Reflection Article

Deepfakes as Communicative Acts: Ethical Challenges to Consent and Privacy*

Deepfakes como actos comunicativos: desafíos éticos sobre consentimiento y privacidad

Tomás Balmaceda[1], Laura F. Belli[2], Diana I. Pérez[3]

Abstract

This article examines the ethical implications of deepfake technology, focusing on the interplay between consent and communicative acts. Enabled by generative artificial intelligence, deepfakes blur the boundary between authentic and synthetic media, creating images and videos that appear real while depicting events that never occurred. Although these technologies can serve legitimate artistic, educational, or therapeutic purposes, their use in the production of non-consensual pornographic content poses serious ethical and social risks. Building on debates in the philosophy of language and technology ethics, the paper argues that sexually explicit deepfakes are not ethically neutral: their circulation constitutes communicative acts—such as insult, silencing, or humiliation—that inflict harm regardless of their factual veracity. Beyond individual autonomy, these practices endanger fundamental rights to privacy, image, and dignity, while also eroding public trust in audiovisual evidence. The analysis concludes by proposing an ethical framework that places informed consent, legitimate purpose, and respect for fundamental rights at the center of any responsible use of deepfake technology.

Keywords: pornographic deepfakes, generative artificial intelligence, derogatory language, responsible use of AI.

How to Cite

Balmaceda, T., Belli, L. F., & Pérez, D. I. (2026). Deepfakes as Communicative Acts: Ethical Challenges to Consent and Privacy. *Trilogía Ciencia Tecnología Sociedad, 18*(38), art. e3684. https://doi.org/10.22430/21457778.3684

* This paper was prepared as part of the Executive Unit Project "Artificial Intelligence, Scientific Studies of the Human Mind, Technology, and Ethics: Challenges for the 21st Century" from the Institute for Philosophical Research (IIF), funded by the National Scientific and Technical Research Council (CONICET) and the Argentine Society for Philosophical Analysis (SADAF).

[1] Universidad de San Andrés, Victoria, Argentina, tbalmaceda@udesa.edu.ar
[2] Universidad de Buenos Aires, Buenos Aires, Argentina, laurafbelli@gmail.com
[3] Universidad de Buenos Aires, Buenos Aires, Argentina, dperez@filo.uba.ar

Received: October 4, 2025 | **Accepted:** January 30, 2026

CC BY NC SA OPEN ACCESS

Resumen

Este artículo examina las implicaciones éticas de la tecnología *deepfake*, poniendo el foco en la relación entre consentimiento y actos comunicativos. Los *deepfakes*, posibilitados por la inteligencia artificial generativa, desdibujan la frontera entre lo auténtico y lo sintético, creando imágenes y videos que parecen reales aunque representen eventos que nunca ocurrieron. Si bien estas tecnologías pueden emplearse con fines artísticos, educativos o terapéuticos legítimos, su uso en la producción de contenidos pornográficos no consentidos expone riesgos éticos y sociales de gran envergadura. A partir de distintos debates en la filosofía del lenguaje y la ética de la tecnología, el trabajo sostiene que los *deepfakes* con contenido sexual no son éticamente neutrales, sino que su circulación constituye actos comunicativos (como insultar, silenciar o humillar) que generan daños, al margen de su veracidad fáctica. Más allá de las cuestiones de autonomía individual, estas prácticas ponen en peligro derechos fundamentales como la privacidad, la imagen y la dignidad, al mismo tiempo que erosionan la confianza pública en la evidencia audiovisual. El análisis concluye proponiendo un marco ético que ubica el consentimiento informado, la finalidad legítima y el respeto por los derechos fundamentales en el centro de cualquier uso responsable de los *deepfakes*.

Palabras clave: *deepfakes* pornográficos, inteligencia artificial generativa, lenguaje denigrante, uso responsable de la IA.

INTRODUCTION

In recent years, the ability to generate content using Artificial Intelligence (AI), whether in the form of text, images, or sound, has revolutionized Western societies. In this sense, deepfakes are one of the most sophisticated applications of the technology known as Generative Artificial Intelligence (GAI). Unlike usual images, audio, or videos representing some form of recording of real sounds or objects, GAI tools create digital artifacts with false data. The constant refinement of this technology is making it increasingly difficult to distinguish between AI-generated materials and authentic texts, images, sounds, or videos. The former includes photos of unreal people; paintings, music, or texts "in the style of" an artist; fabricated videos of public figures; and fake sex videos featuring individuals who never filmed such scenes. Instances of the latter are photos of real-world events; paintings, music, or texts created by actual artists; and videos of real subjects performing various actions they certainly did. Furthermore, greater accessibility to these tools, through increasingly affordable and user-friendly applications, has led to a surge in the creation of this type of representations. In combination with the internet and social media, which globally disseminate attention-grabbing content that quickly goes viral, GAI used for this purpose is transforming not only the digital environments but also the lives of millions of people.

