

NEW FORMS OF GENDER-BASED VIOLENCE: BETWEEN *DEEPREAL* AND *DEEFAKE*

NUEVAS FORMAS DE VIOLENCIA DE GÉNERO: ENTRE LA *DEEPREAL* Y LA *DEEFAKE*

Roberta Casagrande

Universidad de Mesina, Italia
ORCID: 0000-0003-2366-4304
robertaliacasagrande@gmail.com

Umberto Spaticchia

Universidad de Mesina, Italia
ORCID: 0009-0001-8785-7159
umbertospatichia@gmail.com



| Abstract |

This article constitutes an exploratory contribution to the understanding of the intersection between digital technologies and gender-based violence, with a particular focus on the emerging threats posed by *deepfake* technologies. It combines theoretical insights and empirical analysis based on data collected through a non-probabilistic opinion survey. The fundamental objective of this article is to analyze how *deepfake* and *deepreal* phenomena are perceived in the context of gender-based violence and to identify patterns in public awareness and attitudes toward these technologies. The methodological approach involves qualitative analysis supported by survey data, aimed at capturing trends rather than making generalizable conclusions. Through this approach, the article seeks to offer a critical discussion on the ethical, social, and cultural ramifications of *deepfake*-related gender violence, paving the way for future research and regulatory recommendations.

Keywords: Digitalization; AI; Artificial intelligence; Deepfake; Gender based violence; Reality.

| Resumen |

Este artículo constituye una contribución exploratoria para comprender la intersección entre las tecnologías digitales y la violencia de género, con un enfoque particular en las amenazas emergentes que plantean las tecnologías de *deepfake*. Combina perspectivas teóricas con análisis empírico basado en datos recogidos a través de una encuesta de opinión no probabilística. El objetivo fundamental de este artículo es analizar cómo se perciben los fenómenos de *deepfake* y *deepreal* en el contexto de la violencia de género

e identificar patrones en la conciencia pública y las actitudes hacia estas tecnologías. El enfoque metodológico implica un análisis cualitativo respaldado por datos de la encuesta, con el fin de capturar tendencias más que hacer conclusiones generalizables. A través de este enfoque, el artículo busca ofrecer una discusión crítica sobre las ramificaciones éticas, sociales y culturales de la violencia de género relacionada con los *deepfakes*, allanando el camino para futuras investigaciones y recomendaciones regulatorias.

Palabras clave: Digitalización; IA; Inteligencia artificial; Deepfake; Violencia de género; Realidad.

| Introduction |

In the era of digitalization, where the veracity of information is constantly challenged, deepfake becomes one of the most insidious threats. Distorted content, crafted using sophisticated *deep learning* techniques, stands as one of the most controversial technological developments of the past decade (Chesney & Citron, 2019). The advanced artificial intelligence (AI) techniques behind *deepfakes* use neural networks to encode the features of a face or voice, overlaying them onto another video or audio track (Goodfellow et al., 2020). While these techniques can have endless applications in fields like cinema or animation, it's clear they also have the capacity to produce deceptive and potentially defamatory content. Concretely, while the internet offers new spaces of freedom, such as the ability to create online communities and promote empowerment, which may reduce the gender gap, online violence against women takes on new and unprecedented characteristics. Indeed, in the new logic of the influencer system, the consumption patterns broadcasted influence the behaviors of young people, especially young women, who often find themselves navigating between two opposing realms: on one side, a culture of formal independence and self-determination, and on the other, dynamics that are still strongly male-dominated and misogynistic. The consequences can be numerous: blackmail, extortion, privacy breaches, gender violence, and loss of personal dignity (Waisbord, 2018).

In recent years, the evolution of digital technologies has led to a growing awareness of new challenges related to gender-based violence. In the context of this rapidly changing landscape, new and complex manifestations of violence are emerging, redefining the very concept of gender-based aggression. This article aims to explore the evolving phenomenon of new digital technologies, particularly focusing on the dynamics of *deepreal* and *deepfake*. Through in-depth analysis, this research is based on data collected through an opinion survey distributed to a non-probabilistic sample. The unit of analysis consists of individuals aged 20-40 from Europe, with data collected first-hand between November 23 and December 6, 2023. The sampling method is non-probabilistic based, employing a purposive sampling strategy to gather perspectives from individuals with varying familiarity with *deepfake* technologies. This approach allows for the observation of trends and preliminary conclusions rather than generalizable results. The survey aims to illuminate public perceptions of new forms of gender-based violence by examining personal experiences, awareness, and opinions regarding the *deepreal* and *deepfake* phenomena. Through this strategy, we aim to shed light on emerging patterns of gender-

based violence, paving the way for a critical and informed discussion on the social, cultural, and ethical implications of these increasingly pervasive phenomena. A detailed explanation of the unit of analysis and the sampling method, along with the type of conclusions that can be drawn from this methodological approach, is provided in the annex.

| Visual manipulation: the rise of the *deepfake* phenomenon |

For several years, in the dense context of discussion about digital-related risks, there have been emerging conversations about the potential dystopian scenario that could develop through the application of Artificial Intelligence (AI) technology (focused on the creation of *deepfakes*). We examined how excessive similarity (Mori, 1970) can become problematic when the subject is no longer able to distinguish what he or she is perceiving. Thanks to the evolution of new technologies and advances in artificial intelligence, we are now able to generate images and videos that look incredibly real, to the point of appearing authentic, but are actually nothing more than imitations and reproductions created from scratch. These are hyper-realistic images that may or may not have a reference model, such as photos of individuals artificially generated by artificial intelligence who do not physically exist, or they may use photos of real people who are extremely similar but completely artificially altered. In this context, we refer to *deepfakes*, or deep falsifications of reality and the individuals involved.

