VISUAL REVIEW | Vol. 18, No. 3, 2026 | ISSN 2695-9631
International Visual Culture Review / Revista Internacional de Cultura Visual
https://doi.org/10.62161/revvisual.v18.5893

GENERATIVE AI AND EFFECTIVENESS OF DEEPFAKE DETECTION
Analysis of the Effectiveness of Deepfake Detection Capabilities Between Artificial Intelligence and the Human Brain.

NATHALIE ALEJANDRA RODRÍGUEZ EGAS [1]
nathalir@ucm.es
DANIEL GÓMEZ GONZÁLEZ [2]
dagomez@ucm.es
JAVIER SIERRA SÁNCHEZ [1]
javier_sierra@ucm.es
NARCISA JESSENIA MEDRANDA MORALES [3]
nmedranda@ups.edu.ec
[1] Universidad Complutense de Madrid, España
[2]Universidad Complutense de Madrid, Instituto de Estadística y Ciencia de Datos, España
[3] Universidad Politécnica Salesiana, Ecuador

KEYWORDS

Deepfake
Artificial intelligence
OpenAI's
Detection capacity
Deep learning
Neural networks
Identity theft

ABSTRACT

Artificial intelligence has revolutionized multiple industries, particularly in the generation of images and videos. The emergence of tools capable of creating hyper-realistic videos poses serious risks in identity theft. This article examines the phenomenon of deepfakes and, factors that influence their identification, and presents an experimental study that evaluates the effectiveness of detecting these contents using deep neural networks. It compares the performance of a group of people with a tool based on deep learning technology, providing a detailed analysis of the strengths and limitations of both approaches.

Received: 11/ 06 / 2025
Accepted: 15/ 09 / 2025

1. Introduction

Artificial Intelligence (AI) is a field of computer science focused on creating systems capable of performing tasks that typically require human intelligence, such as learning, reasoning, and perception (Recovery, Transformation and Resilience Plan, 2023).

This refers to digital systems that can reason and solve problems even with large volumes of data, facilitating tasks such as content creation, statistical predictions, or automation. AI can be defined from various perspectives, depending on its typology, models, or languages employed. According to its typology, the philosopher John Searle (1980) classifies artificial intelligence into weak AI and strong AI. He defines weak AI as that which operates through programming, lacking its own causal powers or intentionality, and thus does not simulate the mind, meaning it performs specific functions with processes distinct from those of the brain. Strong AI, in contrast, is described by the author as a mind in itself, as it possesses traits equivalent to human brain processing, autonomy in decision-making, and learning capacity. However, Searle's main hypothesis is that no type of programme or machine can match the causal capacities of the human brain: "Any attempt to create intentionality artificially, strong AI, could not succeed merely by designing programmes but would need to replicate the causal capacities of the human brain, and no programme alone is sufficient for thinking" (p. 1).

Turing (1980), in his writings, argued that a computer exhibits intelligent behaviour, likening it to cognitive development that emulates the human mind, which he termed the imitation game, and foresaw the importance of machine learning in AI development. In 1956, the mathematician McCarthy introduced the term Artificial Intelligence for the first time during the Dartmouth Conference, where pioneers and visionaries such as Minsky, Shannon, and Rochester, among others, convened. Aware that the computers of that era had limited capacity to mimic human brain functions, they anticipated their evolution through the automation of computers. They considered that advancements would be significant and identified seven key issues: automation; computer use of language; neural networks; computational efficiency; self-learning; abstractions; and the relationship between chance and creativity. In their proposal, they defined AI as the learning of concepts and abstractions that consolidate algorithms similar to human intelligence with such precision that they can improve or solve problems effectively (McCarthy et al., 2006). At this conference, McCarthy et al. (2006) developed the information theory and antonomasia theory, proposing models that resemble the human brain. In line with this, Russell and Norvig (2004) propose a classification based on the cognitive model approach, delving into systems that think and act like humans. This aligns with the theory of mind advanced by Woodruff and Premack (1979), which discusses the brain's inference system, attributing capacities such as consciousness and the intent of the mind's deductive system.

Kosinski (2024), a renowned Stanford researcher, analysed current AI language models such as ChatGPT and, while acknowledging their preliminary nature, highlighted their significant potential in the theory of mind. "Let us not forget that we are witnessing exponential progress, as AI models double their performance every year," Kosinski stated in an interview with *Infobae* (Fernández, 2024), emphasising that machines now interpret and capture a previously unseen social component.

López de Mántaras and Meseguer (2017) define the main AI models as symbolic, connectionist, evolutionary, and embodied. The symbolic model operates with abstract representations modelled through non-embodied representation languages based on mathematical logic and its extensions, which can be useful for proving theorems. The connectionist model, also called bio-inspired AI, proposed by McCulloch and Pitts in 1943, supports the idea that a neuron is fundamentally linked to logic. The evolutionary model, based on concepts of evolutionary computation, focuses on improving computers' ability to solve problems for which they were programmed through operators and programme development, with solutions being improved versions. However, these authors highlight several challenges and limitations, which are better explained in section 1.1, The Limitations of Generative AI.

