VISUAL REVIEW | Vol. 18, No. 2, 2026 | ISSN 2695-9631
International Visual Culture Review / Revista Internacional de Cultura Visual
https://doi.org/10.62161/revvisual.v18.5930

ETHICS, PERSONAL IMAGE, AND DISINFORMATION IN THE ERA OF DEEPFAKES

ÁNGEL FERNÁNDEZ FERNÁNDEZ [1]
amfernandez@thecoreschool.com
IREIDE MARTÍNEZ DE BARTOLOME RINCÓN [2]
evaireide.martinez@universidadunie.com
BORJA MORGADO AGUIRRE [3]
morgado@um.es

[1] The Core School / UNIE Universidad
[2] UNIE Universidad
[3] Universidad de Murcia

KEYWORDS
Deepfake
Artificial intelligence
Ethics
Disinformation
Personal image

ABSTRACT

The growing significance of digital visual culture raises significant questions regarding image curation, respect for image rights, and the protection of privacy. In this context, the development of deepfake technology exacerbates these issues by enabling unprecedented audiovisual manipulation. This article analyses the ethical, legal, and social implications of the creation and dissemination of deepfakes, emphasizing the challenges posed by identity falsification. It also examines the control and regulatory measures implemented by digital platforms to detect and limit the distribution of such content, highlighting the need for clear ethical guidelines. Furthermore, it addresses the capacity of deepfakes to reinforce biases and discriminatory narratives, undermining trust in visual information and perpetuating harmful stereotypes in the collective imagination

Received: 27/ 06 / 2025
Accepted: 07/ 09 / 2025

1. Introduction

This research aims to explore the ethical and legal repercussions arising from the creation and dissemination of deepfakes, analysing the responsibility of those who share, edit, or compile such images and proposing tools and ethical guidelines to protect privacy and equity in an environment increasingly affected by visual manipulation. The analysis will focus on two main areas: the impact of deepfakes on individual rights and personal image, and the capacity of this technology to reinforce biases and discriminatory narratives. Deepfakes represent an unprecedented threat to public trust, as they challenge the validity of images and foster large-scale disinformation. It is urgent to address this issue from an interdisciplinary perspective that integrates ethics, law, and technology. (Chesney and Citron, 2019).

1.1. Background

The increasing prominence of visual culture in digital environments has profoundly transformed contemporary communication and representation dynamics. Currently, images, particularly photography and video, have achieved unprecedented communicative significance, shaping our gaze and conditioning how we interpret reality. Since the late 20th century, when digital media were still in their infancy, Baudrillard (1999) already referred to the promiscuity and ubiquity of images, highlighting their capacity for cultural contamination and virality. Today, we live in a state of hypervisuality (Buxó, 1999), a form of digital visual culture in which images exert exceptional symbolic influence. In this new empire of the visual, the image has consolidated itself as the primary communicative medium, permeating and transforming all aspects of our experience (Fernández, 2017; Westerlund, 2019).

This new model of image interpretation poses profound ethical challenges that impact our perception of reality and compel us to reconsider diverse aspects, such as the role of images in constructing our identity, the control of privacy in digital environments, and the management of disinformation (Chesney and Citron, 2019; Vaccari and Chadwick, 2020). These challenges have intensified with recent advancements in Generative Artificial Intelligence (AI), particularly the emergence of deepfakes—fake yet highly realistic images generated through AI systems (Ajder et al., 2019; Bañuelos, 2022). This technology enables the copying, iteration, or substitution of a person's image and voice with another's using deep learning neural networks (Bode et al., 2021).

Currently, the rapid evolution of the technology underpinning deepfakes is challenging traditional notions of authenticity and verisimilitude, undermining trust in the referential nature of images. Although the capabilities of artificial intelligence (AI) are not yet fully defined and regulation of its boundaries remains limited, it is evident that deepfakes are significantly impacting the structure of digital media. While they can be used for legitimate purposes and offer innovative applications in fields such as audiovisual creation, entertainment, or education, their potential for misuse raises serious ethical and legal challenges (Paris and Donovan, 2019). In this regard, Bayer et al. (2021) note that the negative consequences of deepfakes resemble other forms of manipulation and propaganda, but their ability to create highly convincing forgeries amplifies their disruptive potential. The popularisation and virality of deepfakes facilitate their reception, regardless of their veracity, exacerbating this phenomenon (Vaccari and Chadwick, 2020). This raises fundamental questions about consent and the misuse of personal content, as well as the rapid dissemination of manipulated material that can violate the right to one's own image and cause significant reputational harm. The malicious use of deepfakes as a tool for disinformation and manipulative propaganda has the potential to profoundly undermine trust in information and destabilise the social fabric in a particularly severe manner. The verisimilitude of deepfakes clashes with the conventional perception of the image as irrefutable evidence—this exists because it has been photographed—generating concern regarding their use within the framework of disinformation while transforming truth into a mutable reality (Schick, 2020).

1.2. Objectives

This research is guided by the general objective of analysing the ethical and legal repercussions arising from the creation and dissemination of deepfakes, with a specific focus on the issue of identity falsification. This analysis will centre on three specific objectives:

1. To identify how deepfakes can violate the personal image, dignity, and privacy of individuals, causing reputational and psychological harm to victims, perpetuating stereotypes, and fostering disinformation.
2. To review the policies and self-regulatory mechanisms implemented by social media platforms to detect and limit the distribution of deepfakes, evaluating their effectiveness and proposing improvements.
3. To propose ethical and regulatory recommendations that promote transparency, accountability, and the protection of fundamental rights through collaboration between industry, academia, and regulatory bodies.

1.3. Justification

The novelty and originality of this research lie in its systematic approach to the ethical and legal implications of deepfakes within the framework of digital visual culture and social media. While specific studies on image manipulation and disinformation exist, there remains a need to integrate the perspective of personal rights (image, privacy) with the urgency of safeguarding social trust in visual information. Vaccari and Chadwick (2020) note that deepfakes are a powerful tool for political disinformation, underscoring the need to develop effective regulatory strategies. Furthermore, Ajder et al. (2019) highlight the increasing exploitation of this technology for illicit purposes, emphasizing the urgency of coordinated legislative action. "The regulation of deepfakes must balance the protection of fundamental rights with the promotion of technological innovation" (Chesney and Citron, 2019, p. 1795).