Like all technology, GAI can be used for noble or ethically questionable purposes. De Ruiter (2021) cites among its potential beneficial applications the artificial regeneration of voices of patients who have lost their ability to speak due to diseases such as amyotrophic lateral sclerosis. This use improves their quality of life and restores some of their independence. One might also mention the preservation of deceased individuals' legacy, though there are many factors to take into account in this regard.

Finally, educational initiatives could implement these approaches, as demonstrated by the Dalí Museum in Florida, United States, which deploys a deepfake of the artist to guide visitors and create an immersive educational experience.

This article focuses on an ethically reprehensible use: the production of deepfakes with sexual content.[1] Accordingly, the ethical complexities posed by these technologies are examined through this use case, with the aim of extending the discussion to other applications and establishing ethically appropriate regulatory boundaries. To address this issue, Section 2 presents a brief overview of deepfakes, the technology underlying their production, and their effects on consumers. The discussion is divided into several topics to outline the ethical challenges posed by these outputs. Section 3 analyzes the process of generating this imagery and the consent required to use the materials that are employed for creating the simulation. Section 4 considers the various applications of these representations once they are created, emphasizing its communicative purposes. Section 5 reviews the rights that are affected or violated and proposes some guidelines for the proper deployment of deepfakes. These include whether the image or voice of the individuals involved is employed with their permission; whether the use is morally acceptable (for educational, artistic, or therapeutic purposes, clarifying that such media is a synthetic reproduction and not a true depiction); and whether any rights are infringed. These principles can be extended to all applications of deepfakes, regardless of its content.

WHAT ARE DEEPFAKES?

Deepfakes constitute a highly advanced application of GAI to create synthetic media for specific purposes. It entails the use of enhanced techniques to generate fake content—including images or sounds or a combination of both—with such a high level of realism that it can hardly be distinguished from original materials (Whittaker et al., 2020).

Its name comes from the fusion of "deep learning," a type of AI architecture, and "fake," making direct reference to the underlying technology that uses this type of algorithm to produce fabricated images, audio, or video. Thus, deepfakes do not capture the real world but reproduce content created through digital GAI-powered tools (Schick, 2020)—what is depicted is not reality but material generated from processed data.

The origins of the technology behind deepfakes dates back at least to 1997. In that year, computer scientists Christoph Bregler, Michele Covell, and Malcolm Slaney modified images from a video of a person speaking to make them appear as though they were pronouncing words from a different audio track. In the years that followed, various researchers continued to develop algorithms for facial tracking and matching. In the 3rd International Conference on Face and Gesture Recognition, the Active Appearance Model computer vision algorithm was presented. This approach enabled the alignment of a statistical model of an object's shape and appearance with a new image, making it the most significant precursor to this technology (Edwards et al., 1998). In 2016, the Face2Face platform was

[1] Approximately 96 % of online deepfakes are pornographic (Ajder et al., 2019).

created, which allowed the mouth area of a subject in a video to be swapped in real time with that of another person. A year later, it was one of the technologies used to create the video of Barack Obama insulting Donald Trump (Whittaker et al., 2020).

Today, this type of technology has become increasingly accessible and its results are sometimes very convincing. Thanks to open-source tools, powerful AI-trained models, and relatively inexpensive hardware, anyone with basic computer skills can now create hyper-realistic videos featuring altered faces or voices.

In technical terms, deepfakes are based on deep neural networks, particularly on models known as autoencoders and Generative Adversarial Networks (GANs). Autoencoders are used to encode an image (for example, a face) into a compressed representation that is then reconstructed with a "decoder." For face swapping—just as happened with the video of Obama allegedly insulting Trump (BBC News Mundo, 2017)—an autoencoder is trained to recognize and reconstruct the facial expressions of two different individuals. For their part, GANs work with two competing networks: a generator (which creates images) and a discriminator (which tries to distinguish the fake images from the real ones). As they compete, both improve, resulting in increasingly realistic videos, in which one person's face is superimposed over another's, mimicking gestures, glances, and even subtle expressions.

The data used to train these deep neural networks is key to their success. In the case of the Obama video, hundreds or thousands of images of his face were required, along with the video of the body onto which it was superimposed—in this case, that of an actor. This data is typically obtained from freely available videos on YouTube, social media, or imagery indexed by search engines like Google. Specialized software is then used to segment the faces, identify key points (eyes, nose, mouth), align the faces, and train the models to reproduce expressions, lighting, and angles. Training can take from a few hours to several days, depending on the desired quality and computational capacity.

With regard to hardware requirements, only a laptop with a modern graphics processing unit—such as those in the Nvidia RTX series—is necessary, as well as enough random-access memory. As for software, the most common programs and applications vary over time because the industry is constantly evolving. However, some of the most popular environments for creating deepfakes at the time of writing this article include DeepFaceLab, an open-source tool; FaceSwap, another open-source project with a more user-friendly interface; and Avatarify, which can create deepfakes in real time during video calls and when using applications.