According to Douglas Eck, Senior Research Director at Google, the most prominent AI language models are: Predictive: This uses a machine learning system for decision-making in future scenarios and can classify information about language, but it cannot generate new data from its programming. Generative: This is trained on large volumes of data; an example is Gemini, which is working on creating images from text and automation sequences (Google Cloud, 2025). Machine learning inherent to the generative model has the capacity to create entirely new content based on the programming and data available (Eben, 2023). Large language models (LLMs) are a type of generative AI, as they create new combinations of text in the form of natural language and images.

According to their developmental stages, as outlined on the official Google Cloud portal (2025), four phases are defined: reactive machines; limited memory; theory of mind; and self-aware machines. Adapted to the current context, weak AI corresponds to reactive machines, which respond to programmed rules but lack memory and cannot learn new data—for example, Deep Blue, the computer programmed to play chess. Strong AI, conversely, can be trained with new data through artificial neural networks, generally evident in limited memory models or LLMs based on machine learning and deep learning. The other two categories remain undeveloped at present, but they involve AI based on the theory of mind, which simulates human brain reasoning through decision-making and emotion recognition capabilities. A step beyond this theory is self-aware AI, with feelings and self-perception akin to a human being (Google Cloud, 2025).

This article presents an exploration of the capabilities of artificial intelligence and humans in detecting deepfakes, along with the limitations and strengths of both perspectives. It addresses ethical concerns, existing regulations, and risks associated with deepfake use, such as identity impersonation. Additionally, the text highlights the growing prevalence of deepfakes, particularly in non-consensual pornographic content, and the need for greater digital literacy and accessible detection tools. The questions it seeks to address are as follows:

P1. What are the limitations of generative AI as a mechanism for deepfake detection?

P2. What is the success rate of a target group compared to an algorithm in detecting a deepfake?

P3. What factors can enhance deepfake detection in digital environments?

P4. What prior experiments and research analyse this issue? To address these questions, the deepfake phenomenon is analysed with current figures, theories, and prior studies comparing the reasoning and detection capabilities of an algorithm versus the human brain, aiming to pursue a line of research on the determining factors for detecting photorealistic images and videos.

1.1. The limitation of Generative AI

The most recurrent limitations relate to potential risks such as toxic language and social biases. It is imperative to ensure that necessary measures for confidentiality and data access have been established when using these technologies.

AI can also produce hallucinations, which occur when responses are generated out of context. Its outputs depend on the data it has been trained on and the quality of that content.

Von Neumann (1980) analyses the parallel processing of information in machines compared to human intelligence, attempting to understand the complexity of the human brain and the neural functions that have informed the programming of machines and systems based on artificial intelligence algorithms, such as machine learning and deep learning.

There are programmes capable of generating intelligent problem-solving behaviours, namely evolutionary programming models aimed at general intelligence. Regarding this premise, Lopez and Meseguer (2017) state the following:

> "The reality is far more complex, and this approach has numerous limitations. One of the strongest criticisms of these non-embodied models is that an intelligent agent requires a body to have direct experiences with its environment, rather than a programmer providing abstract descriptions of that environment" (López de Mántaras and Meseguer, 2017, p. 14).

This confirms the idea that direct interaction can generate internal representations for rational decision-making, wherein the agent and the environment-context play a significant role, known as situated cognition. This relates to the limitations of representation and reasoning based on mathematical logic. AI has not yet overcome these limitations; however, this has opened a line of research within the subfield of robotic AI development.

The computational architecture model, as expressed by the author in his book *The Computer and the Brain*, has been considered. This establishes foundations for the convergence of neuroscience and computer science. The most relevant differences highlighted in this article are as follows:

1.2 Logical reliability and arithmetic precision:

Statistical systems inherent to machines, whether digital or analogue, depending on information processing, tend to exhibit greater arithmetic precision. However, their results can be distorted or altered by a minor error in programming, leading to hallucinations in AI responses, which renders them less reliable compared to the human brain, as the latter possesses greater logical reliability (Von Neumann, 1980).

1.3 Limitations of Serial Architecture:

"The brain compensates for its lack of logical depth by exploiting logical breadth, that is, massive parallel processing" (Von Neumann, 2012, as cited in Kurzweil, 2012, p. 40). This section compares parallelism versus serialism and asserts that the brain cannot respond with the same speed and precision as a machine with serial architecture, such as computers. Consequently, the logical breadth and distinction of human intelligence can surpass that of machines in detecting information.

One similarity identified by Von Neumann (1980) is the so-called prima facie digital nature of the nervous system, where he explains how dendrites conduct to neurons and compares nerve impulses from axons to binary markers in a digital system. In essence, Von Neumann (1980) acknowledges that computational architecture and the brain share some similarities, though they also differ substantially in precision, reliability, and speed.

