lee, 2019)The content spread widely, leading to an explosion of fake content initially targeting actors and singers before expanding to include politicians and others. This spread accelerated as easy-to-use tools were developed, allowing individuals, even those without technical expertise, to produce fake videos. As a result, deepfake technology transitioned from a rare innovation to a widespread phenomenon, introducing a range of ethical and practical concerns.

However, what sets deepfake technology apart from traditional manipulation lies in several fundamental aspects. Traditional manipulation tools demande advanced technical skills and many hours of manual labor, whereas deepfake relies on AI algorithms and deep learning, making the process faster and more efficient. Furthermore, deepfake technology is capable of producing highly realistic visual and audio content, making it increasingly difficult for even experts to distinguish between real and fake material. This critical difference underscores the heightened risks associated with deepfake technology compared to earlier techniques.

- **Deep-Learning and Realism Enhancement**

Deepfake technology leverages artificial intelligence to analyze images and videos using deep learning algorithms. These algorithms process extensive sets of visual and audio data (AlokeEjike & AbahJos, 2023) building models that can generate new content by mimicking learned patterns. This technology represents an unprecedented tool for creating fake content, as it enables the substitution of faces and voices, making the resulting content appear entirely authentic. (Bahar Uddin Mahmud, et al, 2019)

The more data fed into the system, the more accurate and realistic the fake content becomes. This self-learning capability allows deepfake technology to

evolve (Nguyen, Thanh Thi,et al, 2020) making it increasingly challenging to detect deepfakes.

Face-swapping in videos is one of the most common applications of deepfake, making the targeted individual appear to say or do things they never actually did (Bahar Uddin Mahmud, et al, 2019) Furthermore, deepfake technology can mimic voices, adding another layer of complexity to the process of detecting deception. As a result, deepfake technology has been widely applied to impersonate public figures and politicians, creating vast opportunities for information manipulation.

The Internet and Social Media: Key Drivers of Deepfake Technology's Spread

The rapid proliferation of fake visual content has been driven by advancements in digital processing systems and the continuous increase in processing speeds. Furthermore, the internet and social media have played a significant role in the widespread adoption of deepfake technology, making it easier, faster, and more affordable to produce videos using this method. This accessibility has been facilitated by numerous user-friendly software programs and services, both paid and free, that enable users to create realistic fake videos with tools such as "Faceswap" and "DeepFaceLab" (Arash Heidari, et al, 2022) accelerating the technology's global spread. Additionally, the open structure and high speed of the internet have facilitated the rapid dissemination of fake videos across various platforms, including Facebook, Twitter (X), and TikTok. These videos spread with remarkable speed, making it challenging to control their distribution or mitigate their impact.

- **Implications of Deepfake Technology**

It is essential to recognize that deepfake technology has numerous positive applications and valuable contributions, significantly enriching fields such as art, culture, media, and beyond. The technology has had a notable impact on the internet, particularly through the creation of satirical videos, comedic content, and humorous memes, which have gained wide popularity among users. Additionally, deepfake has been creatively applied in e-commerce, illustrating that this technology is not limited solely to negative applications, as is often assumed; rather, it encompasses many beneficial uses that add value to the digital community. For example, this technology was employed to recreate the character of Paul Walker after his passing during the filming of the Fast and Furious series and to synthesize the voice of John F. Kennedy, delivering his final, unspoken speech on the day he was assassinated. Furthermore, deepfake technology allows for synchronized, creatively dubbed videos, offering significant potential for use in cinematic and artistic productions, underscoring its positive and constructive applications.

De Reuter has noted that while deepfake technology may be ethically questionable, it is not inherently unethical. There are three primary factors that define unethical uses of deepfake: representing individuals without their consent, intentionally deceiving viewers, and harboring malicious intent. Based on these factors, ethically acceptable uses of deepfake technology should not be entirely dismissed (Eberl, 2022).

Conversely, deepfake technology has fundamentally altered how individuals interact with images and videos, directly impacting the trustworthiness of visual content shared online. Previously, images and videos were regarded as reliable means of documenting events; however, deepfake technology has undermined this trust. On a social level, deepfake can serve as a powerful tool for disinformation and manipulation of public opinion, where it may be employed to create fake statements by political or public figures with the aim of spreading lies or inciting violence. Globally, numerous cases have arisen where deepfake technology has been used to sow discord between groups or to disseminate misleading information. For instance, during the 2020 election campaign, conflicting deepfake content was circulated regarding Presidents Donald Trump and Joe Biden, which was believed to contribute to an atmosphere of fear and distrust (Masood & Nawaz, M, 2021).

- **Risks and Threats Posed by Deepfake Technology**

Deepfake technology presents an escalating threat due to its potential for malicious use, which can inflict severe harm on individuals, communities, and even nations. The primary risks associated with deepfake technology can be summarized as follows:

- **The Spread of Misinformation**: One of the most significant dangers of deepfake technology is its capacity to generate false information, which can lead to various crises and foster the spread of hatred.
- **Privacy Violations and Blackmail**: Deepfake technology can be used to create fake videos of individuals in compromising or inappropriate situations, aimed at defamation or blackmail. Such actions severely damage individuals' reputations. For instance, in 2019, an application called DeepNude emerged, producing videos of undressed women through digital manipulation (Taeb & Chi, 2022).
- **Security Threats**: Deepfake technology poses a national security threat, as fake videos can be created to depict fabricated statements by political leaders or public figures, thereby creating confusion (AlokeEjike & AbahJos, 2023). Intelligence agencies may also use deepfake to create content aimed at influencing policies or security decisions, potentially impacting international relations or escalating tensions between governments.
- **Legal Implications**: Deepfake technology presents a substantial legal challenge. Visual evidence, once considered indisputable in investigations and legal cases, has now become unreliable, as images and videos can be manipulated through artificial intelligence, making them difficult to trust as conclusive evidence in investigations and trials.