The level of realism and the deceptive capacity of a deepfake depend on several factors. One key element is the quantity and quality of the training data—the greater the diversity of facial images, the more effectively the model can learn to reproduce natural expressions. Video resolution also plays a significant role, as lower-resolution source and output videos tend to obscure imperfections, making detection more difficult. In addition, lip synchronization is essential to ensure that facial movements align convincingly and that the face integrates seamlessly with the body and background—simply overlaying a face is insufficient. Finally, deepfakes require a careful matching of skin tones, lighting, shadows, and depth.

According to Lewis et al. (2023), humans have a very limited ability to identify deepfakes, even when explicitly informed of their presence. Without prior warning, most individuals fail to detect them. Meanwhile, when told that at least one out of five videos has been digitally manipulated, only 21.6 % are able to correctly identify it. In addition, their study report that nearly half of the respondents mistakenly identified at least one authentic video as fake, indicating a rise in skepticism toward audiovisual content. This is a phenomenon known as the "liar's dividend"—as people become aware of the existence of deepfakes, they become more likely to doubt the authenticity of any video, including genuine ones (Lewis et al., 2023, p. 2). Other studies indicate that factors such as familiarity with the subject depicted, self-confidence, or heavy use of social media contribute to inaccurate detection of deepfakes (Köbis et al., 2021).

Considering what has been presented so far, it is worth noting that deepfakes are barely distinguishable from photos and videos that document everyday situations and events. Therefore, a common initial reaction is to accept them as images and sounds that provide information about real-life situations and subjects. As Crippen (2023) argues, since deepfakes are seen as photographs, which have a causal relationship with the photographed object, the truth content of a deepfake is really strong. In other words, just seeing a deepfake as a photograph leads individuals to believe that what the image shows is exactly that. In this sense, deepfakes raise ethical issues regarding the (pseudo)knowledge they appear to provide (Kerner & Risse, 2021). If one doubts—due to the liar's dividend—one doubts all the images received because discerning whether a visual representation is fabricated or not becomes complicated. In short, with repeated exposure to this imagery, people become either too gullible or too distrustful. Thus, it seems that being able to distinguish deceit from veracity is one of the pillars of a reasonable life in society. With this technology, it appears that this ability is being called into question.

Turning to the topic that will be the focus of this paper, deepfakes with pornographic content became a hot topic starting in 2017, following allegations by several popular Hollywood actresses regarding audiovisual manipulation (Holliday, 2021). By the end of that year, the first wave of pornographic deepfakes went viral, and many of those videos were posted on various pornographic websites. One of the most well-known at the time was the video posted by a Reddit user named *u/deepfake*. This user uploaded a video to the *subreddit r/deepfakes* in which photographs of Gal Gadot (actress and star of the Wonder Woman franchise) were superimposed onto the body of a real-life pornographic actress. The video depicted the celebrity having sex with her stepbrother (Fletcher, 2018). This event sparked a global discussion about the ethical risks of deepfake technology, not only in terms of misinformation, but particularly because of its use to harm and exploit women.

Subsequently, there was a surge in videos in which this technology (especially face swapping) was employed to depict famous women in explicit sexual situations, exploiting their image without their consent. Actresses such as Scarlett Johansson and Emma Watson and singers such as Katy Perry were victims of these practices. In 2024, the musician Taylor Swift was the target of fake pornographic images that are believed to have originated from *4chan*—a website that allows users to post content anonymously—using Microsoft Designer. The material was then shared via a Telegram channel and

through various user accounts on the social media platform X. This case once again brought the issue of how this technology is used into the media spotlight (Goujard, 2024; Montgomery, 2024).

THE ISSUE OF CONSENT

Consent is a central issue in research on new technologies. In the age of social media and internet-based communication, people often agree to relinquishing both their rights and control over substantial amounts of personal information and imagery, including those of their close relations. It is difficult not to consent because in many situations some privacy has to be given up to be able to access desired or needed services, platforms, or websites. When accepting the terms and conditions, individuals also agree to share certain personal data.

Among the uses consented to when accepting the terms and conditions of the platform to which content is uploaded, there are some to which people would not otherwise allow. Given that the quantity and quality of training data are crucial for producing convincing deepfakes, the way in which such materials are obtained is of profound importance.

Leaving aside for the moment the controversies that may arise regarding why the creation of pornographic deepfakes can be objectionable (van der Nagel, 2020), this is a paradigmatic example in which consent—or the lack thereof—is particularly complex. The creation of pornographic material from non-sexual images, such as photographs or videos, violates the type of authorization that the subjects depicted gave when their image was made available for others to view (Umbach et al., 2024). Consent is one of the fundamental pillars that allows individuals to distinguish between lawful and unlawful sexual practices and productions and, in some cases, between morally permissible and impermissible. In recent years, discussions about the importance of consent regarding various sexual practices have taken center stage in public debate (Brigham et al., 2024), particularly in response to demands from feminist and LGBTQI+ activism and hashtag movements such as #MeToo or #NiUnaMenos.