As this type of technology continues to advance, there is no doubt that more specific regulations will be necessary for each case. The disparity between the rapid growth of these open-access technologies, such as Deepseek, ChatGPT, or Gemini, competing in an environment filled with diverse monetary, political, and social interests, combined with limited regulation and lack of knowledge for their detection, are factors that, according to Smith, must be improved alongside the obligation to protect the security and cybersecurity of nations (*The Guardian*, 2023). Sam Altman, CEO of the startup behind ChatGPT and the image generator DALL-E 2, aligns with Smith on the importance of creating licences and tests for the development of these technologies.

The impact of these technologies has led to the creation of the first comprehensive regulation on artificial intelligence. Regulation (EU) 2024/1689 of the European Parliament and of the Council, of 13 June 2024, establishing rules applicable to the field of artificial intelligence, applies to both public and private entities. It establishes a series of obligations and sanctions based on risk levels, classified into the following four categories:

Level 1: Unacceptable or prohibited, such as biometric manipulation of individuals, human behaviour manipulation with AI, social scoring, emotion recognition systems in the workplace, or those exploiting human vulnerabilities. This regulation aims to ensure the safety of individuals; however, certain experiments are permitted in controlled environments for research purposes.

Level 2: High risk, permitted with strict requirements, such as the processing of medical data and recruitment in the workplace.

Level 3: Limited risk, permitted with transparency; for example, companies must provide prior notice of the AI tools they will employ and must not violate copyright law.

In recent years, the use of deepfakes for identity impersonation has increased.

In 2019 alone, over 900 Facebook pages with fake photographs of American journalists were identified, using hyper-realistic AI technology to engage in disinformation tactics. The administrators of these pages were traced to Vietnam and linked to a far-right US media association, Epoch Media Group (Bhuiyan, 2023).

Technology giants such as Google, Meta, X, TikTok, and other companies have committed to the responsible use of AI. In 2018, this became a priority for these companies, with Google developing a programme for governance and ethical responsibility in the application of this technology. In 2022, following the pandemic, awareness of the importance of detecting false information increased significantly, leading to the implementation of information verification systems. For instance, technology leaders and academic experts collaborated to create the Deepfake Detection Challenge (DFDC) project, aimed at developing new methods for detecting deepfake videos. This research resulted in the creation of a public dataset of over 100,000 videos from 3,426 paid actors, produced using various GAN-based and face-swapping methods, making it the largest public dataset to date (Dolhansky et al., 2020).

In April 2023, Brad Smith, President of Microsoft, expressed significant concerns about artificial intelligence and deepfakes to the world.

Despite the efforts and planned investments, the landscape shifted in early 2025 when CEOs Elon Musk and Mark Zuckerberg announced the discontinuation of their information verification programmes, placing the responsibility on users to filter content. They justified this by stating their desire to respect freedom of expression and avoid errors in content moderation (Raya, 2025).

1.4 The Deepfake Phenomenon. Theories and Previous Studies

According to Giansiracusa (2021), the term *deepfake* originated in 2017 from an anonymous user who used this pseudonym as their nickname. With this fake profile, videos were created that swapped the faces of celebrities, such as Gal Gadot (star of the film *Wonder Woman*), using deep learning algorithms. This event marked a significant milestone in non-consensual pornography.

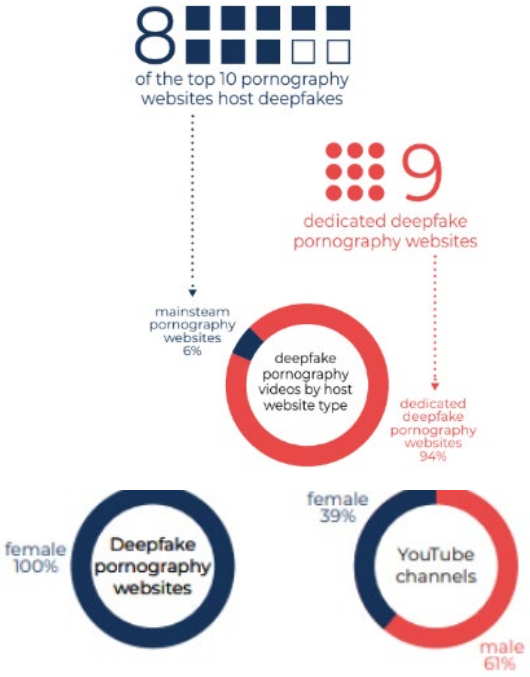
The author classifies different types of deepfakes based on their application: pornographic, entertainment, fashion, sports, business, and politics.

With technological advancements and the accessibility of numerous applications, initially for entertainment purposes, creating deepfakes no longer requires programming knowledge or any specific training. Notable open-access applications dedicated to creating these fake videos include FakeApp, DFaker, FaceSwap, FaceSwap-GAN, and others.