In a context where artificial intelligence technologies are rapidly evolving and regulations remain nascent, this work contributes to a multidisciplinary understanding of the phenomenon and provides recommendations for both legislation and the self-regulation of digital platforms.

2. Methodology

The formal object of this research is the phenomenon of deepfakes, understood as the creation of hyper-realistic audiovisual content generated through deep neural networks (GANs), and the adopted perspective considers the ethical and legal consequences of their implementation. The study examines the intersection between the technology enabling image manipulation and the implications of such manipulation for contemporary digital visual culture.

In this regard, this research adopts a qualitative approach, grounded, on the one hand, in a systematic review of the existing literature (Godinho Bilro & Correia Loureiro, 2020; Moher et al., 2009), with the aim of gathering relevant data to identify and conceptualise the main research trends related to the topic under study (Molina Montoya, 2005; Vargas & Calvo, 1987). Thus, a selection of essays exploring deepfakes as phenomena situated at the boundaries of our cultural value system and technology is analysed.

On the other hand, the analysis of deepfakes in their real-world context is based on a case study approach (Castro Monge, 2010; Yin, 1981), which seeks to illustrate the scope of their effects, proving fundamental for understanding their deeper and broader implications, meanings, and possibilities (Gosse & Burkell, 2020; Paris & Donovan, 2019), and delving into theoretical generalisation (Berenguel Fernández & Fernández Gómez, 2018; Rodríguez et al., 1996).

Furthermore, our understanding of deepfakes will deepen as we explore their historical continuities and points of rupture with older practices of media manipulation, technological mediation, fragmentation, and commodification of human images (Bode et al., 2021).

3. Analysis

Data collection for this research was conducted following the criteria of a systematic literature review, gathering information published between 2018 and 2024 on the topic of deepfakes and digital visual culture. Concurrently, case studies were selected based on their media relevance and the legal and ethical debates they sparked, to illustrate the diversity of uses and the magnitude of the risks involved. Once the information from the literature review and the case studies was processed, a procedure for categorising emerging themes (image rights, privacy, digital platforms, disinformation, biases, and discriminatory narratives) was followed, linking the findings to existing theories on digital visual culture and AI ethics. This triangulation between literature, empirical data, and legislative proposals facilitated the identification of recommended lines of action.

3.1. Approach to the Deepfake Concept

The term deepfake is a contraction of "deep learning" and "faked imagery," highlighting the connection between neural network technology and the intentional audiovisual manipulation of discourse. The origin of the term dates back to late 2017, when a Reddit user named Deepfakes posted pornographic videos in which the faces of famous actresses were replaced with those of other women. From a technical perspective, these early deepfakes required large amounts of data and intensive graphical processing. However, in January 2018, the emergence of FakeApp simplified the process, making it accessible even to individuals without advanced technical knowledge.

Although the term deepfake applies to a wide range of formats, it typically refers specifically to photographs and videos created using deep neural networks. In contrast, the concept of cheap fake (Paris & Donovan, 2019) refers to the alteration of images using conventional editing tools (e.g., Adobe Premiere, Final Cut) without a high level of technological sophistication.

Today, most commercial generative artificial intelligence models operate under strict control filters to prevent deepfakes. This is the case with Midjourney, DALL-E 3, Leonardo, and Adobe Firefly, some of the most widely used image-generating AIs, which do not allow creations to closely resemble real faces. However, other AI platforms operate using FLUX models (Black Forest Labs) in their various versions, such as Grok, a model owned by entrepreneur Elon Musk, which has been made freely available to all users of the X platform a few months before the American presidential elections. A subscription-based model, Freepik's AI, uses the FLUX Pro model, the most powerful of these. Generative AIs using this model can achieve levels of fidelity to the original reference that are difficult to detect, complicating the task of discerning their veracity. Most deepfakes are generated on such platforms, but particularly through local and refined AI models, such as ComfyUI and others, using the FLUX engine for processing faces in still images or audiovisual creations. Since AI engines can be downloaded for free, creators install these local models on personal computers costing a few thousand euros and are free to generate all types of unfiltered content, which they subsequently upload to the internet.

3.2 Legal and Ethical Implications of Deepfakes

Deepfakes have undergone rapid advancement, sparking intense debate in both legal and ethical domains due to their capacity to produce highly realistic yet entirely fictitious audiovisual content (Diakopoulos & Johnson, 2021). This technology raises concerns about the potential violation of fundamental rights, such as privacy and personal life (Rizzica, 2021), and creates tension with the freedom of artistic and scientific creation when manipulating the image or voice of third parties without consent. Their potential to generate highly convincing videos increases the risk of extortion and defamation (Greenough, 2022), perpetuates biases (Mukta, Mustak y Naitali, 2023; Mustak et al., 2023), reinforces negative stereotypes, and has the capacity to undermine political stability (Pantserev, 2020).

However, the mere creation of deepfakes does not, in itself, constitute a crime: the key factor in determining illegality lies in the intent and the potential to cause specific harm (Busacca & Monaca, 2023; Kirchengast, 2020; Montasari, 2024). In this regard, some authors advocate for the development of "conscious" artificial intelligence (Ng & Leung, 2020) and emphasize the role of large-scale data management and interoperability in strengthening privacy protection (Nikolakopoulos et al., 2023). Meanwhile, scientific research continues to refine detection methods based on enhancing the robustness of detectors (Lu & Ebrahimi, 2024) and identifying postural discrepancies (Yang et al., 2019). Similarly, from journalistic and neural network perspectives, strategies are being explored to distinguish authentic faces from simulated ones (Haseena et al., 2023; Shilpa et al., 2023; Sohrawardi et al., 2020). These practices, which require time and resources to achieve higher levels of sophistication, generate epistemic distortions and weaken trust in the veracity of images (Liz-López, 2023; Matthews, 2023).