The advent of deepfake technology and the widespread availability of artificial intelligence tools have intensified these threats, with the most significant impact being on public trust. As people increasingly question the authenticity of visual content, this erosion of trust may result in problematic social norms and behaviors. This environment enables wrongdoers to dismiss allegations against them, arguing that nothing can be definitively proven. Consequently, it may become easier to undermine evidence, reject visual data, and dismiss it as “just another deepfake.”

The challenge of deepfake technology lies not only in disproving false information but also in establishing the authenticity of content. This issue is particularly pressing for news media organizations, which must assess the credibility of videos circulating in an environment where tracking the origin, creator, and distribution of content has become increasingly complex. With these negative impacts, it has become essential and urgent to adopt effective measures to counter the risks posed by deepfake technology, especially as the rapid spread of this phenomenon continues. The majority of literature reviews indicate that addressing these challenges requires a comprehensive approach, focusing on three primary strategies:

1 Technical Approach for Deepfake Detection
2 Legal and Regulatory Approach
3 Media Literacy and Training Approach

III. Approaches to counter the negative impact of deepfake

1 Technical Approach for Deepfake Detection

Detection represents a crucial tool for controlling the proliferation of fake images and videos and mitigating associated risks, particularly given the mounting challenges deepfake technology presents to digital content credibility. Researchers and leading companies are actively developing AI-based mechanisms for identifying such manipulated content.

As generated content continues to evolve in sophistication and realism, the ability to keep pace with these advancements becomes increasingly challenging. Once effective detection methods may no longer prove viable against the latest iterations of deepfake content. For example, in June 2018, computer scientists at the University of Albany, announced a program to detect deepfakes by examining abnormal blinking patterns, based on the observation that deepfake subjects blink significantly less than individuals in authentic videos. This discrepancy is largely due to the algorithm's reliance on publicly available facial images, which typically depict individuals with open eyes, with few instances of closed-eye images. Consequently, deepfake algorithms struggle to create faces with natural blinking patterns. Thus, lower blink rates in deepfake videos can aid in distinguishing real from fake content (Nguyen, Thanh Thi,et al, 2020) .However, once this detection method was adopted, new deepfake generation tools emerged, designed to circumvent this limitation by programming blinks to appear at natural rates, complicating the task for detection specialists. Consequently, experts remain divided on whether digital media forensic analysis alone is sufficient for detecting deepfakes. There is ongoing concern that bad actors can simply update their tools each time a new detection method is released. Therefore, effective efforts to counter misinformation demand continuous enhancement of detection capabilities and the ongoing development of new techniques. Addressing deepfake challenges further necessitates collaboration among diverse stakeholders, including researchers, tech companies, governments, and other invested parties, to drive the successful advancement and implementation of effective detection technologies.

- **Collaborative Efforts and Specialized Datasets in Deepfake Detection**

Several multinational corporations and universities have developed specialized datasets to implement advanced deep-learning systems aimed at identifying fake videos and images (Gupta, 2024). This progress is intended to strengthen deepfake detection capabilities and curb the proliferation of such content.

National security agencies primarily fund the development of these detection tools, with the National Institute of Standards and Technology (NIST) launching the Open Media Forensic Challenge (OpenMFC) in 2017. This initiative allows researchers to tackle media forensic challenges and evaluate available algorithm performance. Similarly, the Defense Advanced Research Projects Agency (DARPA) launched the Media Forensics (MediFor) program in 2018, designed to create tools that detect manipulated images and combat the spread of misinformation.

In 2020, Facebook held a deepfake detection competition, offering a prize of up to one million dollars and a dataset containing 100,000 manipulated images, aiming to stimulate the development of more accurate deepfake detection techniques (Lin, 2024) Additionally, private cybersecurity firms are actively developing advanced solutions to identify deepfakes, build secure platforms for blocking unauthorized bots, and prevent fraud and digital pollution. Despite these efforts, detection technology must continuously evolve to keep pace with improvements in deepfake creation tools (Kietzmann, Jan, et al, 2019).

Various tools have been developed for detecting deepfakes, each with distinct features and specific applications. These tools assist in identifying image and video manipulation by analyzing visual anomalies through traditional methods, combining them with metadata analysis, or applying artificial intelligence techniques. The effectiveness and accessibility of these tools vary; platforms like FaceForensics++ and Sensity AI have shown notable success. One AI-based tool, for instance, was trained on a dataset of half a million altered images, achieving detection results superior to human forensic teams (CDEI, 2019). Other tools, such as InVID and WeVerify, are widely used by journalists and fact-checkers worldwide as free, multilingual verification add-ons suitable for newsrooms (Nygren, Thomas, et al, 2022).

Although numerous modern detection methods are available, but detection accuracy remains highly variable depending on the specific case. For example, many recent content-creating models, such as DeepFaceLab, remain challenging for most detection models to identify. Although promising generalized detection methods have recently emerged, no single detection model has been able to effectively identify all types of deepfake content. Therefore, current systems still rely on a suite of specialized models, each trained on a particular content-creation framework (Kalina Bontcheva, et al, 2024).