In recent years, research on this technology has increased; however, it remains a phenomenon in constant evolution, leaving several perspectives underexplored. According to researchers Vasist and Krishnan (2022), there are several gaps in deepfake narratives, including concerns about definition, limited theoretical grounding, and a lack of comprehensive demographic and geographical coverage in the existing literature on this topic. Their analysis was based on publications, affiliations, and geographical locations. The most prominent indices were led by the United States, with the majority of empirical studies employing a qualitative methodology.

While this phenomenon has primarily been viewed as a problem affecting public figures, such as politicians, actors, influencers, and others, technological advancements have become more accessible and user-friendly, leading to a rise in cases of ordinary individuals falling victim to identity impersonation. This is evidenced by a study conducted by Ajder et al. (2019), which found that, out of ten pornographic websites, at least eight contain deepfakes. This confirms that 94% of such content is pornographic, and its presence continues to grow across various platforms and websites, as shown in the following chart:

Figure 1. The rise of deepfake pornography



Source(s): Ajder et al., 2019

In a subsequent study, Ajder (2019) determined that 96% of deepfakes appearing on the internet consist of non-consensual pornography, using face-swapping technology. The increase continues to rise, as evidenced by a report issued in June 2019. This study classifies the technology according to its application and percentage of proliferation in web environments for the following purposes: pornographic, representing 96%; entertainment, 88.9%; fashion, 21%; sports; business/corporate, 4.1%; politics, 4% (Giansiracusa, 2021, p. 46).

The creation of platforms using AI technology can often generate models based on algorithms that perpetuate inequality, bias, and lead to harmful outcomes for various sectors of society. For this reason, this article proposes an integrative perspective that employs empirical analysis for detecting and interacting with deepfakes, while also noting the converging perspectives between different fields and disciplines of interest.

The accelerated advancement and open access of various applications that generate not only hyper-realistic fake images but also voice impersonation have made this technology a much more everyday concern, even affecting vulnerable populations such as young people and older adults.

Mass cases in which deepfakes are used have become increasingly common. In Badajoz, Spain, during September 2023, around twenty complaints were filed for these fake videos with nudes of underage girls, who, in addition to being victims of identity impersonation, were also bribed and harassed. The accused adolescents were 11 boys under 14 years old, three of whom operated the application that allowed them to create the videos, while the rest handled the dissemination on social networks and chats dedicated to this purpose (Pérez, 2023). Subsequently, a similar case in Seville has investigated five young people for disseminating AI-generated nude images of 20 underage girls (Aguiar, 2024).

In Ecuador, at a Catholic school, 24 underage students were victims of identity impersonation to create sexual content using AI. The incident was not reported to the prosecutor's office; however, the Ministry of Education stated that a protocol of action is being established for these cases, categorised as gender-based violence (Loaiza, 2023).

These cases highlight the importance of continuing to research tools appropriate to each social context that can serve for the detection of this type of content, as researcher Popova (2019) analyses through an ethnographic exploration of two websites dedicated to deepfakes. She distinguishes perspectives on privacy concerns for the person behind a hyper-realistic fake video. According to the study, compared to other communities that produce adult content, deepfake communities are less concerned with issues of privacy and authenticity of the person behind the image. Moreover, they attempt to keep the content within their groups.

2. Methodology

For the methodological development, a phenomenological analysis was employed to examine the evolution of deepfakes, alongside an experiment measuring the learning and detection capabilities of algorithms based on deep learning compared to human abilities. This approach objectively compares performance, using detection methodologies and prior experimental studies described in this section.

The applied technique involved a survey conducted with a target group in Quito, Ecuador, comprising a sample of 116 participants. This sample was selected based on the level of vulnerability, analysed through Michalos' (1985) Multiple Discrepancies Theory, which explains how comparisons can involve both real and empirical parameters, using up to five standards or categories of distinction.

In the first section, the level of prior knowledge about deepfakes was measured, alongside other contextual factors such as accessibility and limitations, using differentiated scales and Likert scales. In the second part, an experiment was conducted to assess the effectiveness of deepfake detection capabilities between artificial intelligence based on deep learning and the target population group.

The hypotheses are as follows:

H1. Generative artificial intelligence faces ethical and technological challenges and limitations.

H2. Generative AI has a broad spectrum of effectiveness in detecting fake images; however, its success rate may be lower compared to the detection abilities of an average person.

H3. Prior knowledge and digital literacy can enhance the detection of fake videos.

H4. Experiments and theories related to the main issue can generate new lines of research.

2.1. Objectives

The objectives of this article are as follows:

Analyse the limitations of generative AI as a mechanism for detection.

Analyse the effectiveness of deepfake detection capabilities between artificial intelligence and the human brain.

Determine which factors can enhance the detection of a deepfake.

Conduct a review of the literature and experiments related to this topic.

2.2. Detection Methodologies

Technological tools designed for image manipulation and detection of hyper-realistic videos have been analysed, alongside the state of the art and its prospective evolution, with technologies employing facial recognition models or systems for deepfake prevention, such as Deepware.