The term *deepfake* originates from the combination of two English expressions: *deep learning* (a set of techniques that enable artificial intelligence to learn to recognize shapes) and *fake* (false, fake news). This neologism refers to a human image creation technique that takes advantage of artificial intelligence. This technique involves combining pre-existing videos and images with other images through a complex algorithm using adversarial generative networks. The use of the term *deepfake* first registered in 2017, when an anonymous *Reddit* user, known by the pseudonym *deepfake*, posted several pornographic videos starring celebrities (Cole, 2017). At that time, the disclosure of these videos caused a scandal of considerable proportions. However, it was later discovered that these videos were not real but had been created by artificial intelligence. As a result, they were removed, but this incident was only one of several cases in which such a phenomenon occurred.

In 2017, an article published in *Vice*, a magazine popularizing *underground* culture, shed light for the first time on the *deepfake* phenomenon, identifying it as one of the rapidly spreading new media. This phenomenon was described as “worrisome” and “frightening” (Cole, 2017) as today it is increasingly easy to find images of people, famous and not, online, including through Google photos and selfies. As a result, everyone is increasingly vulnerable to seeing their image appear in a fake video, whether pornographic or not. The worrying situation lies in the fact that advanced programming skills are not required to create such videos, thus opening the door for anyone to produce *fake porn* content.

In light of this dramatic situation, there is still not enough public awareness, it reflects a very wide accessibility to the creation of these “real fakes of authorship” (Mento & Pira, 2023, p. 19). Communication channels, such as instant messaging apps, convey these materials at a frightening and disturbing rate. Francesco Pira, professor and sociologist at the University of Messina, writes in that book about how a simple bot created for the instant messaging platform Telegram gave the possibility to “completely undress a woman in a virtual way” (ibidem). Needless to say, material like this can generate extreme backlash: blackmail, invasion of privacy, *revenge porn*, extortion for financial gain, and loss of dignity.

Thanks to the use of the complex methodology of *deep learning* and specific neural networks called GANs (Generative Adversarial Networks), it is now possible to use images of an individual's face and samples of his or her voice to produce incredibly realistic videos in which the person depicted utters sentences that are never actually uttered. However, nowadays, the performance of such tools has not yet reached sufficient levels to completely conceal the artificiality of such manipulations. These audiovisual productions currently find their place in the field of comic or grotesque content, eliciting smiles at the hypothetical effect of someone expressing outlandish and inappropriate concepts, situations that under normal conditions would never occur. To understand how *deepfake* videos and techniques work today, it is crucial to first introduce two key concepts: *deep learning*, previously mentioned, and *neural networks*. *Deepfakes* are based on neural networks presented in 2014 in the paper “Generative adversarial networks” by Ian Goodfellow et al. (2020 [2014]), a U.S. computer scientist and researcher, along with other scholars at the University of Montreal. These networks are known as GANs (Generative Adversarial Networks)¹. We can define them as artificial intelligence algorithms designed to solve the generative model problem. The goal of the generative model is to study a set of training examples and learn the probability of distribution that generated them. In the worst-case scenario, some stars in the entertainment industry have had their faces and voices mounted on the bodies of the protagonists of pornographic videos to stimulate the imagination of an audience aware of fiction, but eager to push beyond the regimes of the visible codified by the entertainment industry.

We identified several possible negative implications related to the abuse of *deepfakes*, including their use in political contexts and in the aforementioned fake porn. In addition, we introduced some of the methodologies through which this phenomenon has emerged and continued to develop. Some academics express deep concerns about the rise of these new methods of mystifying reality. According to Schwartz (2018), Hany Farid, a professor at the University of California Berkeley, stresses the need for researchers to be more cautious in the issues they address and disseminate, both socially and politically.

¹ GANs are able to generate additional examples based on the estimated probability distribution. Goodfellow et al. (2020 [2014]) also argue that generative models based on deep learning are widespread, but GANs stand out as among the most successful generative models, especially in their ability to generate realistic high-resolution images. The process involves feeding footage of two individuals into a deep learning algorithm that can swap faces through the face swapping technique. In other words, deepfakes exploit facial mapping technology and artificial intelligence to swap the faces of people in a video in an extraordinarily realistic way.

In addition, Farid, in recent years, has devoted efforts to creating detection methods to identify and, if necessary, counteract *deepfakes*. Despite the challenge of developing effective software in this area, the increase in the prevalence and, in some cases, the danger associated with *deepfakes* has spurred research to create new detection methods.

The association between the sign and the real becomes not only arbitrary but also deeply influential, as the signs acquire a self-sufficient existence, expressing nothing beyond themselves. This change in dynamics challenges our fundamental understanding of representation and authenticity.

In this hyperreal environment, signs exert significant power, dictating the terms to which reality must conform. The arbitrary nature of this association requires a reassessment of our perception and engagement with the world around us. The signs, which manifest themselves as images, videos or information, work with an independent authority, often overshadowing the very reality they aim to represent.

Baudrillard's notion of *hyperreality* (1981) comes to the fore, suggesting that we are living in a world where the simulated version of reality and an idealized humanity take precedence over the original. This raises deep questions about the nature of our existence, as hyperreality becomes a lens through which we interpret and engage with the world.

The era of virtual simulation challenges us to navigate a landscape where the boundaries between the real and the simulated blur. The improvement and enhancement of reality through these simulations raise complex considerations on the authenticity of our experiences. As we face the impact of hyperreality on our perceptions, it becomes imperative to critically examine how we can maintain a sense of authenticity in a world where the virtual often eclipses the real. The era of virtual simulation challenges us to navigate a landscape where the boundaries between the real and the simulated blur are drawn. The refinement and enhancement of reality through these simulations raises complex considerations about the authenticity of our experiences. As we face the impact of hyperreality on our perceptions, the community seems to be perpetuating negative behaviors towards what should be a conscious and ethical use of AI.