There seems to be no disagreement that the use and distribution of sexual photographs or videos obtained without the authorization of the people being depicted—for example, if they are filmed by hidden cameras or from a distance—constitutes an illegal and a morally impermissible act. The same applies when such images are produced by individuals who appear in them but then are illegally obtained—for example, through theft or hacking. Now, why should not the same apply in the case of deepfakes? Just as the viewing or distribution of a sexual act is legal and permissible if the subject involved consents to it, the creation of sexual media should be illegal and impermissible if the individuals represented did not agree on it.

Why does the use of recognizable images of a person in a pornographic deepfake constitute a violation of consent? On the one hand, it violates the right to privacy. These images depict situations that are socially regarded as "private," whether they involve acts or body parts. On the other hand, this imagery could be considered a form of defamation, insofar as they convey a falsehood about the people who appear in them. In turn, in the case of pornography, they can be understood as a violation of the right to the non-alteration of social identity. This is based on the protection of the

individuals' ability to exercise autonomy over how social relationships with others are formed (Story & Jenkins, 2023), especially in terms of choosing to whom to reveal their naked body and sexual preferences. This type of visual representations also constitutes a form of violation of sexual privacy, including the right to identity, intimacy, and equality (Citron, 2019).

If an original photograph without sexual content is used to generate a deepfake and if that image is publicly available and was shared with the consent of the individual depicted, the distribution of the original does not constitute a violation of privacy or a form of defamation. However, why is the simulated image—which is not "real" nor shows a private act—considered a violation? Precisely because what underpins the requirement for consent in the case of photographs and videos is the identification of a person. If pornographic deepfakes depict recognizable subjects,[2] their authorization is also required for distribution. When that permission is lacking, as is exactly the case with these productions, the rights of the people concerned are violated.

The creation of sexual imagery that involves the use of someone's likeness or photographs raises a specific requirement for consent. As mentioned at the beginning of this section, authorization to use certain images is not sufficient to allow those visual representations to be deployed in all possible scenarios. Particularly, they should not be employed as inputs to generate deepfakes about the subject who gave consent.

One could argue that, in certain cases, the person who shared their images gave permission for them to be used in any way. In this respect, the standard of consent, as it appears in many terms and conditions, is heavily criticized (Lane et al., 2014). The reason behind is that it fails to meet almost any of the requirements identified as fundamental by theories of consent. Notably, the possibility to reject agreements is excessively limited because in many cases they are required to access essential services in society, from professional social media to applications for booking medical appointments or completing government paperwork. Additionally, they have serious shortcomings regarding the way information about the permitted uses is communicated. These deficiencies include texts written in technical or legal language that is difficult to understand or documents dozens of pages long whose scope of the individual's decision is hard to discern.

Furthermore, consent required in relation to issues such as sex or pornography does not function in the same way as other types of consent. Besides the reasons mentioned above, this is due to the taboo status of such practices in society and the importance of the right to privacy.

COMMUNICATIVE USES OF DEEPFAKES

Beyond the ethical issues involved in the process of creating deepfakes, which include the issue of consent for the use of data and images, it is necessary to consider how these images will be employed once they have been generated.

[2] In the case of pornographic deepfakes, whether involving public figures or the so-called "revenge porn" (which are often the most widespread forms of this content), "identifiability" is intentional.

Indeed, one might contemplate a strictly private use of deepfakes. In this case, the image would be like a drawing in a personal diary, created solely for the author's eyes. This use could hardly be regarded as unethical because it does not involve third parties. The case of imagery that sexualizes people would be similar to that of a sexual fantasy, which does not seem to be ethically reprehensible in any way.[3] However, it should be noted that unlike a drawing in a diary tucked away in a drawer in someone's home, deepfakes are stored in the cloud or on users' hard drives. Therefore, even if there is no intention to disclose them, their potential for disclosure through hacking or third-party access to these sites is much more likely than in the case of non-digital media.

In addition to strictly private uses (if such uses are possible), the dissemination of these deepfakes through various channels, generally via online platforms, websites, or social media, is more common and problematic. In this way, deepfakes become an act of communication and may be considered offensive, insofar as they elicit an emotional—and often automatic—reaction of rejection in the viewer (Baumgartner, 2018).[4] From this perspective, the offensive nature of a statement depends on individual psychological factors—that is, on the sensitivity that causes the recipient to react emotionally. At the same time, in cases of deepfakes that sexualize people, attribute statements to public figures, or depict attitudes that certain individuals would never adopt, the effect no longer depends solely on individual sensibilities. The offensive nature here stems from a degrading act rooted in social practices and values, independent of individual subjects. Indeed, if someone uses a derogatory term to refer to another person, the offensive nature does not depend on individual sensitivities but on structural realities of society (Anderson & Lepore, 2013). According to many specialists, such realities are encoded in the semantics of the term (Camp, 2013, 2018; Jeshion, 2013). Thus, although the content of this type of imagery is not linguistic, it is interesting to view these communicative acts as imagery in parallel with various types of speech acts that include derogatory terms. However, unlike pejorative terms, these are not words of a language but visual representations. Therefore, instead of considering the most widespread discussions in the field of applied philosophy of language concerning the semantics of such terms (Anderson & Lepore, 2024; Sosa, 2018), the focus could be on pragmatics. In other words, the major component shifts to the type of communicative act performed when publishing these deepfakes containing sexual or humiliating representations or depicting attitudes or actions in which a person would never engage.

