Deepfakes and artificial intelligence in social engineering: Emerging threats in 21st-century cyberfraud*

Deepfakes e inteligencia artificial en la ingeniería social: amenazas emergentes en el ciberfraude del siglo XXI

Deepfakes e inteligência artificial na engenharia social: ameaças emergentes na fraude cibernética do século XXI

Yonni Albeiro Bermúdez-Bermúdez[1]

Cite as:

Bermudez Bermudez, Y. A. (2025). Deepfakes and artificial intelligence in social engineering: Emerging threats in 21st-century cyberfraud. *IUSTA*, (63), 54-71.

https://doi.org/10.15332/25005286.11068

Recibido: 05/04/2025
Aceptado: 05/05/2025

[1] Universidad Cooperativa de Colombia. Correo: yonni.bermudez@campusucc.edu.co. 0000-0001-8766-6953.

[*] This article is the result of research linked to the project ”Reform and Modernization of the State 2022–2026” INV 3701 of the Center for Legal, Political and Social Research of the Law School of the Universidad Cooperativa Campus Bogotá (CIFAD).

Abstract

This essay will address the increasing significance of deepfakes and generative AI in the context of social engineering tactics, emphasizing their role as a developing danger in cyber fraud in the twenty-first century. Recent technological advances have enabled the generation of hyperrealistic content that has the potential for near-undetectable identity impersonation. Therefore, contemporary cases of cyberfraud where deepfakes were used to trick victims, break authentication systems, and get to private information will be looked at using a mixed method that includes criminology, statistics, and specialized doctrine. This essay also addresses the importance of applying the criminological theory of routine activities as a preventive strategy, which can be used to identify everyday routines that can be exploited by cybercriminals. In the 21st century, it is important to have a culture of digital self-protection and self-regulation in digital environments, so that people don’t fall victim to cyber fraud.

Keywords:
deepfakes, cyber fraud, generative artificial intelligence, cybercriminals, social engineering.

Resumen

En el presente artículo se analizará el impacto creciente del uso de los deepfakes y la IA generativa como evolución de las estrategias de ingeniería social, destacando su papel de amenaza emergente en el ciberfraude del siglo XXI. Los recientes avances tecnológicos han permitido generar contenidos hiperrealistas que tienen el potencial de suplantar la identidad de forma casi indetectable. A partir de un enfoque mixto, en el cual se integran corrientes de la criminología, estadísticas y doctrina especializada, se examinarán casos recientes de ciberfraude en los cuales se emplearon los deepfakes para manipular a las víctimas y vulnerar sistemas de autenticación y acceso a información reservada. Asimismo, se discute la pertinencia de aplicar la teoría criminológica de las actividades rutinarias como estrategia preventiva, con la cual se pueden identificar rutinas cotidianas que pueden ser explotadas por los ciberdelincuentes. En el siglo XXI, se destaca la importancia de contar con una cultura de autoprotección y autorregulación digital en los entornos digitales para que las personas no sean víctimas del fraude cibernético.

Keywords:
deepfakes, ciberfraude, IA generativa, ciberdelincuente, ingeniería social.

Resumo

Este artigo analisa o impacto crescente do uso de deepfakes e IA generativa como evolução das estratégias de engenharia social, destacando seu papel como ameaça emergente no ciberfraude do século XXI. Os recentes avanços tecnológicos permitiram gerar conteúdos hiper-realistas que têm o potencial de suplantar a identidade de forma quase indetectável. A partir de uma abordagem mista, na qual se integram correntes da criminologia, estatísticas e doutrina especializada, serão examinados casos recentes de fraude cibernética nos quais foram utilizados deepfakes para manipular as vítimas e violar sistemas de autenticação e acesso a informações confidenciais. Além disso, discutiremos a pertinência de aplicar a teoria criminológica das atividades rotineiras como estratégia preventiva, com a qual é possível identificar rotinas diárias que podem ser exploradas por cibercriminosos. No século XXI, destaca-se a importância de contar com uma cultura de autoproteção e autorregulação digital em ambientes digitais para que as pessoas não sejam vítimas de fraude cibernética.

Palavras-chave:

deepfakes, fraude cibernética, IA generativa, cibercriminoso, engenharia social.

Introduction

Recently, advances in generative artificial intelligence (hereinafter, GAI) have impacted multiple productive, social, cultural, and personal sectors. But this new technology has also been used by malicious individuals to generate new forms of digital crime, which have been referred to by the doctrine as cybercriminals. A recent and constantly growing phenomenon is the use of GAI tools, such as the so-called *deepfakes*, with which they have managed to perfect engineering techniques socially, which would be used to execute cyber frauds increasingly difficult to detect.