New uses and new disturbing stories, such as that of the artificial influencer Emily Pellegrini, are peeping on the horizon, creating worrying implications on the use of AI in modern society.

| Reality as material for *deepfakes*: focus on Emily Pellegrini |

Looking forward, what will be the impact when artificial intelligence, *deepfake*, and generated content merge, becoming a powerful catalyst that can shape a completely illusory view of reality? This was already happening right between the end of 2023 and the beginning of 2024, which saw the arrival on Instagram of a fascinating influencer completely generated through the use of artificial intelligence. Emily Pellegrini's

Instagram profile boasts over 130 thousand followers, with numerous companies that have identified him as a testimonial for their products.

However, it should be emphasized that Emily Pellegrini had no tangible reality; she was a creation of artificial intelligence. The advent and progressive integration of advanced technologies such as artificial intelligence and *deepfake* have catapulted the world of communication and entertainment into a fascinating and, at the same time, disturbing territory. The evolution of these technologies, combined with artificially generated content, is shaping a new dimension in which the perception of reality collides with illusion in an intricate game of digital mirrors. The case of Emily Pellegrini, an influencer figure invented by artificial intelligence, is a paradigmatic example of how these technologies have reached a level that affects not only the online experience, but also the daily life of individuals and even the business world. This phenomenon has led us to question the ethical, social and economic implications of a reality increasingly shaped by the fusion of AI, *deepfake* and generated content.

Let's start by looking at the current state of art. The story of Emily Pellegrini, which has gained popularity on social platforms, highlights how AI can create completely fictional characters that interact with online audiences. These characters, thanks to advanced techniques of text and image generation, can arouse empathy and involvement by users, challenging our perception of authenticity. The ethical implications behind using AI to create fake influencers are considerable. The thin line between reality and fiction is increasingly blurring, raising fundamental questions about trust in the digital world. Users can be deceived, investing emotions and resources in interactions with characters that do not exist in the real world. Manipulation of public opinion, fueled by the ability of AI to create compelling narratives, poses significant challenges for democracy and public debate.

However, the dark side of *deepfake* emerges when you consider manipulating information and spreading false news. AI can be used to create seemingly authentic video and audio of public figures making speeches or doing actions that never happened. This type of manipulation can have devastating consequences on the reputation of individuals and institutions, undermining trust in journalism and the truthfulness of information.

Another crucial aspect is privacy. AI, powered by huge datasets, can generate increasingly compelling *deepfakes* using real people's faces and voices without their consent. This privacy breach raises serious concerns about the security of personal information and the possibility of malicious use of such technologies to harm the lives of ordinary individuals.

To address these challenges, a broad and inclusive debate on how to regulate the use of AI and *deepfake* is essential. Regulations must adapt quickly to keep pace with technological evolution and protect society from potential threats and abuse. At the same time, it is crucial to educate the public about the presence and impact of these technologies, promoting an informed digital citizenship.

According to Vacca (2015) Jean Baudrillard said to Claude Thibaud that “The system of representation is in question. Man’s self-image is virtualized. You are no longer in front of the mirror, you are inside the screen, which is quite another thing.” There is still talk of *hyperreality*, a concept coined by Jean Baudrillard, which suggests reality and the representation of reality are increasingly confused, leading to a nuance between what is authentic and what is simulated.

In the current context, where artificially generated influencers like Emily Pellegrini can accumulate a significant following on social platforms and interact with the public, the barrier between the true and the false becomes increasingly tenuous. Hyperreality emerges when digitally generated experiences can induce emotions and reactions similar to events or real people. The phenomenon of *deepfake*, in particular, raises questions about the credibility of visual and auditory information. With the ability to create compelling video and audio with subjects that don’t exist in reality, we are faced with an advanced form of simulation that challenges our traditional perception of truth. In this context, hyperreality manifests itself when artificial representations become so sophisticated that they are difficult to distinguish from actual reality.

The concept of hyperreality becomes even more relevant when we consider how people interact with digital platforms and social media. Online experiences are increasingly shaped by algorithms that filter and customize content based on our behaviors and preferences. This creates a personalized reality, a customized digital universe that can differ greatly from that of others. The statement that reality disappears and is replaced by an anti-reality, a hyperreality, puts at the center of the reflection the radical transformation that is going through our relationship with what we consider real. The accelerated evolution of digital technologies, the spread of artificial intelligence and the creation of generated content contribute to a landscape in which the line between authenticity and simulation is increasingly blurred.

Hyperreality, in the context described, emerges as a form of existence in which reality itself is supplanted by its distorted counterpart, an exaggerated and altered representation. This anti-reality, fueled by the illusion and counterattack of simulacra, leads us to question the very nature of perception and truth (Baudrillard, 1981).

Virtual simulation, which takes on the role of a perfected copy of reality and humanity, raises deep questions about the meaning of authenticity. We live in an age in which reality seems to bend to the rules of algorithms, the dictates of perfect images and the needs of digital storytelling. In this perspective, real reality is often hidden behind a curtain of artificially created appearances. The relationship between the sign and the real thing becomes increasingly complex. In a hyperreal world, the association between symbol and reality becomes arbitrary. Signs, be they images, videos or information, acquire their own autonomy, divorcing themselves from the need to reflect the surrounding reality. Reality itself is forced to conform to these signs, becoming a representation influenced and distorted by their dominant presence.

The concept of *real thing* crumbles, as the deceptive simulacrum takes on the predominant role. Virtual reality becomes the norm, and our daily perception is shaped by digital narratives, filtered images and simulations of human experiences. This raises crucial questions about the genuineness of our interactions, the meaning of our experiences, and the possibility of distinguishing between what is authentic and what is built. In trying to understand this new phase of virtual simulation, it is essential to consider the impact on different spheres of society. From the cultural to the economic sphere, hyperreality permeates every aspect of our existence. The dynamics of power, social relations and the very concept of truth are redefined in this context, triggering a radical redefinition of the social fabric.