Regarding detection methodology, it is framed within the CNN classification systems described by Mirsky and Lee (2021) in their article *The Creation and Detection of Deepfakes: A Survey*. They classify specific detection methods, outlining seven types of mechanisms: combined spatial, environmental, temporal artefacts in behaviour, physiology, synchronisation, and coherence (p. 28). These authors support the theory that deep neural networks tend to perform better than traditional forensic image tools on compressed images. Furthermore, several authors have demonstrated how standard CNN architectures can effectively detect deepfake videos. Albahar and Almalki (2019), on the other hand, classify detection methods into facial recognition, multimedia forensics, watermarking, and Convolutional Neural Networks (CNNs). According to them, one of the simplest and most distinguishing methods for detecting fake news relates to the gaze in videos: firstly, the rhythm and speed of blinking. An average person blinks every 2–10 seconds, with each blink lasting a quarter to a tenth of a second. Deepfakes, on average, exhibit a slower blinking rate. Secondly, eye colour differentiation is a technique using computer vision, involving the analysis and extraction of colour.

In general, researchers have adopted one of two approaches: classification or anomaly detection. In this case, the anomaly detection approach is used, supported by evidence that, unlike a fully connected (dense) network, a convolutional neural network (CNN) learns hierarchies of patterns in data, making it significantly more efficient in handling images.

The challenges of detecting AI-generated images have led companies like Sensity to develop advanced real-time detection algorithms that analyse visual and contextual signals to identify AI-generated images. These algorithms can detect subtle artefacts, inconsistencies, and patterns indicative of AI manipulation, even in the absence of ground truth or reference data (Sensity, 2023). This pioneering company also presented an analysis revealing that facial recognition and digital identity systems are vulnerable to threats in 95% of cases, indicating that, despite advancements in detection algorithm development, both facial recognition and detection remain at a disadvantage compared to deepfake technology (Sensity, 2022, p. 3).

2.3. Previous Experimental Studies

A detection experiment by Groh et al. (2022) was used as a reference, comparing the detection capabilities of a tool based on a CNN algorithm with those of a group of people to assess detection performance. Two experiments were conducted. In the first, participants were shown a real and a fake version of the same video and asked to determine which was real and which was a deepfake without prior knowledge. In the second part, they were given the opportunity to change their response if desired after learning the detection tool's result. In both experiments, participants were also exposed to random emotional triggers before responding. The results showed that, in the first experiment, 82% of people outperformed the tool, which achieved a precision rate of 65%.

In a more recent study, Lake and Baroni (2023) determined that AI can easily match or surpass human learning abilities and even develop new capabilities. This was elucidated through a neural network developed using a technique they termed Meta-learning for Compositionality (MLC), which they claim can exceed the learning and detection capabilities of any other AI.

Another experiment highly relevant to this research is the "In Event of Moon Disaster" project by the Massachusetts Institute of Technology (MIT) (2020). This involves a deepfake video depicting President Nixon announcing that the Apollo 11 moon landing failed and the astronauts died. The

experiment's purpose is to test a large number of people's detection capabilities, allowing viewers to decide whether the video is real and analysing the factors influencing their responses.

3. Analysis and Results

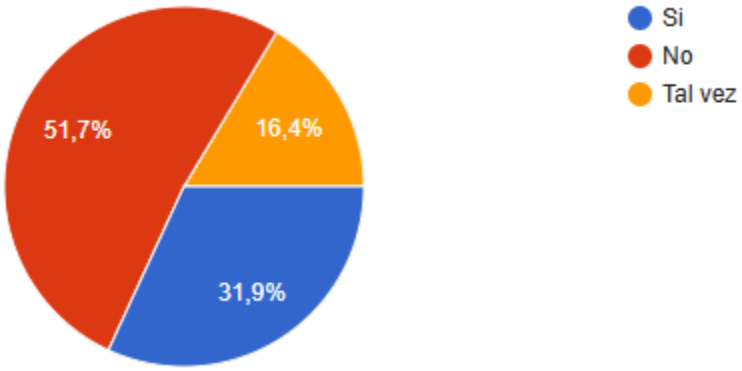
In the first phase of the survey, the data obtained were evaluated based on the following variables: Degree of deepfake recognition among young people. Accessibility of tools and applications for deepfake detection. Factors influencing the capacity to detect deepfakes. The percentage effectiveness of an algorithm compared to humans in detecting deepfakes.

The measurement techniques employed included the Likert scale, alongside a side-by-side matrix to assess the degree of influence of information factors on deepfake recognition.

In the second part of the survey, an experiment was conducted to analyse the deepfake detection capabilities of the target group compared to an AI based on deep learning.

The respondents indicated that 51.7% had no prior knowledge of deepfakes, compared to 31.9% who had heard about the topic.