In response to these growing risks, multiple institutions have taken action. The European Parliament (2021) warned of the threats posed by deepfakes in terms of defamation, intimidation, fraud, and electoral manipulation, noting that between 90 and 95 per cent of such content is related to pornography. For its part, the European Union has implemented various mechanisms: Regulation (EU) 2016/679 (GDPR) establishes the importance of explicit consent for the use of personal data and the right to be forgotten, while the Digital Services Act (Regulation (EU) 2022/2065) imposes responsibilities on platforms to protect users and remove illegal content. Additionally, Regulation (EU) 2022/1925 on digital markets and the proposed Artificial Intelligence Regulation of 2021 reflect growing concerns about the generation of false information for political or fraudulent purposes.

At the national level, Spain published the Proposed Organic Law for the regulation of image and voice simulations of individuals generated through artificial intelligence (Congress of Deputies, 2023), in line with the aforementioned European report. This initiative seeks to balance the right to freedom of expression with the protection of honour, privacy, and personal image. To this end, legal reforms are proposed in Law 13/2022, of 7 July, on General Audiovisual Communication, and Organic Law 1/1982, of 5 May, to classify the dissemination of deepfakes without explicit consent as a very serious infringement or an illegitimate intrusion, unless their artificial nature is unequivocally indicated. Similarly, reforms to the Penal Code are proposed to qualify the recreation of a person's voice or image through deepfakes with the intent to harm their honour as a crime of slander, with penalties aggravated when dissemination occurs on social media. Finally, the introduction of a precautionary measure in the Civil Procedure Law is contemplated, authorizing the immediate removal of such content at the request of the affected individual (Congress of Deputies, 2023).

This landscape underscores the need for a dynamic and ongoing legislative approach (Bode et al., 2021), incorporating ethical considerations and user experiences (Li & Wan, 2023), while legal systems adapt to counter both fraud and the manipulation of digital evidence (Montasari, 2024; Mekkawi, 2023; Mahashreshty, 2023). The ability to impersonate identities with great verisimilitude increases the likelihood of humiliation and affects the construction of personal identity (Ayers, 2021). Even recreational or "light" uses of impersonation can perpetuate exclusionary social roles. Consequently, the academic community and regulatory environments advocate for strengthening accountability mechanisms for the malicious use of deepfakes, while continuing to research and develop new methods for detecting and controlling their social impacts.

3.3. Protection of User Privacy in Relation to Platform Policies

The protection of user privacy in digital environments hosting manipulated content—particularly deepfakes—has become a central element in debates about ethics in digital visual culture. According to Bayer et al. (2021), the proliferation of false or misleading information not only distorts the functioning of the rule of law and democratic processes but also violates individuals' personal spheres by exposing their image and data without consent. Thus, audiovisual manipulation endangers the integrity of personal identity by facilitating the reuse of faces and voices for potentially illicit purposes or without authorization. From this perspective, the detection and limitation of the distribution of manipulated content have become essential priorities. Bode et al. (2021), in *The Digital Face and Deepfakes on Screen*, emphasize the need to create specific policies to ensure the protection of user privacy. They also highlight the importance of collaborating with artificial intelligence experts to develop and continually update tools for identifying deepfakes, preventing their rapid obsolescence as the technology evolves. This collaborative approach—involving technology corporations, independent researchers, and regulatory bodies—is critical for mitigating the spread of manipulated content. However, the effectiveness of these policies presents several challenges. According to Bode et al. (2021), several platforms have implemented AI algorithms to recognize and remove manipulated content; however, the speed at which deepfake generation techniques emerge and improve hinders effective suppression before significant impact occurs. Furthermore, Ayers (2021) emphasizes the low transparency of many platforms' moderation processes, which obstructs independent verification of the effectiveness of detection and removal mechanisms. This lack of clarity undermines the trust of both the academic community and users, complicating an objective evaluation of the results achieved. Ultimately, protecting privacy in the context of deepfakes requires a comprehensive approach that combines robust policy design, systematic collaboration with artificial intelligence experts, and greater platform transparency. As long as the technology for generating manipulated content continues its rapid evolution, strategies aimed at safeguarding personal identity and the right to one's own image must be continually renewed and strengthened, addressing both the ethical and legal dimensions underpinning digital visual culture (Ayers, 2021; Bayer et al., 2021; Bode et al., 2021).

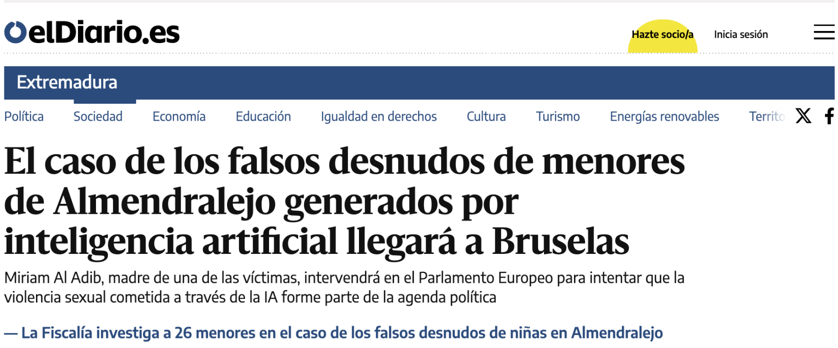
Below, we outline the measures adopted by two of the most prominent platforms, Meta and X, to illustrate the protection of their users' privacy and the evolution of these measures. Meta (formerly Facebook) established a policy aimed at combating the alteration of personal identity through AI, essentially prohibiting the malicious use of deepfakes (Meta Platform, Inc., 2024). This initiative included automated detection tools while allowing the publication of manipulated content for satirical or educational purposes, provided it was clearly labelled. At the end of 2023, Mark Zuckerberg announced the discontinuation of external third-party verification, claiming that independent fact-checking amounted to a form of censorship (Zuckerberg, 2023). Nevertheless, in Spain, organizations such as the Maldita Foundation and Newtral continue to verify posts (García, 2023). For its part, X (formerly Twitter) also implemented labelling policies for potentially manipulated content, alerting users to the altered nature of material before its widespread dissemination (Twitter, 2023). However, in 2023, X relaxed certain restrictions and reduced collaboration with verification entities, leading to an increase in disinformation on the platform. In this context, the European Commission opened a formal investigation against X for alleged violations related to content moderation and transparency under the Digital Services Act (DSA), which requires greater efforts to remove illegal content and preserve privacy and accurate information (Pérez, 2023).