The anti-reality, in which we find ourselves, requires a critical reflection and a growing awareness. As individuals and society, we must explore how we can preserve our authenticity and our understanding of reality in a world where hyperreality seems to prevail. The creation of an informed digital citizenship, the promotion of transparency in online interactions and the search for new forms of authentic expression are key elements in this challenge.

This particular (and sometimes tragic) moment of the era of hyperreality pushes us to deeply re-evaluate our relationship with reality, challenging conventional notions of truth and authenticity. Through a critical exploration of how digital technologies shape our perception, we can develop strategies to navigate this evolving landscape and preserve our connection to what is authentically human.

The main concern is that society nowadays tends to believe *fake news* or *deepfakes* more easily, even when their verisimilitude is lacking (as demonstrated by the intentionally low-quality fake Trump video in Belgium). We are facing a framework of misinformation that can lead to two possible outcomes: either society will react by believing anything, or an opposite trend will develop, namely, to stop believing everything. The main threat is not only that people will be deceived, but that they will come to regard any information as deception, resulting in distrust of the media and institutions. This could gradually lead to nullification and complete epistemic decay.

| *Deepfakes* as catalysts for gender-based violence |

While online offers new spaces of freedom, such as the possibility of creating community and promoting empowerment, which potentially narrow the gender gap, online violence against women presents new and unprecedented characteristics (Öhman, 2020). In this area, law appears crystallized and unable to provide adequate responses.

The digital dimension of violence against women covers a wide range of acts perpetrated online or through technological tools, which are part of the continuum of acts suffered because of their gender, including in the domestic sphere. However, none of the definitions devised so far has the capacity to cover the full range of phenomena and contain the wealth of meanings of such a complex, elusive and changing subject.

The sketch of people who transform on the web consists of individuals who are normal in real life, but who engage in different behaviors online. Among the least dangerous are trolls, those who find pleasure in insinuating dissent and attacking ideas or people by launching provocative comments. Their goal is to get the victim's attention and spark a heated debate. However, the most dangerous are the so-called five-star haters. These individuals not only irritate or offend but also aim to unleash the baser instincts of their interlocutors: their behavior can generate hatred, racism, misogyny, and discrimination and create harmful tensions for society as a whole.

Suffice it to say that Amnesty International's Hate Barometer, when the topic is “women and gender rights”, reports how one in three online comments generates hate speech and sexist speech: in Italy, the Vox Diritti Observatory reported that 63.1 percent of hate tweets in 2019 were directed at women (Amnesty, 2021).

Further poisoning virtual communities is revenge porn, typically defined as the nonconsensual distribution of private, sexual images by an ex-partner; however, it also extends to strangers acting with different purposes, such as exploitation, extortion, humiliation, and mockery. Today we observe that behaviors that in the past were considered marginal have become widespread at a widespread level, involving not only young people but also adults. This phenomenon generates fears and increasingly violent behavior. If we analyze surveys conducted over time on the world of adolescents, a growing awareness emerges, but also a feeling of fear that is accompanied by behavior patterns that tend toward extremes, weak relationships and, in some cases, illicit actions.

In 2020, Save the Children conducted a survey showing that 39% of boys and girls in Italy are exposed online to content that condones violence against women, with the range widening from 31% of boys to 48% of girls. Among girls, 41% have seen content posted by their social contacts that made them feel offended and/or humiliated as a woman, and of these, 10% felt most exposed during lockdown².

Although the phenomenon is widely disseminated because of its implications in the social and political spheres, a worrisome and lesser-attended perspective is the possibility of its being a catalyst for gender-based violence. We could call it technological abuse in that it poses a potential threat to women's identity and safety, resulting in psychological and social consequences, according to our own research (Annex: Pictures 1 and 2).

The first distorted use of deepfake technology dates back to 2017, when pornographic images and videos portraying celebrities such as Emma Watson and Taylor Swift were shared on the Reddit platform, clearly made without their knowledge: now, however, it should be considered a common problem, in fact anyone can now become a victim of it. This leads to a reflection on how obvious is what Pecchinenda (2017, p. 6) argues:

We will not necessarily find a *correspondence* between the representation (the

² <https://www.savethechildren.it/press/violenza-contro-le-donne-il-70-delle-ragazze-dichiara-di-aver-subito-molestie-e-apprezzamenti>

digital-image) of a given material object and the object itself. That is, an image (and the space-time of experience to which it refers) begins to assume a status absolutely independent of the physical reality connected with traditional space-time categories.

From the criminological point of view, *deepfake* can be considered as a 'synthetic alternative to post-relational revenge porn, extortion mode, as indirect blackmail, sexual harassment, as an additional tactic for cyberstalking or as a real integration into the child pornography market or as a submarket other than it.

A further sore point is the private and submerged market of these *deepPorns* in which one only needs to collect a few high-quality images of the unwitting victim in order to be able to purchase at an average cost of \$50 per minute, perfect deepfakes videos customized so that one can make one's sexual fantasies, including sadistic-violent ones, concrete. We see, on the one hand, the rise of a new phenomenon in Porn that generically concerns celebrities. And, at the same time, we observe the continued development of low-cost online services and Apps for *deepNude* that include the possibility of face swapping i.e., the application of one person's face on another's body simply with one's cell phone, which sees the emergence of mobile *deepFaking* as a phenomenon of proximity, not only in cyber-social terms, but in terms of physical proximity of the target audience, which sees all women, in their daily lives, as potential victims.

| Protecting ourselves from *deepfakes* |

Is there a way to protect oneself? Surely that of hiding one's image of the network: delete photos on Instagram, on LinkedIn profile (with its consequences), Facebook profile but even this would not be enough, photos shared by others, containing our face would be enough for the manipulator to use it at will. A study by Citron (2019), analyzed how *deepfakes* can be used for the creation of totally fake videos that turn out to be, at the same time, extremely realistic of women who have never performed certain actions, often they are women of a certain popularity and produced explicitly to be shared online or sold to collectors, and it comes to mind how this is to be understood as the most symbolic example of the objectification of the body. The female body is thus reduced to a mere digital resource and the manipulation given by artificial intelligence is a tool to sexualize it to be disposed of for male pleasure and consumption.