To date, numerous methods have been developed to address the challenges posed by deepfake technology, with most of these approaches relying on deep learning—the very technology used to generate fake content. This has created an ongoing conflict between malicious applications that exploit deep learning capabilities to

produce misleading content and positive applications that utilize this technology to detect forgeries and enhance the reliability of digital content. These efforts focus on two primary aspects: (1) tools for deepfake detection and (2) tools for content authentication and prevention of falsification.

A. Deepfake Detection Tools

Deepfake videos usually can't fool users for long, as their accuracy depends heavily on the creator's skill level. For example, face-swapping in videos often leaves small flaws, like minor quality inconsistencies. The more refined the fake, the harder it is to detect. This level of quality in visual content largely determines the best digital forensic tools to use for detection. However, as it becomes harder to tell real videos from fake ones, detection methods may be less effective, especially for images outside of known datasets.

To achieve stronger results in spotting deepfakes, many platforms use a mix of detection methods. Each tool plays a role in confirming whether content is real or altered, making these platforms valuable for fact-checking and online security. Here are some of the main tools these platforms rely on to fight the spread of deepfakes:

1) Visual Analysis Tools

Image analysis shares certain methodologies with deepfake video analysis; however, the complexity of tasks varies greatly, requiring distinct methods when detecting manipulated videos. Video analysis involves all the steps used in image detection, plus additional processes, such as converting the video into individual frames before running it through detection systems (Kaur, A., Noori Hoshyar, A., Saikrishna, V. et al., 2024).

Researchers have proposed innovative methods to identify fake content by analyzing different characteristics in images, audio, and video. This analysis covers spatial and frequency information in images, as well as temporal and frequency information in audio and video. The idea is that these methods can reveal manipulations not visible to the naked eye (Gupta, 2024).

- **Spatial Analysis**

Spatial analysis is dedicated to understanding and interpreting the distribution of visual features within a single image or frame in a video. This method is mainly applied to detect visual manipulations by analyzing specific details, patterns, and visual components at a particular point in time. It focuses on fixed characteristics within a single frame, such as edges, patterns, and color distribution (Taeb & Chi, 2022)

Key indicators examined in the spatial analysis include inconsistencies in lighting, reflections, and shadows; blurred edges; distorted facial features; unnatural eye direction; and missing facial details like known facial marks or moles. Other tell-tale signs are unnaturally smooth skin, missing details in hair or teeth, facial asymmetry, and inconsistencies in pixel quality. Additionally,

spatial analysis can detect discrepancies in the background, such as mismatched room dimensions or anomalies in other visible elements, like people or objects that should logically align with the scene (CDEI, 2019)

- **Temporal Analysis**

Scientists have detected fake videos through spatial inconsistencies within altered footage, achieving notable results but initially overlooking the continuity of time within video sequences. To address this gap, researchers began using temporal cues to manage issues of discontinuity in deepfake videos, though early methods did not fully capture spatial anomalies. As a result, scientists integrated both spatial and temporal distortions, which ultimately improved the identification of fake videos that were not part of existing datasets (Chin-Yuan Lin,et al, 2024).

In temporal analysis, the focus lies in tracking data changes over time in visual media, especially in videos, by examining and studying transitions in the visible scene. This approach is widely applied to analyze movement and detect manipulations that may be invisible when viewing individual frames alone.

The core idea is to search for discrepancies between "visual markers" or mouth formations and "phonetics." For example, certain sounds like "B," "M," or "P" are challenging to produce without fully closing the lips, allowing researchers to spot inconsistencies. Additional indicators include various artifacts such as facial wobbling, abnormal shine, distortions, unnatural movements of typically static objects like microphone stands, and peculiar behaviors indicating an unnatural or unlikely action (J. Nightingale & A. Wade, 2022; Westerlund, 2019).

- **Frequency Analysis**

Frequency analysis is utilized to examine and interpret the distribution of various frequencies within visual and audio data. This method is widely applied across technical fields related to image, video, and audio processing. In frequency analysis, mathematical transformations like the Fourier Transform are used to convert data from the spatial domain to the frequency domain, creating a frequency map that illustrates how different frequencies are distributed across an image or audio file.

While traditionally employed to enhance the quality of images, videos, and audio signals, frequency analysis has become crucial in deepfake detection, particularly for identifying subtle distortions or noise artifacts that might not be visible in the spatial domain but are detectable in the frequency domain.

However, frequency analysis alone is insufficient for comprehensive detection of deepfakes. For effective results, it must be integrated with other methods, such as spatial analysis, deep learning algorithms, and metadata examination. This combination of techniques provides a more thorough approach to uncovering deepfake content, addressing the unique challenges posed by sophisticated manipulation techniques.

2) Metadata Analysis

Metadata analysis focuses on examining file-associated information, such as creation time, geographic location, and camera type. In cases of manipulation, metadata may reveal inconsistencies when compared with spatial and temporal analyses. It is important to note that there are no exclusive metadata tools specifically designed for detecting deepfakes. Instead, general metadata analysis tools like ExifTool and Amped Authenticate are used as part of a broader manipulation detection process, including deepfake identification. These tools are typically employed in conjunction with visual content analysis tools to enhance detection accuracy, providing a more comprehensive approach to uncovering potential alterations.

3) Deep Learning Tools

Experimental findings indicate that deep learning techniques are highly effective in detecting deepfakes and often surpass traditional methods that don't utilize deep learning (Rana; Nobi, et al, 2022),These tools capitalize on deep learning's capacity to process vast datasets, enabling them to detect intricate patterns within large volumes of data using neural networks. Through training on both real and manipulated images and videos, these models learn to recognize subtle inconsistencies, such as facial texture, lighting, lip movements, and other minor discrepancies that are characteristic of deepfakes. This training allows the models to distinguish between authentic and altered content with high accuracy (Preeti & Bansal, 2024)

However, as deepfake quality improves, current deep learning methods also require continuous advancements to keep pace and effectively detect newer, more sophisticated fakes (Almars, 2021).