Lepore and Stone (2018) argue that the pejorative nature of these speech acts is not semantically encoded in the terms used. On the contrary, it lies in the "tone" applied to deliver the utterance in question—that is, the way the illocutionary act is performed. For their part, Orlando and Saab (2019) consider this to be a distinct and unique type of speech act, which they label as "insulting." Langton et al. (2012) identify three distinct illocutionary forces related to hate speech: (i) propaganda, (ii) attack, and (iii) subordinating authoritarian speech. In the context of discussions regarding pornography as a form of hate speech, these communicative acts are also understood as acts of

[3] Öhman (2020) provides a detailed analysis of the similarities and differences between deepfakes sexualizing someone without their consent and sexual fantasies.
[4] Verrips (2018) identifies three types of offensive images: those related to (1) sex, (2) the sacred, and (3) death.

silencing (Langton, 1993).[5] Placing the discussion about the deployment of deepfakes in the context of the force of the communicative act, an additional ethical dimension can be introduced, given that actions such as insulting, attacking, or silencing are in themselves ethically reprehensible.

It might also be interesting to analyze the communicative acts involving deepfakes in the context of freedom of expression (Howard, 2024), a right that technology companies often invoke to favor the dissemination of all types of content. Many of the discussions in this area concern freedom of expression as it relates to political ideas and artistic expression. One of the central debates on this topic revolves around the tension between the right to freedom of expression and the duty not to cause harm. Therefore, the case of derogatory images could be examined through this lens—both in the political sphere and in cases related to ethnic, religious, or gender identity. Undoubtedly, derogatory images cause harm.

Rini and Cohen (2022) identify various ways in which deepfakes can harm their targets. One such harm involves novel forms of objectification, which undermine individual autonomy by placing persons in fabricated, non-consensual situations. A second form is illocutionary harm, whereby individuals are compelled to engage in speech acts to deny deepfakes attributed to them. A third form concerns existential trauma, arising from the destabilization of personal identity and memory through the circulation of fabricated content.

An additional problem is that the role of GAI companies, platforms, and dissemination networks is unclear, given that they allow such images to be generated and spread. Howard (2024) outlines two opposing viewpoints on this issue. On the one hand, these companies can be seen as private entities that have the right to self-regulate however they deem appropriate. In the case of social media, since participation is voluntary, one can argue that it is up to users whether to be exposed to this type of outputs. Social media would be limiting users' autonomy if it were to moderate content. Likewise, one could affirm that these platforms are merely the vehicle for the expression of the people who participate on them. Therefore, each subject is responsible for the ethically reprehensible acts they commit, absolving technology companies of all moral and legal responsibility.

On the other hand, one might claim that social media has become a public forum. Thus, the same laws that apply elsewhere should govern it in order to preserve individual rights, such as the right not to be harmed or the right to reply. To put it differently, social media should be subject to the laws in force in each country regarding defamatory acts. The challenge in this case is that the subject generating a deepfake, the individual posting it, the GAI platform enabling its generation, the company operating the platform where it was posted, and the person whose reputation was damaged might all belong to different countries. This is why AI and social media governance is not easily confined to national laws. Furthermore, it can also be argued that, given the recommendation algorithms employed by these platforms, users cannot choose not to be exposed to certain types of content they find offensive. Therefore, these platforms are violating their right not to see certain things and undermining their autonomy.

[5] Deepfakes sexualizing images primarily target women, which is the reason for this practice to be considered a specific form of sexual abuse (Tong, 2022).

THE RIGHTS THAT MUST BE PROTECTED

Pornographic deepfakes evidence two central issues in the overall deepfake discussion: that of consent and that of ethically reprehensible practices when understood as communicative acts. These problems must be extended beyond the case mentioned. Furthermore, there are other rights that may be violated in all kinds of deepfakes. In fact, due to their high degree of realism, they pose one of the most concerning threats to fundamental rights such as the right to one's image, the right to privacy, and the right to dignity.

The right to one's image is an expression of the right to personality and is linked to each individual's ability to control the dissemination of their own visual representation. In this sense, the image—understood as one of the primary attributes of a person, insofar as it reveals their unique characteristics—is also an essential component of the free development of personality. In other words, as de Ruiter (2021) suggests, a person's image is an inseparable part of their personality.

Traditionally, the protection of this right has been linked to the need for the consent of the individual depicted, but with the advent of this new technology, that requirement is being objected (Medon Affonso, 2021). Deepfakes pose a direct challenge to the protection of this right because they allow for the manipulation of media without the consent of the subject involved, often using images and videos whose deployment or distribution was initially consensual. As noted, this practice is particularly alarming when fake pornographic videos are created with the faces of real people. In this case, the violation of the right to one's image results in irreparable harm to the reputation and integrity of the affected person and their social circle (McNamara, 2023). However, not only is the individual sphere violated, but the veracity of communication in the public sphere is also put at risk.