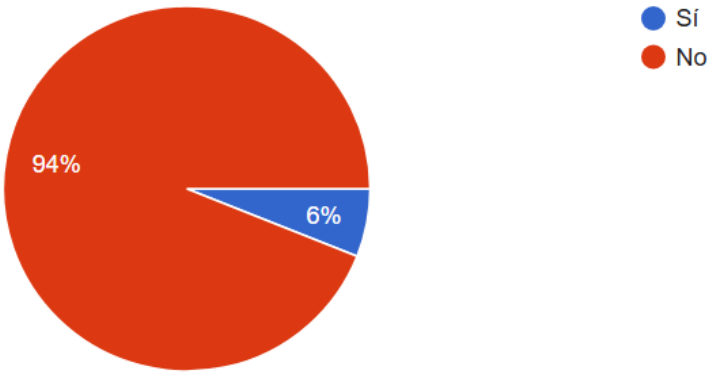
Figure 2. Previous knowledge



Source(s): Own elaboration, 2025.

To the question, "Have you received information on how to detect or prevent deepfakes in your immediate or academic environment?" only 6% of respondents affirmed they had, compared to 94% who responded that they had not.

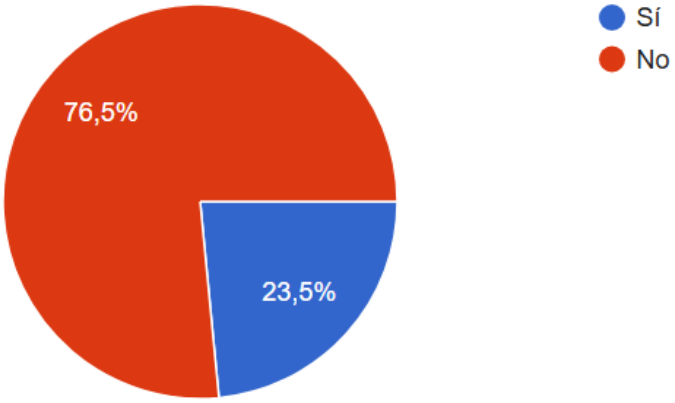
Figure 3. Self-perceived knowledge detection



Source(s): Own elaboration, 2025.

Regarding the accessibility variable, 76.5% of respondents stated they did not have access to information about detection tools.

Figure 4. Accessibility



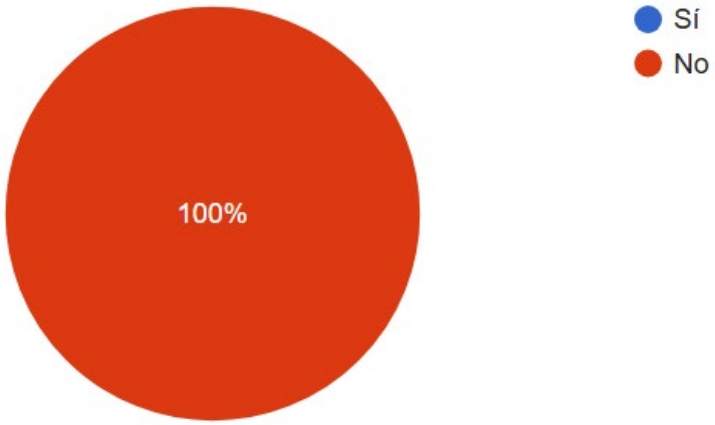
Source(s): Own elaboration, 2025.

For the respondents, the most influential factors were access to detection platforms, while sociocultural factors had the least impact. This question asked respondents to rate the degree of influence on deepfake detection capability on a scale from 1 to 5 (where 1 is the least influential and 5 is the most influential). The results were as follows:

5) Accessibility to detection platforms
4) Previous knowledge
3) Digital literacy
2) Information polarisation
1) Sociocultural factors

The level of concern revealed that 49% of respondents fear their privacy being affected by identity impersonation. Additionally, 100% stated they did not know how to respond if affected by a deepfake.

Figure 5. Action protocol

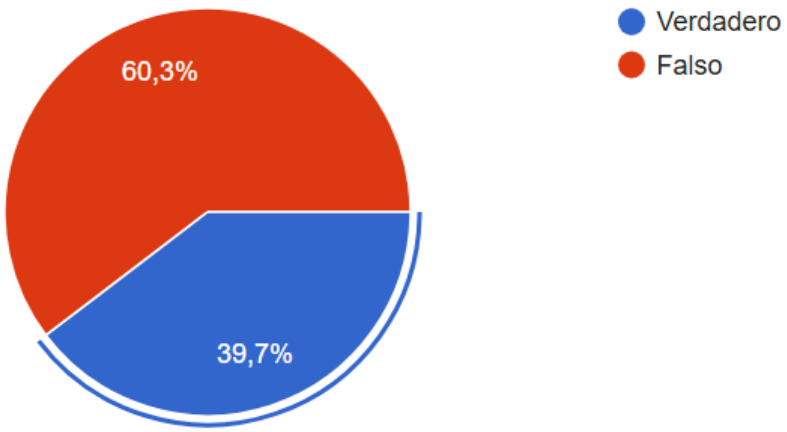


Source(s): Own elaboration, 2025.

In the second part, respondents were presented with a video to analyse and determine whether they considered it true or false, providing justification for their response. The video used was the same as in the "In Event of Moon Disaster" experiment, yielding the following results:

60.3% of respondents stated that the video was false, while 39.7% indicated it was true.