B. Content Authentication and Anti-Forgery Tools

While previously mentioned tools focus on detecting signs of manipulation in visual content after it has been created, there is a growing research focus on prevention—specifically, on techniques designed to preempt the fabrication of images and videos. This preventative approach leverages innovative technologies, such as artificial intelligence, digital analysis, and cybersecurity methods, to restrict the spread of manipulated content. By securing original content through various means, these tools allow for later comparisons with any suspect visual materials, making it easier to verify authenticity. This method plays a significant role in protecting content from tampering. Key tools used in this preventive framework include:

1) Watermarking

Watermarking is a technique used to authenticate images and video content by embedding encrypted information within the file at the time of creation. This information, which can be visible or hidden within the content, serves as a means of establishing ownership and verifying the source rather than examining specific aspects of the video, such as the location or creator. While watermarking is valuable for proving intellectual property rights, it is less effective for long-term

content verification or preventing deepfake manipulation, as advanced tools can remove or bypass watermarks. However, it still presents a barrier for unauthorized reuse, making it more challenging for some to directly repurpose content without permission.

2) Blockchain Technology

The term "blockchain" refers to a sequence of records, each encrypted and linked to the previous block, that securely, transparently, and traceably stores transactions. Originally developed to secure cryptocurrency transactions such as Bitcoin and Ethereum, blockchain technology offers a high degree of security, preventing any unauthorized alterations or tampering with data saved within the blockchain (sultan , ruhi, & lakhami, 2018),

The decentralized and sequential nature of this system ensures that each record, or "block," is connected to the others. This makes altering a single block without a consensus from all users nearly impossible, enhancing the reliability and immutability of the data stored within. Blockchain is now being used as a reliable means to verify the origin and history of specific products, as well as to record the creation and ownership of artistic works, allowing for a permanent record of a digital asset's authenticity. In the future, this technology may be applied more widely for establishing the authenticity of various digital assets, whether written documents, images, videos, or artwork (Floridi, 2018). When users encounter questionable digital content, blockchain could serve as a reference to verify its origin. Although such a solution is currently not widespread, researchers have proposed using blockchain and smart contracts to assist users in identifying fake videos by tracking each digital file's recorded fingerprint. Once a digital fingerprint is stored on the blockchain, it can be referenced to confirm whether the current version of the content matches the original.

One notable example is Truepic, a U.S.-based startup that uses blockchain to sell software that adds a digital (invisible) fingerprint to images and videos upon their creation. This information, which includes GPS data, is encrypted and stored as a file on Truepic's servers (Chesney & Citron, 2019). This allows authenticity inspectors to access the original version and verify it against subsequent copies. Though promising, experts Danielle Citron and Robert Chesney have noted that the effectiveness of such technology depends on its widespread adoption on content-capturing devices, including smartphones and laptops (CDEI, 2019)

Another similar tool aims to provide real-time content verification, such as Amber Authenticate. This technology works on devices that produce original content, like photographs, audio, and video, by generating encrypted, immutable metadata that logs authenticity details such as time and location at the point of creation (Nguyen, Thanh Thi,et al, 2020). This metadata is stored securely, becoming unalterable and enabling traceability of the content's origin during its distribution. By detecting any tampering attempts, these digital fingerprints help ensure the authenticity of original digital content and protect it from unauthorized modifications (Kietzmann & W. Lee, 2019).

As face-swapping technology continues to advance, more challenging and highly convincing deepfake videos are expected, making detection increasingly difficult

with current algorithms. Addressing this challenge calls for developing new detection methods and comprehensive databases to keep pace with deepfake technology. The competition between deepfake methods and detection algorithms is likely to become a technological arms race, one that may require additional non-technical solutions, such as regulatory and legal measures, to prevent harmful misuse (Korshunov & Marcel, 2019).

2. Legal and Regulatory approach

Legislation and regulation play a central role in addressing the challenges posed by AI technologies, specifically deepfake technology. While deepfake technology offers creative potential and contributes to innovative developments, it also raises significant concerns related to privacy, security, and credibility. Historically, videos and photographs have been extensively used as evidence in police investigations and court cases, where they are analyzed by digital forensic experts with backgrounds in computer science and digital information analysis (Nguyen, Thanh Thi,et al, 2020). However, with the growing technical challenges in verifying the credibility of AI-generated evidence, new issues arise in handling such forensic material.

Legal scholars emphasize the gap between how digital evidence is currently managed and the rapid advancements in technology, highlighting an urgent need for reform in how digital evidence is authenticated in court. This shift may require moving beyond traditional human testimonies to incorporate technological solutions and specialized analyses, though this introduces its own complexities, including increased costs (Schlegel, 2024). Consequently, there is a growing demand for legal frameworks to govern the use of deepfake technology, protecting individuals and society from its misuse while balancing technological innovation with public and private rights.
Given the novelty and rapid evolution of deepfake technology, lawmakers face challenges in keeping up with advancements and ensuring the effectiveness of legal provisions. The key questions remain: How can laws be crafted to prevent harmful uses without stifling creative freedom? And how can these regulations be effectively implemented in real-world scenarios?