3.4. Case Studies of Malicious Use of Deepfakes

The following outlines significant cases of malicious deepfakes in Spain, Europe, and the United States, reported by the press and information verification organizations over the past two years. Each case presents distinct characteristics, but all involve victims experiencing repercussions in their personal, social, and professional spheres.

1. Deepfakes of sexual content involving minors in Almendralejo (Badajoz). In the summer of 2023, it was revealed that at least 20 minors were victims of classmates who used artificial intelligence to generate false sexual content from their everyday photographs. The incident, still under investigation, sparked debate about the lack of specific legislation and legal gaps in Spain regarding the use of deepfakes. Calls have been made to reform the Penal Code to address identity impersonation and the generation of false content. This case highlights the vulnerability of adolescents to this type of digital manipulation, which not only violates their dignity but also causes profound psychological consequences, such as shame, guilt, and stigmatization within their school environment. The fear of rejection by peers and the circulation of new versions of deepfakes exacerbate their anxiety, while legal systems still lack sufficiently clear or immediate measures for their protection.

Image 1. Screenshot of the article El caso de los falsos desnudos de menores de Almendralejo.



El caso de los falsos desnudos de menores de Almendralejo generados por inteligencia artificial llegará a Bruselas

Miriam Al Adib, madre de una de las víctimas, intervendrá en el Parlamento Europeo para intentar que la violencia sexual cometida a través de la IA forme parte de la agenda política

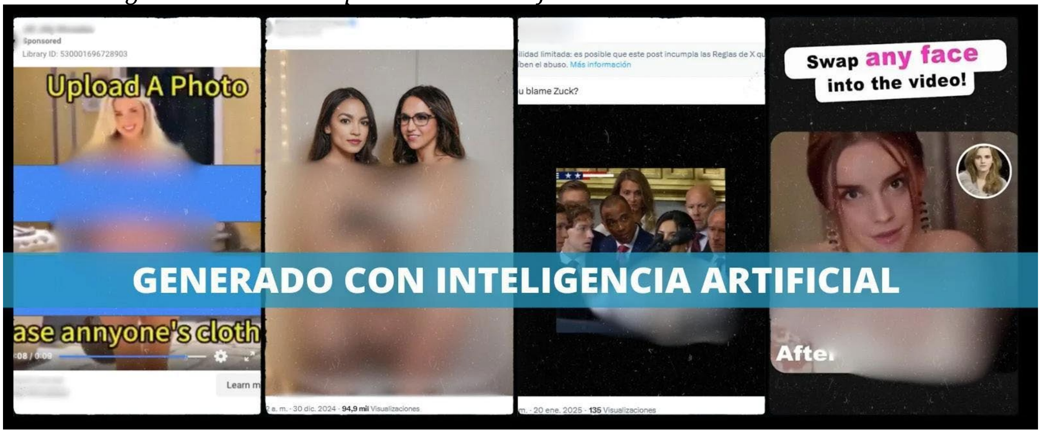
— La Fiscalía investiga a 26 menores en el caso de los falsos desnudos de niñas en Almendralejo

Source: elDiario.es (2023, 8 November). Retrieved from El caso de los falsos desnudos de menores de Almendralejo generados por inteligencia artificial llegar a Bruselas

2. Non-consensual pornographic deepfakes of influencers and public figures. In Europe, cases of non-consensual pornographic deepfakes featuring celebrities and influencers have been reported.

Regarding reputational and psychological harm, the dissemination of false recordings with sexual content causes victims to suffer humiliation, loss of credibility, and professional detriment. This type of gender-based digital violence, along with so-called "revenge porn," exacerbates gender inequality and has a disproportionate impact on women's emotional and social spheres, who are often unfairly blamed for these attacks in public scrutiny. There have been demands for greater legal protection and tools to safeguard victims of these assaults.

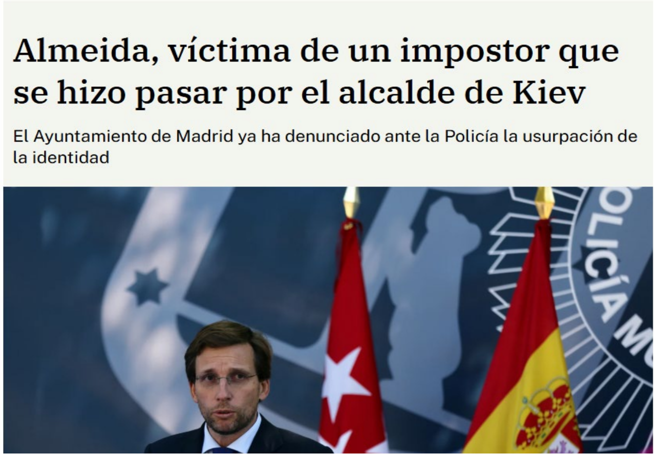
Image 2. Screenshot from the article *From Emma Watson to Alexandria Ocasio-Cortez: how AI is being used to create images and videos that porn or sexualise female celebrities without their* consent.


Upload A Photo
ase annyone's cloth
Swap any face into the video!
GENERADO CON INTELIGENCIA ARTIFICIAL
After
u blame Zuck?

Source: Damn Technology (2025, 22 January). Retrieved from https://maldita.es/malditatecnologia/20250122/ia-imagenes-porno-mujeres-famosas/

3. Deepfakes of public figures. The impersonation of public figures in Europe has been equally alarming, as evidenced by the case of the Mayor of Madrid, who engaged in a virtual conversation with a deepfake of the Mayor of Kyiv, Vitali Klitschko. Similar incidents affected the Mayor of Berlin and the Mayor of Vienna. These actions aim to discredit public figures and spread disinformation. Furthermore, the fear of being targeted by similar attacks generates significant stress, exposing these figures to constant questioning about the authenticity of their appearances, which hinders their public activities and erodes their reputation and even the credibility of the institutions they represent. The European Union has strengthened cooperation between institutions and is discussing the need to classify the creation and dissemination of malicious deepfakes as a specific offence

Image 3. Screenshot of the article *Almeida, victim of an impostor impersonating the mayor of Kiev.*