DeepNude represents a clear violation of what Citron (2019) refers to as “sexual privacy” or the behavior, expectations, and choices people have about their bodies, intimacy, and sexual activity. This kind of manipulation goes far beyond simple defamation or traditional cyberbullying, as it erodes the credibility and integrity of victims and also undermines their psychological sphere.

According to Paris and Donovan (2019, p. 26) “The Internet at once facilitates access to pornographic materials made with and without consent, and provides multiple unchecked and unregulated environs that “normalize misogyny, hurtful sexuality, racism,

and even seeking revenge on female ex-partners.” The debate and legislation over this phenomenon continue today and sets the background for most legal discussion around *deepfake* pornography.”

The authors pointed out that although *deepfake* technology is not limited to the realm of porn, it presents 96 percent of content attributable to sexually explicit images, which contributes to the perpetuation and normalization of sexual and gender-based violence.

Research published by Flynn et al. (2022) in the *British Journal of Criminology* investigated the abuse of manipulated content, including *deepfakes*, as a new form of image-based sexual assault (IBSA). This study found that this abuse includes a variety of victims and perpetrators, underscoring a poly- victimization that includes vulnerable groups such as women, youth, and racial minorities.

The consequences can be manifold, one finds, for example, the removal of the victim from the family unit, the community in which he or she lives or from his or her workplace thus resulting in the phenomenon of victim blaming.

Instead of focusing on the culprit and the underlying causes of non-consensual sharing of intimate or manipulated material, it very often happens that we focus on the victim and the victim's faults when it is clear that the person who decides to carry out these acts was not actually aware that he or she was holding his or her dignity, reputation, career, and very life in his or her hands.

The concept of victim blaming, introduced by psychologist Ryan (1971), indicates the attitude of justifying inequality by finding fault with the victims themselves. In the 1980s, this attitude was also examined in the context of gender-based violence. It is a cognitive bias that manifests itself in the form of a shortcut devised by the human mind to reassure and defend percipients from sociocultural phenomena that are difficult to understand and accept. This cognitive bias is highly individualistic and can be analyzed from a psychological perspective.

According to contemporary literature, the human tendency to blame the victim can be explained by the *just-world theory* (Lerner, 1980). This theory is based on the dangerous and fallacious assumption that “everyone gets what he deserves and deserves what he gets.” Proponents of the theory are inclined to perceive social reality as naturally just and to believe that negative events happen only to those who, because of their careless decisions or actions or actions that do not conform to social expectations, deserve punishment.

But while blame continues to be placed on victims, the number of *deepfake* content continues to soar and the impact on them can be long-lasting. Clearly on gender-based online harassment, with huge influence on women's freedom of speech (Maddocks, 2018).

A recent journalist report (Burgess, 2023) pointed out that several Twitch streamers who were victims of *deepfake* told of feeling violated, of being exposed to further harassment, others reported the shame they felt as the non-consensual content also reached their family members' devices. According to the same source, C. Flynn, a professor at Monash University, finds “high rates” of psychological disorders such as

anxiety, depression, self-harm or suicide following digital abuse, and concludes that “The potential impact on a person's mental and physical health, as well as on their work, family and social life, can be immense regardless of whether the image is *deepfaked* or real.” The interdependencies with phenomena including red pilling, Incel (pseudo-) ideology, violent extremism, conspiracy, memetic violence, and more generally online hate, prevention, and anticipation of gender-based violence are a priority issue today.

According to recent research, Burgess (2023) echoes that the pervasiveness of *deepfake* porn is increasingly evident. In the past seven years, at least 244,625 videos have been uploaded to the 35 websites created exclusively or partially to host *deepfake* porn videos. In the first nine months of 2023, 113,000 videos were uploaded to these portals, a 54 percent increase from the 73,000 videos added in all of 2022. Thus, we can consider that the number of videos produced in 2023 ultimately exceeded that of all other years combined. Figures that highlight just how crucial the problem of nonconsensual *deepfakes* has become—the severity of the problem is much deeper and includes other types of manipulated images. Further research focuses on 35 different websites that host pornographic *deepfake* videos exclusively or that embed the videos along with other adult material and thus does not examine videos posted on social media, those shared privately or manipulated photos. Most of these sites clearly indicate their *deepfake* porn content, often including the term or its derivatives in the name. The first two portals contain 44 thousand videos each, while five others host more than 10 thousand *deepfakes*. Most of the portals have several thousand videos, while some list only a few hundred. Some clips analyzed have been viewed millions of times. The study also identified another 300 canonical porn sites that incorporate non-consensual *deepfakes* in some way, plus portals devoted to “leaks” and those that republish social media photos.

Trying to give a measure of the extent of online *deepfake* videos and images is nearly impossible, on the one hand because tracking where this content is shared on social media is complicated, and on the other hand because *deepfakes* are also shared in private messaging groups or closed channels, often by people known to the victims. A stunning case shows that, in September 2023, more than 20 girls between the ages of 11 and 17 in the Spanish town of Almendralejo were victims of artificial intelligence tools that generated photos of them naked without their knowledge (Hedgecoe, 2023).

The availability of AI tools for the creation of non-consensual *deepfake* pornographic images has grown significantly and the demand for this type of content has increased on pornographic platforms and online illicit networks, this phenomenon is likely to increase with new generative AI tools.