- **Legal Efforts to Counter the Misuse of Deepfake Technology**

Currently, most countries' civil and criminal laws do not specifically address deepfake technology. However, some states in the U.S. have managed to pass legislation targeting deepfake misuse. For example, Virginia's law against revenge pornography includes "falsely created" images and videos, making their distribution a misdemeanor and thus broadening the law to encompass deepfake content (Chesney & Citron, 2019). California has also enacted legislation prohibiting deepfake creation with the intent to harm a candidate's reputation or deceive voters into supporting or opposing a candidate. Similarly, a Texas law classifies creating or distributing a fake video as a misdemeanor if it is intended to influence elections. Other states, like Massachusetts, have pending legislation that seeks to criminalize the use of this technology for various harmful purposes (O'Donnell, 2021).

Several other countries have subsequently introduced laws criminalizing the use of deepfake technology without the consent of those depicted. South Korea, for instance, has banned such content within 90 days of parliamentary elections due to concerns about its impact. In the United Kingdom, the 2023 Online Safety Bill addresses fake content on the internet but does not explicitly mention deepfake, instead using broader terms such as automated tools (center for new technology and innovation, 2024).

- **Legislative Adaptation for Regulating Deepfake Technology**

Legal experts have suggested adapting existing laws to address issues like defamation, identity fraud, or impersonati (center for new technology and innovation, 2024)on involving deepfake technology. This approach would enable legislators to address legal loopholes by imposing criminal and civil liabilities on those who create or distribute harmful deepfake content targeting individuals. The misuse of deepfake technology could be considered a cybercrime, falling under existing laws on digital fraud or cyber-related offenses, thus holding those who distribute it with malicious intent accountable.

Most legislation addressing these issues does not directly specify "deepfake" as a term rather, it generally refers to "information technology tools or software," indicating a lack of precise terminology for deepfake-specific technologies. Unlike traditional editing tools, deepfake relies on AI-driven generative techniques that are fundamentally different from simpler, shallow-fake tools. Most legislative texts also recognize harm to personal privacy as grounds for legal action, regardless of the specific tools used. This suggests an urgent need for laws that account for the specific characteristics of deepfake technology and mitigate its consequences, addressing it under categories such as:

- **Privacy Violations**: Individuals have a right to privacy in both physical and digital realms. Laws worldwide strive to protect this right, although it remains challenging due to the technological capabilities for tracking, data mining, and using personal images as primary materials for creating deepfake videos and voices. When deepfake technology exploits someone's likeness without consent, privacy and data protection laws can apply.
- **Defamation:** Deepfake content can be categorized as defamatory when it spreads false claims, allegations, or rumors to harm the victim's reputation. Since the damage involves distributing fabricated content, it may be treated as digital defamation, which requires tailored legal approaches due to the specific technical complexities of deepfake. For instance, the difficulty of identifying perpetrators and determining a clear standard for defamation poses significant challenges.
- **Impersonation or Identity Theft**: Deepfake technology can be used to impersonate victims for illegal activities, including accessing sensitive information or committing financial fraud, which has already occurred in practice. These actions can be prosecuted under identity theft laws, especially when used for fraudulent purposes.
- **Intellectual Property:** Intellectual property includes creations of the human mind across various fields, such as art, literature, science, and industry, deserving of legal protection. In cases involving deepfake,

perpetrators may use copyrighted materials, such as images or videos, without permission, violating intellectual property rights.

- **Post-Legal Adaptation Challenges**

A significant challenge in tackling these violations is the difficulty of proving allegations against those responsible. Attributing the creation of deepfake content to a specific individual or entity is particularly challenging. Given the rapid technological advancements, this is far from straightforward, as creators can obscure their locations using anonymization tools like Tor. This often makes it virtually impossible to pinpoint their location, especially if the content has circulated widely across different platforms. In systems relying on civil liability, the burden of identifying the creators' identities and whereabouts often falls on the victim—a task that is at best difficult and, at worst, unattainable (O'Donnell, 2021).Furthermore, even if perpetrators are identified, they may reside outside the court's jurisdiction, as is often the case with foreign individuals or governments (Chesney & Citron, 2019).

- **Legal Responsibility of Social Media Platforms**

The legal option faces another problem: the autonomy of social media platforms in making decisions about the content posted on their sites. In this situation, would legal actions be taken against the social media platform hosting the content or against the content creators themselves? (Ramluckan, 2024). Since social media companies currently enjoy broad immunity for user-generated content under U.S. law—specifically Section 230 of the Communications Decency Act of 1996—they cannot be sued for such content. Given the extent of this immunity, some argue that the law should be amended to hold companies accountable for harmful and fraudulent information distributed through their platforms unless they make reasonable efforts to detect and remove it. They suggest that companies protected under this provision should lose their immunity if they knowingly or intentionally allow users to post illegal content (O'Donnell, 2021).

Therefore, the legal solution might involve incentivizing social media platforms to put more effort into identifying and removing fake content by introducing new laws to prevent deepfakes, along with necessary enforcement mechanisms. In other countries, the approach of imposing legal responsibility on platforms has been adopted. For instance, in 2017, Germany enacted a law imposing strict fines on social media companies that failed to remove racist or threatening content within 24 hours of being reported (Chesney & Citron, 2019). Thus, one legislative option could be to roll back the legal immunity granted to social media companies regarding content posted by their users, making platforms more accountable for published materials—not just the users.

However, legislation has a limited impact on malicious actors like foreign states and terrorists, who might launch massive disinformation campaigns against other countries on social media platforms or other websites.