Almeida, víctima de un impostor que se hizo pasar por el alcalde de Kiev

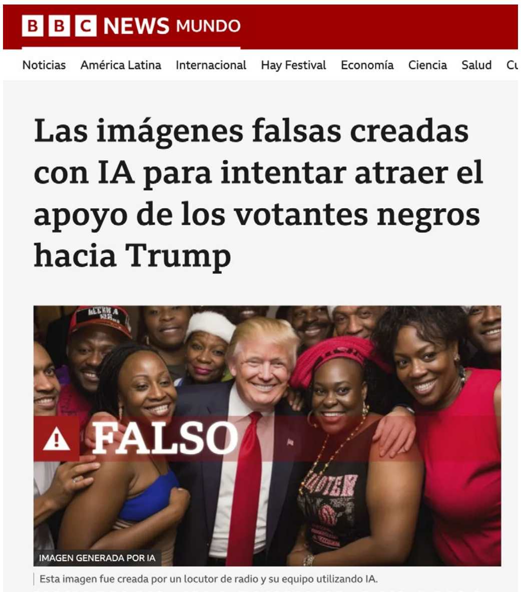
El Ayuntamiento de Madrid ya ha denunciado ante la Policía la usurpación de la identidad

Source: La Razón (2022, 25 June). Retrieved from Almeida, victim of an impostor who impersonated the mayor of Kiev

4. Deepfakes in political campaigns and election disinformation in the United States. In U.S. political campaigns, deepfakes have proliferated for disinformation purposes, manipulating speech excerpts or attributing extremist messages to candidates. In this context, since 2023, fact-checking organizations have warned about the rise of deepfake videos used in political contexts. This malicious use of technology erodes trust in democratic institutions and fosters an atmosphere of scepticism

and political confrontation. Candidates and their teams face additional stress, being forced to constantly debunk falsehoods, which reduces their ability to focus on legitimate political communication. Society, in turn, is drawn into a climate of polarization and suspicion that affects collective mental stability. In response to these events, social media platforms such as the X platform and Meta have strengthened content moderation and labelling policies for manipulated content, although the effectiveness of these measures remains in question

Image 4. Screenshot of the news story *Fake images created with AI to try to attract the support of black voters in the US*



BBC NEWS MUNDO

Noticias América Latina Internacional Hay Festival Economía Ciencia Salud Cu

Las imágenes falsas creadas con IA para intentar atraer el apoyo de los votantes negros hacia Trump

Esta imagen fue creada por un locutor de radio y su equipo utilizando IA.

Source: BBC News World (2023, April 1). Retrieved from https://www.bbc.com/mundo/articles/c3g4l5xgvryo

4. Conclusions and Discussion

Deepfakes constitute a high-impact phenomenon in contemporary digital visual culture, with the capacity to harm image and reputation, violate privacy, and undermine the credibility of audiovisual information. The research conducted demonstrates how the ease of creating hyper-realistic deepfakes exposes both public figures and private individuals to these manipulations, where their faces and voices can be altered without consent, resulting in psychological harm. Furthermore, the increasing accessibility of deepfakes, combined with the hyper-connected digital environment and the virality of social media, accelerates their widespread dissemination, complicating the removal of falsified content. In this regard, deepfakes enhance the effectiveness of smear campaigns or political deception, eroding public trust in the veracity of audiovisual materials and fostering widespread scepticism among citizens. Additionally, it is evident that audiovisual manipulation can target specific groups, perpetuating stereotypes and promoting symbolic violence on social media. Regarding platforms, despite efforts to develop automated detection methods, the rapid evolution of deepfakes widens the gap between creators and platforms. Although increasing efforts are identified in developing automated detection methods and proposing specific regulatory frameworks, their practical application and large-scale effectiveness require further development and harmonization. This legal gap, coupled with the constant improvement of technological tools, hinders the response of digital platforms and leaves victims without adequate protection. Moreover, deepfakes not only threaten individual privacy but also have significant political and social implications, from interfering in electoral campaigns to perpetuating hate speech. Consequently, a comprehensive approach is required, integrating legislation, ethics, the responsibility of digital platforms, and citizen education in content verification and critical consumption of audiovisual materials.

In line with authors such as Chesney and Citron (2019), it is observed how deepfakes contribute to a “new war on disinformation” opened by the possibility of near-undetectable audiovisual manipulation. The persistence of cases involving defamation and abuse corroborates Vaccari and Chadwick’s (2020) reflections on the erosion of public trust in the veracity of information, highlighting the risk of systemic scepticism towards all visual content. Similarly, Bode et al. (2021) underscore the complexity faced by digital platforms in addressing a technology in constant refinement, reinforcing the observed limited effectiveness of moderation policies.

The discussion extends to the legal dimension, where Kirkengast (2020) and Meskys et al. (2020) highlight the urgency of enacting laws that address deep audiovisual manipulation, moving beyond the residual application of traditional offences. Li and Wan (2023) also emphasize the need to consider user experience and ethics in the creation of such content, as not all deepfakes are produced with harmful intent.

In the broader sphere of digital visual culture, these conclusions align with Bañuelos's (2022) thesis on the rapid transformation of the media ecosystem and Ajder et al.'s (2019) observations regarding the capacity of deepfakes to generate highly convincing forgeries. Consistent with Brown and Fleming (2020) and Maddocks (2020), the malicious use of this technology reinforces gender-based violence and other forms of harassment, demonstrating how AI can amplify pre-existing power dynamics and perpetuate biases of all kinds.

Finally, the relevance of Fletcher's (2018) perspective on the "deterioration of truth" in the post-factual era is evident: the mere possibility that content may be a deepfake extends doubt to all audiovisual material. This scenario calls for reflection on the shared responsibility of governments, AI developers, digital platforms, and users to deploy increasingly precise detection techniques (Rössler et al., 2019; Tolosana et al., 2020) while promoting digital literacy to equip society for critical engagement with the images and videos it consumes.