Figure 6. Experiment results



Source(s): Own elaboration, 2025.

In justifying their responses, the elements highlighted by respondents for assessing the veracity of the video were, in order of importance, the quality of audio and video, followed by the source or medium of dissemination, and the number of blinks.

When comparing the detection effectiveness of the Deepware tool, which utilises deep learning technology, the result indicated that the video was 92% likely to be false. Additionally, the tool identified the source, quality, resolution, and publication date with high precision, using only the provided URL:

Figure 7. Detection tool

Source: Deepware, 2025.

However, the result includes a disclaimer: “Deepware aims to provide an opinion on the scanned video and is not responsible for the result. As Deepware Scanner is still in beta version, the results should not be treated as absolute truth or evidence” (Deepware, 2025).

4. Conclusion

The results obtained regarding the limitations and strengths of generative AI as a mechanism for detecting deepfakes indicate that the most prominent strengths are arithmetic precision, data processing capacity, and speed. The processing of large volumes of images and videos, combined with recent advancements in deep learning, has enabled these tools to learn increasingly complex representations present in pixels, producing effective results in audiovisual recognition.

In the case of this technology, the neural networks used to execute its commands can resemble human capabilities in their ability to combine and create new systematic learning. However, AI’s capabilities depend on the complexity and bias of the programmed data. Natural language processing,

using mathematical and statistical methods, is fragile when identifying contextual, socioeconomic, and other abstractions, and it may fail in many cases when handling ambiguities due to a lack of deep message comprehension or a complete understanding of the scene, stemming from the absence of human-like common sense and logical depth. One way to improve the language model is to train it with more data, although limitations related to ethics, context, logic, and regulation persist, positioning human detection capabilities above those of AI.

The consensus of the deep neural network regarding identification was higher, achieving a 92% detection rate, while the detection capabilities of an average person reached 60.3%. It is noteworthy that in both cases, the results were correct, successfully identifying the video as false.

As stated in the disclaimer of the platform itself (Deepware, 2025), the result cannot be considered definitive, as the tool is still in development and continuous learning, which may lead to hallucinations or false positives or negatives.

Regarding self-perceived knowledge and identification of deepfakes, 31% of respondents' self-perception aligns with the percentage who reported having no previous knowledge. The hypotheses proposed that previous knowledge and digital literacy could enhance the identification of fake videos. However, the results indicated that the target group, despite lacking knowledge on the topic, still managed to identify the video as false, relying on two components: the video's quality and the source of the information.

Access to open-source or free platforms was one of the variables most highly rated by respondents, making it concerning that the population lacks access to information on detecting fake videos and protocols for responding to cases of identity impersonation. The cases of victims analysed in the current context highlight the rapidly expanding problem, underscoring the urgent need for accessible resources to facilitate the early identification of fake videos.

The growing issue of deepfakes, particularly in cases of non-consensual pornography and identity impersonation affecting vulnerable populations, highlights the urgency of establishing greater access to free resources and detection protocols.

Finally, it is confirmed that experiments and theories related to the main issue enable the continued development of research lines concerning the deepfake phenomenon, allowing for the analysis of factors that may enhance the detection of this malicious content. This article presents preliminary results for a doctoral thesis conducting a comparative analysis between two target groups in Ecuador and Spain to determine, through digital ethnography, factors that may facilitate the early detection of deepfakes.

5. Acknowledgements

Daniel Gómez González expresses gratitude for the funding provided by the Government of Spain through the R&D&I Plan of the project [PID2021-122905NB-C21].

Nathalie Alejandra Rodríguez Egas expresses gratitude to the Universidad Politécnica Salesiana of Ecuador, the Multimedia Design Programme, and its students for their participation in this research, as well as to her director of the international placement. Special thanks are extended to her thesis supervisors for their support and guidance and to the Universidad Complutense de Madrid.

References

Aguiar, A. R. (2024). La Guardia Civil detecta un nuevo escándalo de menores 'desnudadas' con IA en España: cinco jóvenes, investigados en Sevilla. *Business Insider*. Retrieved 12 04, 2024, from https://www.businessinsider.es

Ajder, H. (2019). Scams And Sabotage: Why Deepfakes Pose An Unprecedented Threat To Businesses. *Medium*. Retrieved October 6, 2023, from https://acortar.link/EHLBz8

Ajder, H. (2020). Tracer Newsletter 58 (09/07/20)-Deepfake Threat Intelligence: A statistics snapshot from June 2020. *Medium*. Retrieved October 6, 2023, from https://acortar.link/D6bA7d

Albahar, M., & Almalki, J. (2019). Deepfakes: Threats and countermeasures systematic review. *Journal of Theoretical and Applied Information Technology*, 97(22), 3242-3250.