Some researchers argue that media platforms undoubtedly bear responsibility for reporting false content and should even reconsider economic incentives for

publishers. Social media companies need to enforce ethical standards and move away from the reality where divisive content, pushed to the forefront, becomes a financial gain by increasing user engagement time for advertisements. Currently, many companies do not remove disputed content but, at best, reduce its visibility to make it harder to find, keeping it less prominent in users' news feeds. Thus, social media platforms should be encouraged to increase efforts in identifying and removing fake or fraudulent content.

Generally, only a few social media companies have policies regarding deepfake content, yet they should collaborate to prevent the misuse of their platforms for spreading misinformation by proactively implementing transparent, shared policies to ban and remove deepfake content. Should they be held accountable for the content on their sites, including deepfakes, or are they inadvertently supporting deepfake technology? As informed distributors, should they be held liable? (Kietzmann, Jan, et al, 2019).

Some companies have taken stronger measures, such as suspending user accounts and investing in faster detection technology. Both Twitter (now X) and Pornhub have banned sharing deepfake pornography and other explicit content (Rana; Nobi, et al., 2022). Regarding the current policy on deepfake pornography, some platforms now label manipulated or synthetic media, cautioning users before sharing the content. However, platforms don't automatically remove it; for instance, Facebook pledged to ban certain deepfake videos, but its efforts have been lukewarm in effectively curbing harmful deepfakes.

Even in cases where platforms have direct knowledge of illegal content, they often fail to act swiftly. For example, in response to a manipulated 2018 video portraying House Speaker Nancy Pelosi as intoxicated (a shallow-fake rather than a deepfake), platforms reacted differently: YouTube removed it, Twitter (X) took no action, and Facebook merely guided users to reports labeling it as fake, stating, "We don't have a policy that the information you post on Facebook must be true" (Kietzmann, Jan, et al, 2019). According to Westerlund, Facebook prevents any content flagged as false or misleading by third-party fact-checkers from generating ad revenue. The platform collaborates as well with over 50 organizations, including academics, experts, and policymakers, to develop new solutions (Shukla & Lyons, 2017).

Users are known to upload massive amounts of content—up to 500 hours per minute on YouTube alone. Meanwhile, Twitter battles around 8 million accounts weekly that attempt to spread content using manipulative tactics. This scale creates significant challenges for technology to sift through all published materials in such a short timeframe (Westerlund, 2019).

To address this vast amount of data, Google has created educational materials designed to help young people recognize fake news online, potentially extending to visual and audio misinformation. However, it remains uncertain whether such efforts will improve people's ability to recognize deepfakes in everyday internet use. For education to be effective in this area, it must focus on "evergreen" detection methods that remain useful long-term rather than relying on techniques that may quickly become outdated. (CDEI, 2019)

It is crucial to understand that legislation alone will not be sufficient to contain the effects of deepfakes or even shallowfakes. Therefore, investing in advanced screening technology that can detect inconsistencies in visual and audio content, alongside educating the public about manipulated material, is essential (Partnership on AI, 2020). Additionally, some suggest the solution lies in promoting media literacy and fostering "critical skepticism" among the public as a way to counter deepfakes (Siling Tekoniemi, et al, 2022).

Media Literacy and Media Training Approach

The previously discussed mechanisms lack the decisiveness needed to fully address the challenges posed by deepfake technology. Relying solely on one approach will not provide a comprehensive solution due to the limitations of these tools and existing gaps that are difficult to overcome—at least for the time being. This is especially critical as misinformation continues to grow, creating what has been termed an “information disorder” that parallels pathological disorders. Consequently, it has become essential to develop individuals' skills in verifying news, a crucial factor in distinguishing between genuine and fake information (Siling Tekoniemi, et al, 2022).

In an era dominated by communication and information exchange through imagess and audio with their own specific rules, it is vital for individuals to enhance their skills in evaluating news and shared content. They must possess adequate knowledge and awareness of how to navigate the vast and complex landscape of digital content, weighing its potential benefits and harms. In this context, media literacy stands as a modern method for providing individuals with “cognitive immunity” against the dangers of misleading content on media and the internet.

- **Media Literacy Concept**

The concept of media literacy in most references pertains to developing a comprehensive understanding of how to use, analyze, evaluate, produce, and disseminate information and media content. It also encompasses enhancing research skills and the ability to differentiate between reliable and unreliable sources, with an emphasis on critically examining media messages. Media literacy includes understanding how to interact with social media and adapt to digital advancements in news distribution, representing a set of skills individuals must acquire to meet the evolving nature of the information they consume.

In the context of tools to combat deepfake technology, media literacy plays an essential role. Most scholars in this field emphasize that the goal of promoting media literacy is to fundamentally enhance individuals’ capacity to apply critical thinking skills across various media. The core of media literacy lies in fostering advanced levels of critical and creative thinking skills, This includes skills like identifying key concepts, connecting diverse ideas, asking relevant questions, and recognizing misconceptions. This comprehensive approach to literacy underpins intellectual freedom and responsible citizenship within a democratic society (Chanda, 2017).

- **The Importance of Developing Digital Literacy and Skills in Modern Technology**

It has become increasingly essential for society to acquire the skills and abilities necessary to use modern information and communication technologies effectively and securely. These skills range from basic levels (such as typing, setting up network connections, selecting an internet service provider, and running software) to advanced skills (like participating in democratic discussions online, critically evaluating open government initiatives, and contributing creatively to digital culture). (Livingstone, 2007).