The second section showed that the unethical use of deepfakes poses a threat to the right to privacy, a fundamental pillar of democratic societies. In the digital era, this right has taken on even greater significance, given the volume of personal information circulating on the internet and the ease with which it can be used without consent (Kugler & Pace, 2021). In the case of this technology, the threat to privacy arises from enabling the creation of false content that infringes on individuals' privacy and damages their reputation. This includes non-consensual pornography already discussed, the manipulation of information about their lives, fraud, or even identity theft. Such a misuse violates the right to privacy and has profound consequences for victims in their personal and professional lives, including their career prospects and personal relationships (Okolie, 2023).

The right to dignity, which is closely linked to protection against defamation, exploitation, and dehumanization, is a fundamental principle that is also threatened when AI-generated videos or images are used to humiliate, manipulate, or discredit people. In the personal sphere, the case of non-consensual pornography analyzed here is a clear example, as it causes profound moral and emotional harm, affecting the victim's self-esteem and psychological well-being (Gautam et al., 2024).

The development and proliferation of deepfakes pose unprecedented challenges to the protection of fundamental rights such as image, privacy, and dignity. Although technology itself is not inherently harmful, its malicious use has proven to have serious consequences for individuals and society as a whole. The problematic deployment of deepfakes goes beyond the creation of fake content to fulfill

sexual fantasies. It also enables tactics such as harassment, intimidation, and the silencing of people, creating a digital environment where fear and instability increasingly prevail. These actions can significantly hinder freedom of expression and inhibit participation in public discourse, particularly among vulnerable groups and individuals (Gautam et al., 2024). It is therefore crucial that society, technology companies, and governments work together to establish effective protection mechanisms. Such cooperation should take the form of a comprehensive and multidisciplinary approach aimed at preventing or mitigating the negative effects of these technologies, while ensuring respect for fundamental rights in the digital age.

CONCLUSION

This study delved into the phenomenon of deepfakes as one of the most complex and disruptive manifestations of GAI. It outlined the technology's evolution, from its earliest stages to its current ubiquity and accessibility, as well as how its increasing realism and availability have amplified its social and political impacts. Its most problematic uses were examined in detail, particularly in the context of non-consensual sexual content and political manipulation, highlighting how this technology can infringe upon fundamental rights such as privacy, image, and dignity.

Additionally, the study explored the technical mechanisms that enable the creation of deepfakes, the most widely used software, and the type of data required, revealing that the training of these models often relies on materials obtained without consent. The article also analyzed how, even when the generated content is not authentic, real harm can result—the mere suspicion that something has been manipulated is enough to undermine public trust, enabling phenomena such as the liar's dividend.

However, deepfakes are not inherently damaging. Their ethical evaluation depends on how they are employed and the context in which they circulate. Drawing on the analysis conducted, it is proposed that the ethical use of this technology requires compliance with three simultaneous conditions:

- Informed consent from the individuals involved for the use of their image or voice, the training of the models, and the distribution of the content.
- A legitimate and morally justified purpose, such as artistic creation, education, cultural experimentation, or therapeutic support—not harm, manipulation, or the exploitation of others. In this context, viewers should be warned that they are witnessing artificially generated material that does not correspond to real events. Accordingly, some type of indelible "watermark" may be incorporated into the AI-generated media to distinguish it from authentic recordings.
- Respect for people's fundamental rights, including their privacy, their integral identity, and their right not to be humiliated, misinformed, or coerced through synthetic fakes.

As noted in the introduction, these technologies have a wide range of applications, many of which bring enormous benefits to humanity. A prime example is the application of deepfakes for voice restoration in patients with neurodegenerative diseases or stroke-related damage, allowing them to reclaim their voices with their own tones and nuances, and even create new artistic works. Another possible ethical scenario is the recreation of historical figures in immersive educational, cultural, or artistic experiences, provided that the synthetic nature of the content is explicitly stated and the

memory and dignity of the person depicted are respected. In these situations, deepfakes can be a powerful tool for accessibility, education, or creative expression.

There are also cases of complex deepfake legitimacy. For instance, in the 2020 Delhi Assembly elections, the Bharatiya Janata Party, a Hindu nationalist party, deployed deepfakes of its leader, Manoj Tiwari, to reach more voters. Party members created AI videos in which Tiwari spoke in languages he does not speak fluently in real life, such as Haryanvi and English, and sent them to potential voters via WhatsApp (Jee, 2020). While the person involved gave his consent and the use was not derogatory and did not violate any rights, it is debatable whether it was clear to recipients that the video had been digitally altered. Therefore, the idea that this leader was a competent speaker of the recipients' language could be considered a belief induced by deception. Although one might think that the content—and not the language—of a statement is what matters in convincing voters, when it comes to issues concerning identities (ethical, religious, political, gender-related, etc.), language is critical.