Bhuiyan, J. (2023, May 16). OpenAI CEO calls for laws to mitigate 'risks of increasingly powerful' AI. *The Guardian*. https://lc.cx/vE8-g8

Dolhansky, B., Bitton, J., Pflaum, B. L., Howes, R., Wang, M., & Ferrer, C. C. (2020). *The deepfake detection challenge (dfdc) dataset*. arXiv preprint arXiv:2006.07397.

Eben, C. (2023, April). Ask a Techspert: What is generative AI? Google. https://lc.cx/yD6o_fGiansiracusa,

Fernández, M. (2024, Noviembre). *La teoría de la mente: un experimento probó que la IA tiene una capacidad humana que se creía imposible*. Infobae. https://goo.su/SXhUzau

Giansiracusa, N. (2021). *How Algorithms Create and Prevent Fake News: Exploring the Impacts of Social Media, Deepfakes, GPT-3, and More*. Apress. https://doi.org/10.1007/978-1-4842-7155-1

Google Cloud. (2025, Marzo). *¿Qué es la inteligencia artificial (IA)?* .https://lc.cx/0YFGOG

Groh, M., Epstein, Z., & Picard, R. (2022). Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, 119(1), e2110013119. https://doi.org/10.1073/pnas.2110013119

Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45). https://doi.org/10.1073/pnas.2405460121

Lake, B. M., & Baroni, M. (2023). Human-like systematic generalization through a meta-learning neural network. *Nature*, 623(7985), 115-121. http://dx.doi.org/10.1038/s41586-023-06668-3

Loaiza, Y. (2023, October 5). Estudiantes de un colegio de Quito utilizaron fotografías de sus compañeras para crear contenido sexual con inteligencia artificial. *Infobae*. Retrieved November 22, 2024, from https://lc.cx/nO2NO0

López de Mántaras, R., & Meseguer, P. (2017). *Inteligencia artificial.* CSIC Los Libros de la Catarata.

Massachusetts Institute of Technology. (2020). Tackling the misinformation epidemic with In Event of Moon Disaster. *MIT News Office*. https://lc.cx/tUb3GL

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. *AI magazine*, 27(4), 12-42.

Michalos, A. C. (1985). Multiple discrepancies theory (MDT). Social indicators research, 16, 347-413.

Mirsky, Y., & Lee, W. (2021). The Creation and Detection of Deepfakes: A Survey. *ACM Comput*, 54(1), 1-41. https://doi.org/10.1145/3425780

OpenAI. (2025). *Deepware* (versión del 07 de febrero) [Modelo multimodal grande]. https://scanner.deepware.ai/

Pérez, E. (2023, September 18). Falsos desnudos de menores generados por IA: la Policía investiga en Almendralejo el primer caso masivo en España. *Xataka*. Retrieved September 27, 2023, from https://lc.cx/HYGdfz

Plan de Recuperación, Transformación y Resiliencia. (2023, April 19). *Qué es la Inteligencia Artificial*. Retrieved March 4, 2025, from https://planderecuperacion.gob.es/noticias/que-es-inteligencia-artificial-ia-prtr

Popova, M. (2019). Reading out of context: Pornographic deepfakes, celebrity and intimacy. *Porn Studies*, (7). https://doi.org/10.1080/23268743.2019.1675090

Raya, A. (2025, January 21). Elon Musk y Mark Zuckerberg juegan a dos bandas: X y Meta mantendrán la lucha contra las noticias falsas en la UE. *El Español*. https://surl.li/jkukcq

Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) n.° 300/2008, (UE) n.° 167/2013, (UE) n.° 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828

(Reglamento de Inteligencia Artificial). Diario Oficial de la Unión Europea, de 13 de junio de 2024. http://data.europa.eu/eli/reg/2024/1689/oj

Russell, S. J., & Norvig, P. (2004). *Inteligencia artificial: un enfoque moderno*. Pearson Educación.

Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), 417–424. doi:10.1017/S0140525X00005756

Sensity. (2022). Deepfakes vs biometric KYC verification [Report]. *Sensity AI*. https://sensity.ai/reports/

Sensity. (2023, May 10). How to detect AI generated images with Sensity in 2023. Sensity AI. Retrieved April 2, 2024, from https://lc.cx/taJOpi

The Guardian. (2023, May 25). Deepfakes are biggest AI concern, says Microsoft president. The Guardian. https://lc.cx/inUT0b

Turing, A. M. (1980). Computing machinery and intelligence. *Creative Computing*, 6(1), 44-53.

Vasist, P., & Krishnan, S. (2022). Deepfakes: An Integrative Review of the Literature and an Agenda for Future Research. *Communications of the Association for Information Systems*, 51(556), 556-557. 10.17705/1CAIS.05126

Von Neumann, J. (1980). *El ordenador y el cerebro*. Antoni Bosch.

Von Neumann, J. (2012). *The computer & the brain* (R. Kurzweil, Foreword; 3ra ed.). Yale University Press. https://lc.cx/5WsZ-M

Woodruff, G., & Premack, D. (1979). Intentional communication in the chimpanzee: The development of deception. *Cognition*, 7(4), 333-362.