In summary, following the review of existing literature and analysis of case studies, it can be concluded that the convergence of empirical findings and theoretical debates confirms the need for a comprehensive approach to curb the malicious use of deepfakes in digital visual culture, focusing on three priority lines of action:

Firstly, adequate legislation is required that not only prohibits the creation and dissemination of malicious deepfakes but also continuously adapts to technological advancements. This includes imposing significant fines on platforms that allow the proliferation of such content without effective detection or labelling tools. An example is the European regulation, which, through the Digital Services Act (DSA), mandates very large online platforms to identify, analyse, and mitigate systemic risks such as disinformation, hate speech, and illegal content. This includes working with fact-checking organizations to detect and limit the spread of false information. As an additional example, Chinese legislation mandates that AI-generated content include visible watermarks or identifying metadata, penalizing platforms that fail to comply with these requirements (Cyberspace Administration, 2024). Similarly, initiatives like California's AB 730 law, which restricts deepfakes in electoral contexts, demonstrate the potential of specific regulations to mitigate risks in sensitive areas (AB 730, 2019). The creation of a global regulatory framework that holds digital platforms accountable and ensures transparency is an urgent need.

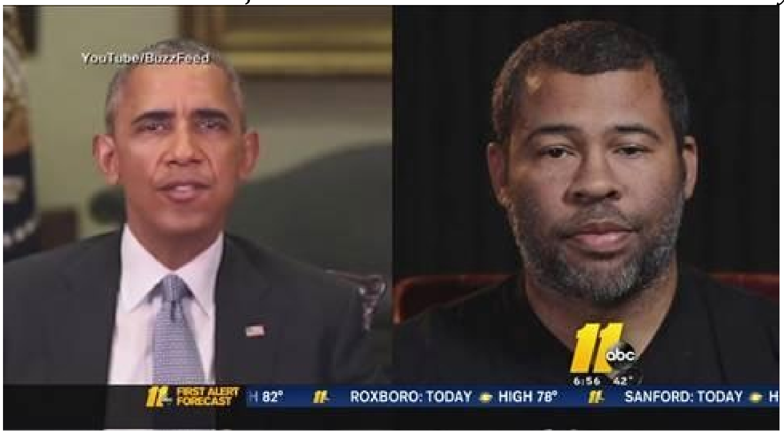
Secondly, the technology industry and external stakeholders must take an active role by implementing transparency standards and developing advanced detection tools (Sanchez et al., 2024). Protocols such as those developed by Adobe's Content Authenticity Initiative enable the tracking and verification of AI-generated content through digital watermarks. Additionally, initiatives like Microsoft's Video Authenticator, designed to analyse videos for manipulations, show how technology companies can contribute to controlling this phenomenon (Microsoft, 2020). Collaboration among major technology players, as in the case of the Partnership on AI consortium, establishes a model of cooperation that should be expanded to include a global repository of solutions accessible to governments and users (Partnership on AI, 2020).

Lastly, it is essential to enhance visual education and media literacy to equip citizens with critical tools against deepfakes. Educational programmes like those implemented in Finland, which integrate media literacy across all levels of the education system, stand out for their effectiveness in fostering critical and resilient citizens against disinformation (European Commission, 2019). These initiatives should be expanded globally, incorporating specific modules on content verification, teacher training, and public awareness campaigns. Educating the population to recognize digital manipulations is thus essential to strengthen society against disinformation (Ajder et al., 2019).

In the realm of awareness-raising and media literacy regarding deepfakes, one of the first notable examples was the production by Jordan Peele and BuzzFeed (Peele & BuzzFeed, 2018), which featured former President Barack Obama making false statements. Using artificial intelligence to synchronize

Peele's voice and gestures with Obama's image, this project aimed to demonstrate, for educational purposes, the ease of manipulating audiovisual content.

Image 5. Screenshot of the video in which Jordan Peele *You Won't Believe What Obama Says in This Video!.*



Source: Peele & BuzzFeed, (2018).

Similarly, Joan Fontcuberta's (2017) exhibition *A chupar del bote*—attributed to a fictional reporter—presented manipulated documents and portraits to underscore the urgency of critical consumption of images (Fontcuberta y Fernández, 2017).

Image 6. Image from the exhibition *A chupar del bote* by Ximo Berenguer (Joan Fontcuberta).

Source: Fontcuberta, J. (2017). Exhibition at PhotoEspaña 2017.

This concern was also addressed in the exhibition *Fake News. La fábrica de mentiras* (2023-2024), displayed at Espacio Fundación Telefónica, which explored disinformation from a historical perspective and proposed a Decalogue to combat it (Fundación Telefónica Movistar, 2024).

Image 7. View of the exhibition *Fake News. El valor de la información* at the Fundación Telefónica Movistar, Buenos Aires, 2024.

Source: Telefónica Movistar Foundation (2024).

Lastly, the initiative *This Person Does Not Exist* (ThisPersonDoesNotExist.com, n.d.) employed generative neural networks to create hyper-realistic faces of non-existent individuals, highlighting the potential of AI to generate fictitious identities and fostering reflection on the ethics and risks associated with these practices.

Image 8. Face generated by artificial intelligence, using AI and not representing any real person.

Source: ThisPersonDoesNotExist.com (n.d.). Retrieved from https://thispersondoesnotexist.com.

Ultimately, addressing the challenges posed by the proliferation of deepfakes requires these three factors, working harmoniously in interdependence. The experiences of countries such as China, the United States, and Finland demonstrate that effective measures can be implemented when acting on multiple fronts. However, international coordination and collaboration among governments, the technology industry, and citizens will be crucial to ensure that these initiatives have a tangible impact on building a safer and more transparent digital ecosystem.

References

Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019). The State of Deepfakes: Landscape, Threats, and Impact. Deeptrace Labs Report. Disponible en: https://regmedia.co.uk/2019/10/08/deepfake_report.pdf

Assem. Bill 730, Ch. 493. (Cal. 2019). https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB730

Ayers, D. (2021). The limits of transactional identity: Whiteness and embodiment in digital facial replacement. *Convergence, 27(4),* 1018-1037. 10.1177/13548565211027810

Bañuelos, J. (2022). Evolución del Deepfake: campos semánticos y géneros discursivos (2017-2021). *Revista ICONO 14. Revista Científica De Comunicación Y Tecnologías Emergentes, 20(1).* https://doi.org/10.7195/ri14.v20i1.1773

Baudrillard, J. (1999). *Sur la photographie.* París: Sens & Tonka.