Analysis of Google searches in Canada, Germany, Japan, the United States, Brazil, South Africa, and Australia uncovered the fact that *deepfake* sites are prominently displayed in search results. Those targeted are often celebrities, streamers, and content creators. Maddocks (2018) states that the spread of *deepfakes* has now become endemic exactly what many researchers feared when the first *deepfakes* rose to prominence in December 2017, as can be seen in Burgess (2023). The first laws criminalizing sharing are beginning to arrive, but similarly to the ever improving and usable nature of the tools needed

to create these videos. The proliferation of *deepfake* apps combined with increased reliance on digital communications in the Covid-19 era and the inability of laws and policies to keep up has created the perfect storm. Certainly, better awareness of new technologies combined with measures to stop the spread of malicious tools, including by the companies that host the sites and search engines, such as Google and Bing, is to be seen as necessary. According the last source, *deepfake* and generative AI expert Henry Ajder, through monitoring the spread of these technologies, believes that adding more “friction” may reduce their spread: it is hard to be optimistic given the volume and scope of these operations, it would be helpful if platforms that have historically not taken these issues seriously would begin to address them. The same source notice that Google spokesperson Ned Adriaance says that as with any search engine, Google indexes content that exists on the web, but designs its ranking systems to avoid shocking people with harmful or explicit content that they don't expect and wouldn't want to see, which links to Google's blog page where the company explains the instances where it removes content from its search results. Google's support pages state that it is possible to request removal of “unintentional fake pornography.” The removal form asks users to manually enter the URLs in question and the search terms used to find the content. As the phenomenon continues to evolve, Google says it is actively working to add additional safeguards to help protect people, building on systems built for other types of non-consensual explicit images. At the same time, Courtney Gregoire, chief digital safety officer at Microsoft, says the latter does not allow *deepfakes*, rather it allows their reporting through its web forms (Burgess, 2023).

Social implications and methods of law enforcement

But is society adequately informed? According to respondents to our non-probabilistic survey, little (42.4 percent) or very little (54.1 percent), as shown in Annex Picture 3. Perhaps this is because in Italy, the first glimpses came only last April 2023, when in Rome two boys stripped their female companions using Bikini Off. An app born in France in 2022 and spread worldwide in a very short time, needless to point out the download records among 20-year-olds. October 2023, a similar case in Treviso: a girl attending middle school (so let's imagine a maximum age of 13) and a high school girl reported the incident, and the investigation into the nudes generated with the AI and who might have spread them is being investigated by the postal police.

According to Uk Government, in September 2023 Minister Suella Braverman explained that it will be essential to first and foremost eliminate AI-generated child pornography by arguing that AI-generated child sexual exploitation and abuse content is still illegal, regardless of whether or not it depicts a real child. The Online Safety Act will also require companies to take proactive actions in addressing all forms of sexual abuse of children and women online³. This could be one of the forms of contrast, in addition to those hypothesized by the respondents in Annex Picture 4.

³ <https://www.gov.uk/government/news/uk-and-us-pledge-to-combat-ai-generated-images-of-child-abuse>

Computing powerhouse Intel Corporation announced in November 2022 the creation of FakeCatcher, which is a tool detection and identification of distorted content 96 percent of the time by detecting blood flow in depicted faces⁴. This was the reason to introduce the question which results are in Annex Picture 5.

The thinking of turkish researchers Ilke Demir and Ciftci (2021, p. 2) is to analyze the color of facial tissues that microscopically change color as blood is pumped into the veins: “Although there is a few techniques focusing on gaze reenactment, artifacts remain when head is not front-facing and inconsistencies exist between pose and perspective.” These minute features are still not found to be replicable by artificial intelligence software.

| Conclusion |

Reviewing these cases and data, the question that arises is: what characterizes human beings? Are we heading towards a dystopic endoapocalypse due to the digital imaginary as Frezza (2015) suggests? The approach is no longer that of *deep learning*, but instead we try to find what is true in a video. Additional tools are for example AiOrNot, Reality Defender, Attestiv, Blackbird.AI, Sensity AI or the old Microsoft Authenticator. The latter was developed in 2020 and is capable of assigning a reliability score to a piece of content through the analysis of individual frames.

Clearly, however, the AI has found a way around the “quality control” in question, causing its reliability rate to plummet. This is clear evidence of how in this context, unfortunately, there are no solutions to solve the problem, only attempts to untangle this Ariadne's Web. In the same way that Ariadne provided Theseus with a thread to guide him out of difficulties, experiences with AI provide us with new “threads” of understanding to help us navigate its increasing complexity. Each move we make, however, allows the manipulation to adapt and change, making the path increasingly circuitous.

| References |

Amnesty (2021). *Barometro dell'odio. Intolleranza pandemica*. Amnesty International Sezione Italiana. <https://www.amnesty.it/barometro-dellodio-intolleranza-pandemica/>

Baudrillard, J. (1981). *Simulacres et simulations*. Galilee.

Burgess, M. (16 October 2023). Deepfake Porn Is Out of Control. *Wired*. <https://www.wired.com/story/deepfake-porn-is-out-of-control/>

⁴ <https://www.intel.com/content/www/us/en/newsroom/news/intel-introduces-real-time-deepfake-detector.html>

Chesney, R., Citron, D. (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*, 107, 1753–1820. <https://doi.org/10.15779/Z38RV0D15J>

Citron, D. (2019). Sexual Privacy. *Yale Law Journal*, 128, 1870–1938. https://scholarship.law.bu.edu/faculty_scholarship/620/

Cole, S. (2017, December 11). AI-Assisted Fake Porn Is Here and We're All Fucked. *Vice*. <https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn>

Demir, I., Ciftci, U. (2021). Where Do Deep Fakes Look? Synthetic Face Detection via Gaze Tracking. *ETRA '21 Full Papers: ACM Symposium on Eye Tracking Research and Applications*, 6, 1–11. <https://dl.acm.org/doi/10.1145/3448017.3457387>

Flynn, A., Powell, A., Scott, A., Camas, E. (2022). Deepfakes and Digitally Altered Imagery Abuse: A Cross-Country Exploration of an Emerging form of Image-Based Sexual Abuse. *British Journal of Criminology*, 63(6), 1341–1358. <https://doi.org/10.1093/bjc/azab111>

Frezza, G. (2015). *Endoapocalypse, The Walking Dead, The Digital Imaginary, The Post-Human*. Area-blu Editions.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2020). Generative Adversarial Networks. *Communications of the ACM*, 63(11), 139–144. <https://doi.org/10.1145/3422622> [First published, 2014]

Hedgecoe, G. (24 September 2023). AI-generated naked child images shock Spanish town of Almendralejo. *BBC News*. <https://www.bbc.com/news/world-europe-66877718>

Lerner, M. (1980). *The Belief in a Just World: A Fundamental Delusion*. Plenum Press.