Deepfakes can create hyper-realistic audiovisual representations that impersonate voices, faces, and identities with a degree of veracity that is challenging even the most sophisticated identity verification systems. Therefore, the technological advance of deepfakes is being exploited for financial fraud, identity theft, information manipulation, and digital extortion, among other forms of cyber fraud. However, these new criminal modalities not only violate individual and corporate security but also pose significant legal challenges.

As a criminal tactic, social engineering seeks the psychological manipulation of victims; with the emergence of GAI, this has acquired a new dimension, as this convergence has redefined the traditional tactics used by malicious individuals to deceive their victims. As a result, the cybercriminal has been provided with a tool that allows him to increase his capacity for persuasion and anonymity exponentially. Faced with this new and daunting landscape, it becomes urgent to examine how these new modalities of cyberfraud are evolving, to determine which strategies domestic legal systems should adopt to address them.

This article aims to analyze the impact of the use of GAI, especially deepfakes, on the social engineering strategies used in cyber fraud. To fulfill the proposed objective, the proposed research question is: How does the use of GAI, particularly deepfakes, impact the evolution of social engineering strategies used in contemporary cyber fraud? The conceptual aspects and the criminological and legal theories around this recent issue will be explored, as well as statistics and recent cases on cyber fraud related to GAI and deepfakes. Finally, the emerging challenges that arise in the detection, prevention, and self-regulation of these practices in the current digital environment will be addressed.

Through a mixed research approach, the main results show that deepfakes are increasingly difficult to detect by the human eye because of the development of generative AI. Therefore, the way forward is prevention through self-protection and self-regulation of potential victims, for which we start from the postulates of the routine activities theory, which states that crime occurs when three elements coincide, namely: (i) a motivated offender, (ii) a victim or potential target, and (iii) the absence of a capable guardian. In this context, strengthening the presence of the "capable gatekeeper" implies promoting safe practices in the digital environment that make it difficult to manipulate and disseminate false content, thus reducing the opportunities for victimization.

I. Conceptual approach: Deepfakes, GAI, cyber fraud, and social engineering

To properly understand the problem addressed, it is necessary to carry out a conceptual clarification of some key definitions that support the object of study, namely: deepfakes, GAI, cyber fraud, and social engineering. First, the neologism deepfake was born at the end of 2017 from the combination of two terms, (ii) *deep learning*, and (ii) *fake*, i.e., false (Amerini *et al.* , 2022). However, the term deepfake is part of a group of terms that have recently come to public light because of the construction of false and ill-intentioned discourses, including: (i) fake news, (ii) cheapfakes, (iii) fake nudes, among others (Cerdán & Padilla, 2019).

In general terms, the specialized doctrine on the subject states that the term *deepfake* is used to refer to a realistic-looking image, voice, or video which has been edited, altered, modified, or created without the consent of the owner and using generative AI (Yavuz, 2024). In a similar vein, the term *deepfake* is also used to denote the result of realistically generated videos, images, and/or voices (Deressa et al., 2024). While deepfakes have gained great popularity in educational, cultural, and artistic fields due to their high creative potential, these recent technological developments have not been exempt from being used to carry out illicit activities.

Due to their capacity to accurately mimic the appearance and voice of a real person, deepfakes have become one of the most effective tools with which you can currently consummate fraudulent activity, such as identity theft, digital extortion, and cyber fraud. These behaviors have become more popular over the years and have attracted the attention of the community in general. One of the most notable concerns lies in the ability to affect the credibility of videos, audio, information media, etc. This brings with it a substantial risk of disinformation, defamation, and the uncertainty in which we are immersed these days (Cerdán & Padilla, 2019).

The growth of this disinformation campaign is a consequence of the inadequate use of GAI, understood as the generation of content resulting from training certain large, advanced language models such as the Large Language Model (LLM) (Brown et al., 2020). These generate automatic responses once an analysis of a large data corpus has been performed. But the generated answers, despite being consistent, are not always correct (García-Peñalvo et al., 2024). Recently, generative AI is not being used to answer questions but for the automated production of high-quality textual, graphic, audio, and audiovisual content (Franganillo, 2023).

Generative AI reached a turning point in 2023, when it became widely popular due to its ability to produce content that simulated with remarkable accuracy style, coherence, and creativity, which at the time was only characteristic of humans (Corredera, 2023). This type of generative AI, based on advanced language models such as Generative Pre-trained Transformer (GPT) and image generation systems such as DALL-E, managed to generate texts, images, music, and even videos that are difficult to distinguish from those created by humans, which marked a milestone in the history of technological development. However, every advance brings about a debate around the limits of its use, copyright, and disinformation, among others.