Bayer, J., Holznagel, D. B., Katarzyna, L., Pace, A., Szakács, J., & Uszkiewicz, E. (2021). Disinformation and propaganda: Impact on the functioning of the rule of law and democratic processes in the EU and its Member States: 2021 update. *European Parliament Policy Department for External Relations.* https://bit.ly/3WHc8ZM

Berenguel Fernández, J., & Fernández Gómez, J. (2018). La eficacia de la comunicación en la convergencia mediática: propuesta de metodología de estudio y aplicación de casos. *Trípodos, (43),* 37-56.

Berenguer, X. (2016). *A chupar del bote*. RM Verlag

Bode, L., Lees, D., & Golding, D. (2021). The Digital Face and Deepfakes on Screen. *Convergence, 27(4),* 849-854. 10.1177/13548565211034044

Brown, M., & Fleming, E. (2020). Celebrity headjobs: Or oozing squid sex with a framed-up leaky. Porn Studies, 7(4), 357-366. http://dx.doi.org/10.1080/23268743.2020

Busacca, A., & Monaca, M. A. (2023). Deepfake: Creation, Purpose, Risks. En Marino, D. y Monaca, M. A. (eds.), *Innovations and Economic and Social Changes due to Artificial Intelligence: The State of the Art* (pp. 55-68). Springer Nature Switzerland. 10.1007/978-3-031-334610_6

Buxó, M.J. (1999). *De la investigación audiovisual. Fotografía, cine, vídeo, televisión.* Barcelona: Proyecto A Ediciones.

Castro Monge, E. (2010). El estudio de casos como metodología de investigación y su importancia en la dirección y administración de empresas. *Revista Nacional de Administración, 1(12),* 31-54. doi: https://doi.org/10.22458/rna.v1i2.332

Chesney, R., & Citron, D. (2018). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review, 107(6),* 1753-1816. DOI: https://doi.org/10.2139/ssrn.3213954

Chesney, R., & Citron, D. (2019). Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics. *Foreign Affairs.* Disponible en: https://www.jstor.org/stable/26798018

Citron, D. (2014). *Hate crimes in cyberspace.* Harvard University Press.

Congreso de los Diputados. (2023). Proposición de Ley [BOCG, Serie B, 15-B-23-1].https://bit.ly/40CHHFn

Cyberspace Administration of China. (2024). Measures for the Identification of AI-Generated Synthetic Content [Medidas para la identificación de contenido sintético generado por IA]. http://www.cac.gov.cn/

Diakopoulos, N., & Johnson, D. (2021). Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media & Society, 23(7),* 2072-2098. 10.1177/1461444820925811

European Commission. (2019). Youth policies in Finland: 2019. European Education and Culture Executive Agency.

Fernández, Á. (2017). Relatos híbridos: El papel de la narratividad en la visualización de información interactiva [Tesis doctoral, Universidad Europea]. Repositorio Abacus https://193.147.239.238/handle/11268/6981

Fletcher, R. (2018). *The truth behind fake news in a post-factual age.* Palgrave Studies in Communication.

Fontcuberta, J. (2017). A chupar del bote [Exhibición]. Galería Fernando Pradilla, Madrid, España.

Fontcuberta, J., y Fernández, H. (2017). A chupar del bote. Editorial RM.

Fundación Telefónica. (2023). *Fake News. La fábrica de mentiras* [Exposición]. Madrid, España: Espacio Fundación Telefónica. Recuperado de https://bit.ly/3EeEBQo

García, M. (2023, octubre 15). *Zuckerberg elimina verificación de datos en Facebook e Instagram, modelo de Elon Musk en X*. ElDiario.es. https://bit.ly/40TiUhM

Godinho Bilro, R., & Correia Loureiro, S. (2020). A consumer engagement systematic review: synthesis and research agenda. *Spanish Journal of Marketing - ESIC, 24(3),* 283-307.

Gosse, C., & Burkell, J. (2020). Politics and porn: how news media characterizes problems presented by deepfakes. *Critical Studies in Media Communication*, *37*(5), 497-511.

Greenough, C. J. (2022). Make it, Fake it and Get Away with it? The Role of Toxic Masculinity and Threat Perception within Cases, Policies, and Legislation Surrounding Deep Fakes. Master Thesis, University of Auckland. https://hdl.handle.net/2292/61261

Haseena, S., Saroja, S. & Nivetha, A. (2023). TVN: Detect Deepfakes Images using Texture Variation Network. *Inteligencia Artificial*, *26*(72), 1–14. https://doi.org/10.4114/intartif.vol26iss72pp1-14

Kirchengast, T. (2020). Deepfakes and image manipulation: Criminalisation and control. *Information & Communications Technology Law, 29(3),* 308-323. 10.1080/13600834.2020.1794615

Li, M., & Wan, Y. (2023). Norms or fun? The influence of ethical concerns and perceived enjoyment on the regulation of deepfake information. *Internet Research, 33(5),* 1750-1773. 10.1108/INTR-07-2022-0561

Liz-López, H, Keita, M., Taleb-Ahmed, A. Abdenour, H., Huertas-Tato, J., Camacho, D. (2023). Generación y detección de contenidos audiovisuales multimodales manipulados: Avances, tendencias y desafíos abiertos. *Fusión de Información*, pp.102-103

Lu, G., & Ebrahimi, T. (2024). A New Approach to Improve Learning-based Deepfake Detection in Realistic Conditions. *arXiv preprint arXiv:2203.11807*.

Maddocks, S. (2020). Deepfake Porn, Revenge Porn, and Gendered Violence. *Feminist Media Studies, 20(7),* 976–990.

Mahashreshty, V. S. (2023). *Implications of deepfake technology on individual privacy and security* (Culminating Projects in Information Assurance No. 142). St. Cloud State University. https://repository.stcloudstate.edu/msia_etds/142

Matthews, T. (2023). Deepfakes, fake barns, and knowledge from videos. *Synthese, 201* (2). https://doi.org/10.1007/s11229-022-04033-x.

Mekkawi, L. (2023). The challenges of Digital Evidence usage in Deepfake Crimes Era. *Journal of Law and Emerging Technologies*, *3*(2), 176-232.

Meskys, E., Kalpokiene, J., Jurcys, P., & Liaudanskas, A. (2020). Regulating deep fakes: legal and ethical considerations. *Journal of Intellectual Property Law & Practice, 15(1),* 24-31.