Maddocks, S. (2018). From nonconsensual pornography to image-based sexual abuse: Charting the course of a problem with many names. *Australian Feminist Studies*, 33(97), 345–491. <https://doi.org/10.1080/08164649.2018.1542592>

Mento, C., Pira, F. (2023). *La violenza in un “click”. Profili sociologici, psicologici e giuridici del revenge porn*. Franco Angeli.

Mori, M. (2012). The Uncanny Valley: The Original Essay by Masahiro Mori. *IEEE Robotics & Automation Magazine*, 19(2), 98–100. [First published 1970, in Japanese]. <https://doi.org/10.1109/MRA.2012.2192811>

Öhman, C. (2020). Introducing the pervert’s dilemma: a contribution to the critique of Deepfake Pornography. *Ethics Inf Technol* 22, 133–140. <https://doi.org/10.1007/s10676-019-09522-1>

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

Pecchinenda, G. (2017). Il reale è virtuale, in *Exagère. Rivista di sociologia, psicologia, pedagogia, filosofia*, 5–6, p. 6.

Ryan, W. (1971). *Blaming The Victim*. Orbach & Chambers.

Schwartz, O. (2018, November 12). You thought fake news was bad? Deep fakes are where truth goes to die. *The Guardian*. <https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth>

Vacca, V. (2015). La cultura iper connessa. Il Museo 3.0. *Unclosed*, 2(7). <https://www.unclosed.eu/rubriche/restore/restore-restauro-ricomposizioni-riciclaggio-remake/107-la-cultura-iper-connessa-il-museo-3-0.html>

Waisbord, S. (2018). Truth is What Happens to News: On journalism, fake news, and post-truth. *Journalism Studies*, 19(13), 1866–1878. <https://doi.org/10.1080/1461670X.2018.1492881>

| Biographical Notes |

Roberta Casagrande is a PhD student at the Department of Political Science at the University of Messina. Since October 2023 she has been a subject expert in Sports Journalism-SPS / 08. She has obtained a 2nd Level Master's Degree in Political and Institutional Communication and Marketing from the Luiss Business School. She has a Master' Degree Course in Methods and Languages of Journalism, a course of study that allowed her to obtain the qualification of Communication Expert. She is part of the Order of Journalists of Sicily as a Publicist and in the past, she had various experiences in the Communications sector. She has always had a passion for the world of sport and for this very reason, her master's thesis is entitled “Pink Power: Gender Discriminations in the World of Sports Journalism and Sport”, a work which allowed her to win the 2020 National Ethic Award for Information and Communication Ethics.

Umberto Spaticchia is a PhD student at the Department of Political Science at the University of Messina. His path is completely hybrid: he obtains a bachelor's degree in information science: Journalism and a master's degree in Contemporary Philosophy, with a thesis on the pedagogical aspect on education through role-playing and participates, with the same topic, in the international project Open Innovative resources for high quality inclusive digital higher education (OIR) alongside the universities of Lublin (Poland) and Oviedo (Spain), as well as that of Messina (Italy). In 2023, he obtained a PhD in Political Science with a project to analyze the rise of Artificial Intelligence on the European social and political landscape on which he is currently working.

| Annex: Survey method and pictures |

The unit of analysis in this study was limited to individuals aged 20 to 40, residing in Europe, in the Autumn 2023, through a Google Forms sent to a first mail list and disseminated by waterfall method. That age group represents a particularly active demographic in terms of digital technology and social media use, platforms where *deepfake* and *deepreal* phenomena are most frequently observed. Additionally, Europe was chosen as the geographical focus due to the significant impact that digital regulation, gender violence awareness, and the adoption of new technologies are having in this region, making European participants a population of particular interest for this research.

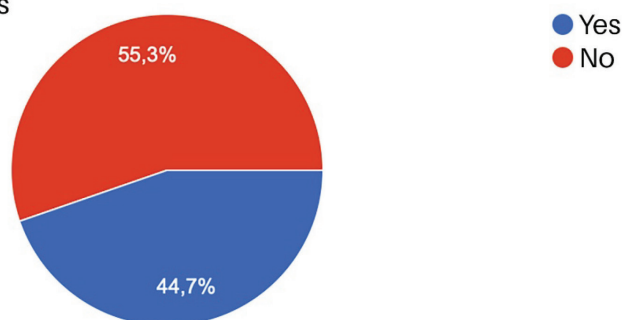
A non-probabilistic sampling method, specifically purposive sampling, was used. This approach was adopted to ensure the collection of diverse perspectives from individuals with varying levels of familiarity with *deepfake* technologies. Given the exploratory nature of the study, the primary objective was not to generalize the findings to the entire population but rather to identify emerging trends and formulate preliminary hypotheses about public perceptions of new forms of gender-based violence in the context of *deepfake* technologies.

The conclusions drawn from this methodological approach are limited to describing patterns and trends that emerged from the collected data. As the sampling was non-probabilistic, the conclusions cannot be considered representative of the general population. However, the findings provide valuable insights for further investigation and contribute to an informed discussion on the social, cultural, and ethical implications of the *deepfake* and *deepreal* phenomena. They also suggest avenues for future intervention to prevent gender-based violence in online environments.

Picture 1

Are you aware of the fact that DeepFakes are often used to carry out forms of gender violence?