This involves asking fundamental questions about what is being presented, noticing what is missing, and fostering a sense of inquiry into the motivations, finances, values, and ownership behind media production—factors that shape content. Media education encourages an exploratory approach, training individuals to ask logical questions about media messages in general, such as: "Who is this message intended for?" "Who wants to reach this audience, and why?" "Whose perspective is this story told from, and who is left out?" and "What strategies are used to capture my attention?" (Malik, 2008)**.**

These skills and insights are particularly relevant to deepfake technology. Despite widespread media coverage and authorities' concerns, there is still a need to raise public awareness about the risks of deepfakes and the potential misuse of artificial intelligence.This technology provides cybercriminals with new tools for social engineering, which requires a high level of vigilance and cybersecurity resilience. Governments, regulators, and individuals need to understand that the outputs of this technology may not accurately represent reality and to learn how to identify signs of deepfake content.

It is recommended to teach critical thinking and digital literacy in schools, as these traits help children detect fake news and interact with each other more respectfully online. Similarly, these skills should be strengthened among older or less technologically savvy people, as everyone needs to be highly aware to critically assess the authenticity and social context of the media they encounter. People should also understand that video quality is not a reliable indicator of authenticity. As technology advances, fewer original images will be needed to create convincing deepfakes, meaning that anyone sharing even one selfie or a video capturing 30 frames per second on social media is potentially at risk.

While the best preventive approach is to avoid posting personal photos or videos online, simple measures like wearing glasses or waving a hand in front of the face can provide some level of protection. Companies and governments that use facial recognition technology and store large amounts of facial data for security purposes must address the risk of identity theft if these databases are ever compromised (Westerlund, 2019).

The most effective way to manage deepfake technology is through education and media literacy. Some individuals need to be made aware of the existence and risks of deepfakes, leading them to engage with online content more thoughtfully and cautiously. Looking back, when Adobe Photoshop was introduced in 1990, many

believed it would be used to distort the truth and hide wrongdoing. However, viewers largely adapted by recalibrating their understanding of images, moving away from the assumption that 'seeing is believing.' A similar adjustment may occur with AI-manipulated visual and audio content and the underlying software tools (CDEI, 2019). Therefore, it is essential that information verification becomes a 'way of life' for every internet user

- **Media training**

When it comes to promoting media literacy and raising awareness among the public, media institutions hold significant responsibility in combating the rise of deepfakes. There is an increasing need for protocols and tools that can effectively identify and prevent the spread of false content (Reardon, 2024). Traditional media, ideally, still carries a sense of trust and credibility in contrast to the misleading and often unreliable information on digital platforms. As such, it should serve as an objective, honest, and transparent alternative.

In journalism, verifying sources should be an inherent part of the editorial process. Constant updating is essential, equipping journalists with tools and resources to ensure accurate information. Adapting to rapid technological advancements requires media organizations and educational institutions in the field to provide journalists with specialized training, enabling them to confront the challenges posed by emerging technologies and deliver ethical, trustworthy content.

For instance, The Wall Street Journal has reportedly formed a forensic media analysis team of 20 members to guide its reporters on identifying fake videos. This team employs detection techniques such as cross-referencing locations with Google Maps and invites academics to discuss the latest innovations in deepfake detection (CDEI, 2019). Similarly, The Washington Post has adopted a standardized approach to combating deepfakes, mirroring its methods for other forms of fake news but with the addition of a team of video experts. Reuters has collaborated with Facebook to detect deepfake content and maintains a dedicated blog focused on debunking deepfakes (Hokkanen, 2021).

Conclusion

Deepfake technology has become a powerful tool for blurring the lines between truth and fake, impacting society in profound ways. Tackling these challenges requires a multi-faceted approach rather than a single solution. Effective approaches include technical methods for digital detection, regulatory and legislative frameworks, and comprehensive media literacy and training. Despite these efforts, significant challenges remain in combating deepfake technology, including:

Rapid advancement of deepfake techniques As new detection tools are developed, deepfake technologies are continually improved to circumvent them, necessitating constant updates to detection methods.

Variety in content manipulation The diversity in methods and types of deepfake content makes it challenging to create a universal tool that addresses all scenarios. This calls for ongoing research across various fields.

Cost and resource requirements: Developing and implementing detection tools and training for their use can be costly, which may limit their widespread adoption.

Lack of accountability on social media and publishing platforms: Platforms often show inconsistency in preventing deepfakes and misinformation. While some efforts are made, economic incentives may encourage them to deprioritize such issues, suggesting a need to revise laws granting them blanket immunity to better serve the public interest.

Legal enforcement challenges: Laws targeting digital misuse and deepfakes are difficult to enforce due to the challenge of identifying the creator, who may also be located in a different jurisdiction, underscoring the need for international cooperation to establish effective deterrents.

Through these integrated approaches, society can mitigate the harmful effects of deepfakes and enhance security and credibility within the digital landscape.

Bibliographie

J. Nightingale, S., & A. Wade, K. (2022, sep). Identifying and minimizing the impact of fake visuals. p. 4. doi:10.1017/mem.2022.8

Shukla , S., & Lyons, T. (2017, aug 28). Blocking Ads From Pages that Repeatedly Share False News. Récupéré sur https://about.fb.com/news/2017/08/blocking-ads-from-pages-that-repeatedly-share-false-news/

Almars, A. M. (2021, may 19). Deepfakes Detection Techniques Using Deeplearning. *Computer and Communications, 9*, pp. 20-35. doi:/10.4236/jcc.2021.95003

AlokeEjike , E., & AbahJos, J. (2023, nov). Enhancing the Fight against Social Media Misinformation: An Ensemble Deep Learning Framework for Detecting Deepfakes. *International Journal of Applied Infor, 12*. doi:10.5120/ijais2023451952

Arash Heidari, et al. (2022, oct). Deepfake detection using deep learning methods:. *ADVANCED REVIEW*, p. 7. doi:10.1002/widm.1520

Bahar Uddin Mahmud, et al. (2019, dec). Deep Insights of Deepfake Technology: A Review. Récupéré sur https://www.researchgate.net/publication/351300442_Deep_Insights_of_Deepfake_Technology_A_Review/citations

CDEI. (2019, september). deepfakes and audio-visual disinformation., (p. 16).