For all these reasons, the ethical debate should not focus on whether to ban or allow deepfakes in the abstract. Instead, it should center on building regulatory, cultural, and philosophical frameworks that help discern when, how, and for what purposes their use can be compatible with a just, pluralistic, and democratic society. Otherwise, a tool with enormous transformative potential could risk becoming an instrument of control, humiliation, or mass manipulation. In the face of this threat, critical thinking, ethical commitment, and informed regulatory action are more necessary than ever.

CONFLICTS OF INTEREST

The authors declare that there are no financial, professional, or personal conflicts of interest that could have inappropriately influenced the results or the interpretations presented in this manuscript.

AUTHOR'S CONTRIBUTIONS

The three authors contributed equally to the completion of this paper.

REFERENCES

Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019). *The State of Deepfakes: Landscape, Threats, and Impact*. Deeptrace.

Anderson, L., & Lepore, E. (2013). Slurring Words. *Noûs*, *47*(1), 25-48. https://doi.org/10.1111/j.1468-0068.2010.00820.x

Anderson, L., & Lepore, E. (2024). *The Oxford Handbook of Applied Philosophy of Language*. Oxford University Press.

Baumgartner, C. (2018). Is There Such a Thing as an 'Offensive Picture'? In B. Meyer, C. Kruse & A. M. Korte (Eds.), *Taking Offense: Religion, Art, and Visual Culture in Plural Configurations* (pp. 317-339). Brill.

BBC News Mundo. (2017, July 7). *El falso Barack Obama creado con inteligencia artificial capaz de hablar como si fuera el original*. https://www.bbc.com/mundo/media-40632577

Brigham, N. G., Wei, M., Kohno, T., & Redmiles, E. M. (2024). "Violation of My Body:" Perceptions of AI-Generated Non-Consensual (Intimate) Imagery. In *Proceedings of the 20th Symposium on Usable Privacy and Security (SOUPS 2024)* (pp. 373-392). Usenix Association. https://www.usenix.org/system/files/soups2024_full_proceedings_interior.pdf

Camp, E. (2013). Slurring Perspectives. *Analytic Philosophy*, *54*(3), 330-349. https://doi.org/10.1111/phib.12022

Camp, E. (2018). Slurs as Dual-Act Expressions. In D. Sosa (Ed.), *Bad Words* (pp. 29-59). Oxford University Press.

Citron, D. K. (2019). Sexual Privacy. *Yale Law Journal*, *128*(7), 1870-1960. https://www.jstor.org/stable/45098020

Crippen, M. (2023). Conceptual and Moral Ambiguities of Deepfakes: A Decidedly Old Turn. *Synthese*, *202*(1), art. 26. https://doi.org/10.1007/s11229-023-04250-y

de Ruiter, A. (2021). The Distinct Wrong of Deepfakes. *Philosophy & Technology*, *34*(4), 1311-1332. https://doi.org/10.1007/s13347-021-00459-2

Edwards, G. J., Taylor, C. J., & Cootes, T. F. (1998). Interpreting Face Images Using Active Appearance Models. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition* (pp. 300-305). https://doi.org/10.1109/AFGR.1998.670965

Fletcher, J. (2018). Deepfakes, Artificial Intelligence, and Some Kind of Dystopia: The New Faces of Online Post-Fact Performance. *Theatre Journal*, *70*(4), 455-471. https://doi.org/10.1353/tj.2018.0097

Gautam, A., Joshi, R. K., Narula, A., & Sharma, N. (2024). Mitigating Human Rights Violations Caused by Deepfake Technology. *Library Progress International*, *44*(3), 4628-4637.

Goujard, C. (2024, February 6). *Taylor Swift deepfakes nudge EU to get real about AI*. Politico. https://www.politico.eu/article/europe-eye-fix-taylor-swift-nude-deepfake/

Holliday, C. (2021). Rewriting the Stars: Surface Tensions and Gender Troubles in the Online Media Production of Digital Deepfakes. *Convergence*, *27*(4), 899-918. https://doi.org/10.1177/13548565211029412

Howard, J. W. (2024, January 19). *Freedom of Speech*. Stanford Encyclopedia of Philosophy (Spring 2024 Edition). https://plato.stanford.edu/archives/spr2024/entries/freedom-speech/

Jee, C. (2020, February 19). *An Indian politician is using deepfake technology to win new voters*. MIT Technology Review. https://www.technologyreview.com/2020/02/19/868173/an-indian-politician-is-using-deepfakes-to-try-and-win-voters/

Jeshion, R. (2013). Expressivism and the Offensiveness of Slurs. *Philosophical Perspectives*, *27*(1), 231-259. https://doi.org/10.1111/phpe.12027

Kerner, C., & Risse, M. (2021). Beyond Porn and Discreditation: Epistemic Promises and Perils of Deepfake Technology in Digital Lifeworlds. *Moral Philosophy and Politics*, *8*(1), 81-108. https://doi.org/10.1515/mopp-2020-0024

Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. *iScience*, *24*(11), art. 103364. https://doi.org/10.1016/j.isci.2021.103364

Kugler, M. B., & Pace, C. (2021). Deepfake Privacy: Attitudes and Regulation. *Northwestern University Law Review*, *116*(3), 611-650. https://scholarlycommons.law.northwestern.edu/nulr/vol116/iss3/1

Lane, J., Stodden, V., Bender, S., & Nissenbaum, H. (Eds.). (2014). *Privacy, Big Data, and the Public Good: Frameworks for Engagement*. Cambridge University Press.