Now, it is recognized that GAI plays a secondary role, since it was created with the purpose of support and to allow a more efficient management of some tasks and/or activities assigned to human beings. However, every development implies new challenges for the law, because some unscrupulous subjects have not given GAI proper management, but rather have used it to carry out illicit activities; a clear example are the so-called cyber frauds in which GAI is used to deceive victims.

We must understand that the concept of cyber fraud arises from the combination of two terms,: (i) cyber and (ii) fraud. The first refers to the relationship with computer networks (RAE, 2025), while the second is understood as an action against the interests of another (RAE, 2025). Thus, in general terms, when the term cyber fraud is used, it refers to deceptions that use computer networks to carry out illicit activities (Domínguez & Vázquez, 2022). This recent criminal dynamic employs technological means with malicious intent to manipulate, deceive, or mislead a natural or legal person to obtain an undue benefit, which is almost always economic.

This type of cyber fraud can be manifested in various forms, such as those involving bank cards: misuse, cloning, among others (Estupiñan, 2002). It occurs when the software or computer system of an entity is altered to change, activate, cancel, eliminate, or increase a banking product to the detriment of a third party (Lara & Alban, 2017). More recent modalities are phishing, impersonation, and online banking fraud, among others. These novel forms of deception take advantage of the victim's naivety and of technological vulnerabilities to obtain confidential data and cause economic loss for the victims.

Several techniques used by cybercriminals to conduct their illicit objectives have been fully identified, among which the following can be highlighted, without ignoring the fact that in an interconnected world, new techniques are emerging that are increasingly sophisticated and difficult to detect. The most common techniques are: (i) the use of malware, keyloggers, spyware, scams, and hacking (Quevedo, 2017; Alvarez, 2020).

Regarding the term *social engineering*, it was born in the business world from the Dutch businessman and philanthropist J. C. van Marken in 1894, and it was used at first to refer to the help received by employees regarding their personal problems. Subsequently, the term was adopted by Edward L. Bernays, publicist and journalist, but with a divergent meaning, as social engineering began to be used to manipulate people (López, 2015). Then, with time, the concept of social engineering evolved and was adopted in the field of computer security, where it acquired a new meaning. The term was first used in the field of computer security by Kevin Mitnick, a well-known hacker (López, 2015). It is used to refer to the different techniques used by cybercriminals to deceive or get potential victims to disclose confidential information or perform actions that compromise the security of computer systems (Rincón, 2023). This set of techniques seeks to take advantage of trust, ignorance, or fear to gain access to sensitive data, such as passwords, credit card numbers, or institutional information.

This practice has become one of the most effective techniques of contemporary cyber fraud, since it does not necessarily require the vulnerability of technological systems but rather takes advantage of human weaknesses to gain access to confidential information that can cause damage to the victim. Moreover, being a technique based on persuasion and deception, the distinctive feature of social engineering is that it adapts easily to various scenarios and communication channels, as would be the case of emails, phone calls, social networks, or altered websites; hence, it has become an imminent and growing risk in our digital environment.

In the current context of cybercrime, the terms mentioned correlate with each other, since GAI and social engineering techniques have significantly enhanced the various forms of cyber fraud. The use of deepfakes has allowed the creation of false identities and to impersonate real people to deceive victims; thus, deepfakes have become a new and advanced form of social engineering through which it is possible to emotionally manipulate humans by simulating genuine interactions, which are a latent risk of victimization.

II. Criminal techniques, statistics, and recent cases of cyber fraud related to GAI and deepfakes

The main criminal techniques employed by cybercriminals using social engineering are: (i) phishing, (ii) ransomware, (iii) cryptojacking, and (iv) cyber fraud; among others (Amerini *et al.*, 2022). The term phishing comes from fishing, since cybercriminals are likened to fishermen, who use bait; in the case of phishing, social engineering messages are used to steal (fish) personal information from victims. According to the Anti-Phishing Working Group (APWG), the term *phishing* dates to 1996, when hackers began to attack America OnLine accounts to be later traded in exchange for hacking software (Khonji *et al.*, 2013).

In the current context, phishing is used to deceive people generally through emails or text messages which contain a Uniform Resource Locator (URL) that redirects them to a fake website, with the main objective of obtaining confidential information from the victim such as: passwords, card numbers or bank details (Benavides *et al.*, 2020). These are subsequently used by cybercriminals to impersonate individuals before legitimate entities and perform financial transactions to the detriment of the victims, such as unauthorized transfers, credit applications, and online purchases, among others.