Center for new technology and innovation. (2024, oct 11). Synthetic Media & Deepfakes. Consulté le oct 10, 2024, sur https://innovating.news/article/synthetic-media-deepfakes/

Chanda, N. (2017). Media Education and Media Literacy: Conceptualising the significance of critical and twenty-first-century literacies in media education. *Journal of Content, Community & Communication*, 23.

Chesney, R., & Citron, D. (2019, jan/feb). Deepfakes and the New Disinformation War. *Foreign Affairs*, 147. Récupéré sur https://www.jstor.org/stable/26798018

Chin-Yuan Lin,et al. (2024, april). Video Detection Method Based on Temporal and Spatial Foundations for Accurate Verification of Authenticity. *Electronics*, p. 2. Récupéré sur https://doi.org/10.3390/electronics13112132

Eberl, A. &. (2022). Using deepfakes for experiments in the social sciences - A pilot study. *Frontiers in Sociology*, p. 3.

Floridi, L. (2018, aug). Artificial Intelligence, Deepfakes and a Future of Ectypes. *Philos. Technol, 31*, 321. doi:https://doi.org/10.1007/s13347-018-0325-3

Gupta, G. (2024). A Comprehensive Review of DeepFake Detection Using Advanced Machine Learning and Fusion Methods. *Electronics*, p. 2.

Hokkanen, J. (2021, apr 20). How media outlets and Internet companies fight deepfakes. Récupéré sur https://journalismresearchnews.org/how-media-outlets-and-internet-companies-fight-deepfakes/

Kalina Bontcheva, et al. (2024). *Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities.* University of Sheffield.

Kaur, A., Noori Hoshyar, A., Saikrishna, V. et al. (2024, may). Deepfake video detection: challenges and opportunities. *Artif Intell*, p. 20. Récupéré sur https://doi.org/10.1007/s10462-024-10810-6

Kietzmann, J., & W. Lee, L. (2019, dec). Deepfakes: Trick or treat? *Business Horizons, 63*, 10. doi:https://doi.org/10.1016/j.bushor.2019.11.006.

Kietzmann, Jan, et al. (2019, nov). Deepfakes: Trick or treat? *Business Horizons 63(2)*, p. 3.

Korshunov, p., & Marcel, s. (2019). Vulnerability assessment and detection of Deepfake videos. *International Conference on Biometrics (ICB).* crete. doi:10.1109/ICB45273.2019.8987375.

Lin, C.-Y. (2024, may). Video Detection Method Based on Temporal and Spatial Foundations for Accurate Verification of Authenticity. *Electronics*, p. 10.

Livingstone, S. (2007, may). What is media literacy? p. 18. Récupéré sur http://www.iicom.org/intermedia

Malik, S. (2008). *Media Literacy and its importance.* islamabad: Society for Alternative Media and Research.

Masood, M., & Nawaz, M. (2021, februry). Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. p. 4.

Nguyen, Thanh Thi,et al. (2020, july). Deep Learning for Deepfakes Creation and Detection: A Survey. *arXiv*. Récupéré sur https://dev.arxiv.org/abs/1909.11573v2

Nygren, Thomas, et al. (2022, april). Teachers' views on disinformation and media literacy supported by a tool designed for professional. *SN Social Sciences*, p. 3.

O'Donnell, N. (2021). Have We No Decency? pp. 11-12.

PartnershiponAI.org. (2020). *The Deepfake Detection Challenge:.*

Preeti , R., & Bansal, S. (2024, aug). Exploring Deepfake Detection: Techniques, Datasets and Challenges. *Computing and Digital Systems, 6*. doi:10.12785/ijcds/160156

Ramluckan, T. (2024). Deepfakes: The Legal Implications. Dans 1. I. Security (Éd.)., (pp. 282-). KwaZulu-Natal. doi:10.34190/iccws.19.1.2099

Rana; Nobi, et al. (2022, dec). Deepfake Detection: A Systematic Literature Review. doi:10.1109/ACCESS.2022.3154404

Reardon, S. (2024, jul 10). The Impact of Deepfakes on Journalism. Récupéré sur https://www.pindrop.com/blog/impact-deepfakes-journalism

Schlegel, S. (2024, sep 2). Deepfakes in Court: Real-World Scenarios and Evidentiary Challenges. Consulté le 10 1, 2024, sur https://judgeschlegel.com/blog/deepfakes-in-court-real-world-scenarios-and-evidentiary-challenges

Siling Tekoniemi, et al. (2022, jun 22). Fact-checking as digital media literacy in higher education. *Seminar net*. doi:10.7577/seminar.4689

sultan , k., ruhi, u., & lakhami, r. (2018). Conceptualizing Blockchains: Characteristics & Applications. *Computer Science, ArXiv*, 2.

Taeb , M., & Chi, H. (2022, mar). Comparison of Deepfake Detection Techniques through. *Cybersecurity*, pp. 89-106. doi:10.3390/jcp2010007

Verdoliva, L. (2020, january 18). Media Forensics and DeepFakes: an overiew. *arXiv*, pp. 1-24.

Westerlund, M. (2019, oct). The Emergence of Deepfake Technology: A Review. *Technology Innovation*, p. 44. doi:10.22215/timreview/1282