Meta Platforms, Inc. (2024, 5 de abril). *Our approach to labeling AI-generated content and manipulated media*. Meta. https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/

Microsoft. (2020, septiembre 2). Microsoft anuncia novedades en su programa Defending Democracy para luchar contra las campañas de desinformación. Microsoft News Center. https://news.microsoft.com/es-es/2020/09/02/microsoft-anuncia-novedades-en-su-programa-defending-democracy-para-luchar-contra-las-campanas-de-desinformacion/

Ministerio del Interior (2023). Informe de Cibercriminalidad en España 2023. *(Referencia sin URL oficial.)*

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., ... PRISMA-P Group. (2009). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev, 4(1).*

Molina Montoya, N. P. (2005). Herramientas para investigar: ¿Qué es el estado del arte? *Ciencia y Tecnología para la Salud Visual y Ocular, (4),* 73-75.

Montasari, R. (2024). Responding to Deepfake Challenges in the United Kingdom: Legal and Technical Insights with Recommendations. En Montasari, R. (ed.), *Cyberspace, Cyberterrorism and the International Security in the Fourth Industrial Revolution: Threats, Assessment and Responses* (pp. 241-258). Springer International Publishing. 10.1007/978-3-031-50454-9_12

Mukta, N., Mustak, M., & Naitali, A. (2023). *An investigation of the effectiveness of deepfake models and tools. Journal of Sensor and Actuator Networks, 12(4), 61. https://doi.org/10.3390/jsan12040061*

Mustak, M., Salmien, J., Mäntymäki, M., Rahman, A. & Dwivedi, Y.K. (2023). Deepfakes: Deceptions, Mitigations and Opportunities. Journal of Business Research, 154.

Ng, A., & Leung, M. (2020). Toward a strong AI for deepfake governance: Ethical implications. (Referencia sin datos.)

Nikolakopoulos, A., Segui, M., Pellicer, A. B., Kefalogiannis, M., Gizelis, C., Marinakis, A., Nestorakis, K., & Varvarigou, T. (2023). BigDaM: Efficient Big Data Management and Interoperability Middleware for Seaports as Critical Infrastructures. Computers, 12(11), Article 11. https://doi.org/10.3390/computers12110218

Pantserev, K. A. (2020). The Malicious Use of AI-Based Deepfakes Technology as the New Threat to Psychological Security and Political Stability. En Jahankhani, H., Kendzierskyj, S., Chelvachandran, N. y Ibarra, J. (eds.), Cyber Defence in the Age of AI, Smart Societies and Augmented Humanity (pp. 37-55). Springer International Publishing. 10.1007/978-3-030-35746-7_

Paris, B., & Donovan, J. (2019). Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence. Data & Society Research Institute. Disponible en: https://datasociety.net/library/deepfakes-and-cheap-fakes/

Partnership on AI. (2020, 12 de marzo). The Deepfake Detection Challenge: Insights and recommendations for AI and media integrity. https://partnershiponai.org/wp-content/uploads/2021/07/671004_Format-Report-for-PDF_031120-1.pdf

Peele, J., & BuzzFeed. (2018). You Won't Believe What Obama Says In This Video! [Video]. YouTube. https://www.youtube.com/watch?v=cQ54GDm1eL0

Pérez, J. (2023, diciembre 18). *Bruselas abre una investigación formal contra X por vulnerar las normas sobre moderación de contenidos.* Cinco Días. https://bit.ly/3El4b6n

Rizzica, F. (2021). Protección del bienestar en la era de los deep fakes. (Referencia sin datos concretos.)

Rodríguez, G., Gil, J., & García, E. (1996). Metodología de la investigación cualitativa. Ediciones Aljibe.

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Niessner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Sánchez, M., Palella, S., & Fernández, A. (2024). Implementación de herramientas de Inteligencia Artificial en la detección de vídeos falsos y ultrafalsos (deepfakes): Caso de Radio Televisión Española (RTVE). VISUAL REVIEW. International Visual Culture Review/Revista Internacional de Cultura Visual, 16(4), 213-225.

Schick, N. (2020). *Deepfakes: the coming infocalypse.* Grand Central Publishing.

Shilpa, A., Varma, K., & Sur, S. (2023). Evaluating convolutional neural networks for deepfake detection. (Referencia sin datos concretos.)

Sohrawardi, S. J., Seng, S., Chintha, A., Thai, B., Hickerson, A., Ptucha, R., & Wright, M. (2020). Defaking DeepFakes: Understanding journalists' needs for DeepFake detection. In *Proceedings of the Computation+ Journalism 2020 Conference* (Vol. 21). Northeastern University.

Tolosana, R., Vera-Rodríguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deep Fakes and beyond: A survey of face manipulation and fake detection. Information Fusion, 64, 131-148.

UNESCO (2022). Recomendación sobre la Ética de la Inteligencia Artificial. https://unesdoc.unesco.org/ark:/48223/pf0000380455

Vaccari, C., & Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society, 6(1).*1-13. https://doi.org/10.1177/2056305120903408

Vargas, J., & Calvo, V. (1987). El estado del arte en la investigación educativa: su definición y significado. Perfiles Educativos, 35, 5-15.

Wagner, T. & Blewer, A. (2019). "The Word Real Is No Longer Real": Deepfakes, Gender, and the Challenges of AI-Altered Video. Open Information Science, 3(1), 32-46. https://doi.org/10.1515/opis-2019-0003

Wang, P. (2019). This person does not exist [Portal web]. Recuperado de https://thispersondoesnotexist.com

Westerlund, M. (2019). The Emergence of Deepfake Technology: A Review. Technology Innovation Management Review, 9(11), 39-52. DOI: https://doi.org/10.22215/timreview/1282

X. Yang, Y. Li, & S. Lyu (2019). Exposing deep fakes using inconsistent head poses. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 8261–8265.

Yin, R. (1981). The case study crisis: Some answers. Administrative Science Quarterly, 26(1), 58-65

Zuckerberg, M. (2023, octubre 20). *Actualización sobre las nuevas funcionalidades de Facebook* [Video]. Facebook. https://www.facebook.com/zuck/videos/1525382954801931