Langton, R. (1993). Speech Acts and Unspeakable Acts. *Philosophy & Public Affairs*, *22*(4), 293-330. http://www.jstor.org/stable/2265469

Langton, R., Haslanger, S., & Anderson, L. (2012). Language and Race. In G. Russell & D. G. Fara (Eds.), *Routledge Companion to Philosophy of Language* (pp. 753-767). Routledge.

Lepore, E., & Stone, M. (2018). Pejorative Tone. In D. Sosa (Ed.), *Bad Words. Philosophical Perspectives on Slurs* (pp. 132-154). Oxford University Press.

Lewis, A., Vu, P., Duch, R. M., & Chowdhury, A. (2023). Deepfake detection with and without content warnings. *Royal Society Open Science*, *10*(11), art. 231214. https://doi.org/10.1098/rsos.231214

McNamara, T. S. (2023). Is It Time to Recognize the Right to an Image? *Beijing Law Review*, *14*(4), 2109-2135. https://doi.org/10.4236/blr.2023.144117

Medon Affonso, F. J. (2021). O Direito à Imagem na Era das Deep Fakes. *Revista Brasileira de Direito Civil*, *27*(1), 251-277. https://rbdcivil.ibdcivil.org.br/rbdc/article/view/438

Montgomery, B. (2024, January 31). *Taylor Swift AI Images Prompt US Bill to Tackle Nonconsensual, Sexual Deepfakes*. The Guardian. https://www.theguardian.com/technology/2024/jan/30/taylor-swift-ai-deepfake-nonconsensual-sexual-images-bill

Öhman, C. (2020). Introducing the Pervert's Dilemma: A Contribution to the Critique of Deepfake Pornography. *Ethics and Information Technology*, *22*(2), 133-140. https://doi.org/10.1007/s10676-019-09522-1

Okolie, C. (2023). Artificial Intelligence-Altered Videos (Deepfakes), Image-Based Sexual Abuse, and Data Privacy Concerns. *Journal of International Women's Studies*, *25*(2), art. 11. https://vc.bridgew.edu/jiws/vol25/iss2/11

Orlando, E., & Saab, A. (2019). Términos peyorativos de grupo, estereotipos y actos de habla. *Crítica. Revista Hispanoamericana de Filosofía*, *51*(153), 31-58. https://doi.org/10.22201/iifs.18704905e.2019.1147

Rini, R., & Cohen, L. (2022). Deepfakes, Deep Harms. *Journal of Ethics and Social Philosophy*, *22*(2), 143-161. https://doi.org/10.26556/jesp.v22i2.1628

Schick, N. (2020). *Deepfakes: The Coming Infocalypse*. Hachette UK.

Sosa, D. (Ed.). (2018). *Bad Words. Philosophical Perspectives on Slurs*. Oxford University Press.

Story, D., & Jenkins, R. (2023). Deepfake Pornography and the Ethics of Non-Veridical Representations. *Philosophy & Technology*, *36*(3), art. 56. https://doi.org/10.1007/s13347-023-00657-0

Tong, S. (2022). "You Won't Believe What She Does!": An Examination into the Use of Pornographic Deepfakes as a Method of Sexual Abuse and the Legal Protections Available to Its Victims. *UNSW Law Journal Student*, *25*, 22-25. https://www5.austlii.edu.au/au/journals/UNSWLawJlStuS/2022/25.html

Umbach, R., Henry, N., Beard, G. F., & Berryessa, C. M. (2024). Non-Consensual Synthetic Intimate Imagery: Prevalence, Attitudes, and Knowledge in 10 Countries. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)* (art. 779). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642382

van der Nagel, E. (2020). Verifying Images: Deepfakes, Control, and Consent. *Porn Studies*, *7*(4), 424-429. https://doi.org/10.1080/23268743.2020.1741434

Verrips, J. (2018). A Brief Anatomy of Offensive Imagery. In B. Meyer, C. Kruse & A. M. Korte (Eds.), *Taking Offense: Religion, Art, and Visual Culture in Plural Configurations* (pp. 284-316). Brill.

Whittaker, L., Kietzmann, T. C., Kietzmann, J., & Dabirian, A. (2020). "All Around Me Are Synthetic Faces": The Mad World of AI-Generated Media. *IT Professional*, *22*(5), 90-99. https://doi.org/10.1109/MITP.2020.2985492