85 answers



Subject: New forms of gender violence: between DeepReal and DeepFake

Age: 73% 20-30y 27% 31-40y

Gender: 51,7% Woman 47,1% Man 1,1% Genderless

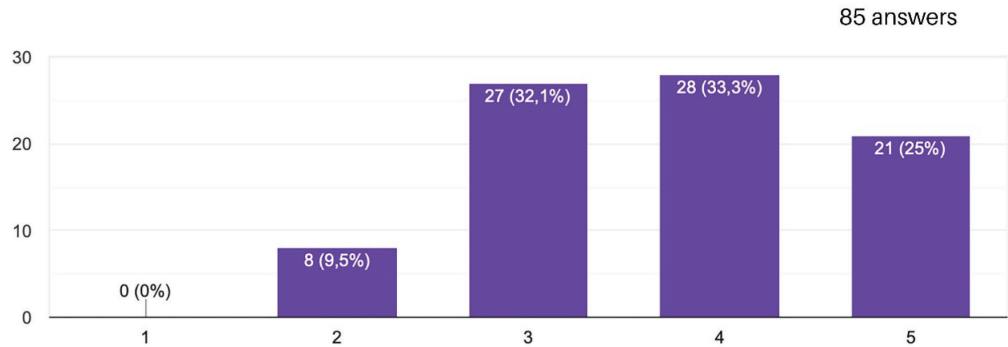
Population: Italians

Age of Research: 2023

Method: Quantitative Research - Google Forms

Picture 2

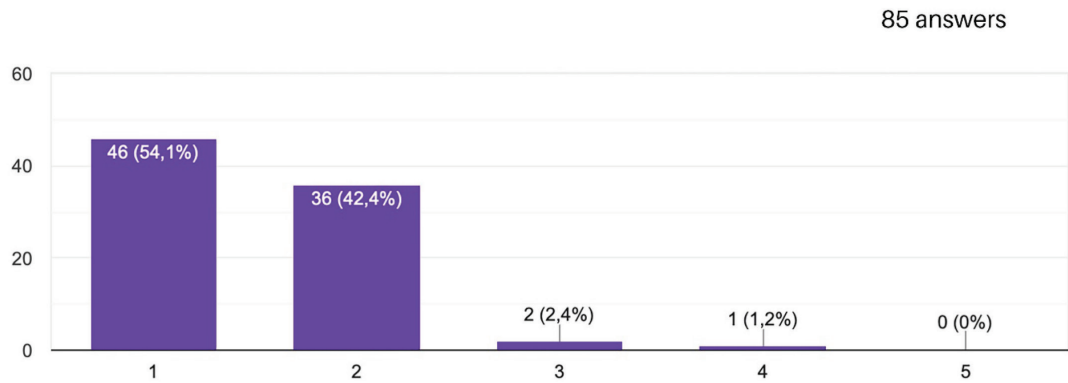
On a scale of 1 to 5, how much of a threat do you think DeepFakes pose to people's safety and dignity?



Subject: New forms of gender violence: between DeepReal and DeepFake
Age: 73% 20-30y 27% 31-40y
Gender: 51,7% Woman 47,1% Man 1,1% Genderless
Population: Italians
Age of Research: 2023
Method: Quantitative Research - Google Forms

Picture 3

In your opinion, is society adequately informed about the risks of DeepFake in relation to gender violence?

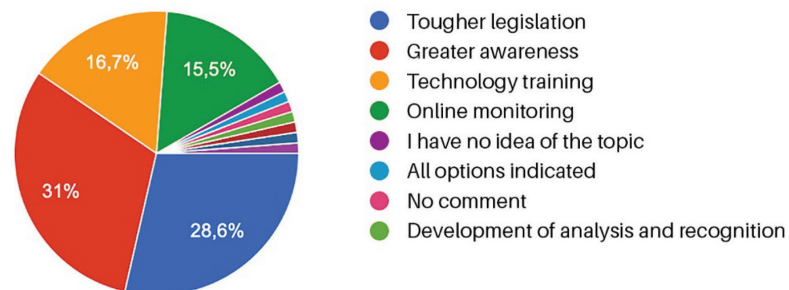


Subject: New forms of gender violence: between DeepReal and DeepFake
Age: 73% 20-30y 27% 31-40y
Gender: 51,7% Woman 47,1% Man 1,1% Genderless
Population: Italians
Age of Research: 2023
Method: Quantitative Research - Google Forms

Picture 4

What measures do you think should be implemented to combat the use of DeepFake in the context of gender violence?

84 answers



Subject: New forms of gender violence: between DeepReal and DeepFake

Age: 73% 20-30y 27% 31-40y

Gender: 51,7% Woman 47,1% Man 1,1% Genderless

Population: Italians

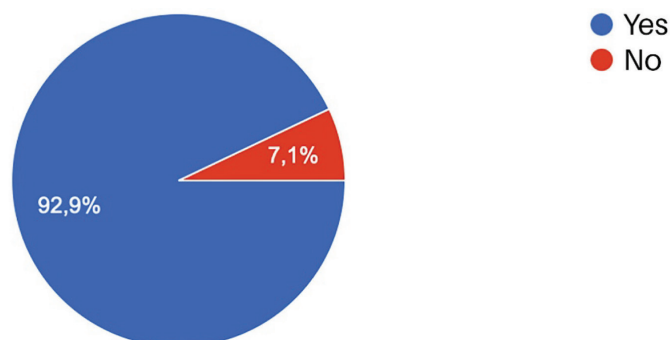
Age of Research: 2023

Method: Quantitative Research - Google Forms

Picture 5

Do you think the use of advanced technologies can help identify and counter DeepFake in the future?

85 answers



Subject: New forms of gender violence: between DeepReal and DeepFake

Age: 73% 20-30y 27% 31-40y

Gender: 51,7% Woman 47,1% Man 1,1% Genderless

Population: Italians

Age of Research: 2023

Method: Quantitative Research - Google Forms