As for ransomware, this technique of intimidation was created by Joseph Popp in 1989, when he created the program AIDS (or PC Cyborg) which was deployed as a Trojan. The modality used by Popp focused on using a diskette containing the AIDS program, which had the potential to encrypt the files on the C drive of the computer where the diskette was inserted, to later request the delivery of 189 dollars, which should be sent to a post office box in Panama to release the encrypted information (O'Kane at al.Carlin, 2018).

This social engineering technique has evolved to the point that today it is used by cybercriminals to block access to the victim's systems or files through data encryption (Komatwar & Kokare, 2020). Once they have control of the information, they proceed to ask the owner of the information for a payment. Generally the payment is requested through cryptocurrencies in exchange for providing the decryption key so that the victim can regain access to confidential and/or sensitive information (Kara & Aydos, 2020). This type of attack is characterized by targeting both natural and legal persons, since it has a high damaging potential on the reputation of the organizations that are victims of these attacks, which brings serious economic and administrative consequences.

Another technique to consider is cryptojacking. This modality was born with the emergence of Bitcoin in 2009; however, it became popular in 2017, possibly because of the boom in popularity and value of cryptocurrencies (Velasco, 2022). Since then, cybercriminals began to use various malicious scripts and specialized software through which they managed to exploit, without the owner's authorization, users' computational resources to mine cryptocurrencies. This practice has led to a significant increase in cryptojacking attacks worldwide, affecting both individual users and large corporations that carry out their transactions through these currencies.

This social engineering technique occurs when the cybercriminal covertly uses the processing resources of a device such as a computer, server, or cell phone to mine cryptocurrencies without the user's consent. This activity consists of three main phases: i) script preparation, ii) script injection, and iii) attack (Tekiner, *et al.*, 2021). Injection can come in three forms, namely: In websites, in RAM, or by embedding it locally in the network; hence, cryptojacking is classified into three categories: (i) In-browser, (ii) In-host, and (iii) In-memory (Varlioglu *et al.*, 2022). This technique seeks to affect the system, increase the consumption of energy, and reduce the useful life of the hardware, all without the victim being aware of what is happening.

Currently, a new criminal technique that employs social engineering has been occupying a relevant position within cybercrime: the so-called cyber frauds. These consist of deceiving victims by means of persuasive and manipulative methods to obtain illicit economic benefits. However, these serious attacks have evolved, and nowadays generative AI is used to deceive victims through the simulation of voices, audios, videos, and even hyper-realistic faces, which has increased the degree of sophistication and credibility of the deception by cybercriminals. This new dynamic has generated great concern among the different authorities, companies, and users, as it increases the vulnerability of traditional verification systems.

The recent phenomenon of cybercrime can jeopardize the internal security systems of every state around the world, and this has been demonstrated in the figures of the last five years, where there has been a 67 % increase in the incidence of security breaches worldwide (Amerini, Baldini & Leotta, 2022). Now, the Colombian case has not been alien to this reality, since according to the figures of the Sistema de Información Estadística, Delincuencial, Contravencional y Operativa de la Policía Nacional de Colombia (Statistical, Criminal, Contravention and Operational Information System of the National Police, SIEDCO), the number of complaints filed in recent years because for alleged computer crimes has shown a significant growth, as evidenced in the following chart:

To calculate the increase in cybercrime in Colombia and to know the panorama in greater detail, a comparison was made between the number of complaints reported in SIEDCO during the years 2020 to 2024; 2025 was not taken since the year is in progress. In 2023, 63,249 complaints were recorded, while in 2024 the figure rose to 74,845 cases. In 2022, 61,993 complaints were registered , while in 2021, 52,224 complaints were made. Finally, in 2020, 49,353 complaints were registered. The following formula was used to calculate the percentage increase:

Percentage increase=

$$\left(\frac{\text{Initial value} - \text{Final value}}{\text{Initial value}} \times 100\right)$$

The results show that the difference between 2023 and 2024 represents an approximate increase of 18.33 %; between 2022 and 2023 an approximate increase of 2.03 %; between 2021 and 2022 an approximate increase of 18.71 %; and between 2020 and 2021 an approximate increase of 5.82 %. These figures show how computer crimes in Colombia have maintained a generalized upward trend in the last four years, reflecting the growing impact of digitization and the sophistication of criminal techniques in the cyber field. From this perspective, it is imperative to analyze how computer crimes have behaved individually during the last five years. To achieve this objective, the figures reported in SIEDCO for the years 2020 to 2025 were taken again, as presented below:

The graphs above show a relative increase in five of the nine crimes analyzed; the same formula mentioned above was used to calculate the percentage increase, . The data analyzed covers the period 2023-2024, where the results show that the increase in the crime of article 269A was 41.95 %; in the crime of article 269B the increase was 16.33 %; in the crime of article 269F the increase was 7.68 %; in the crime of article 269G the increase was 18.83 % and in the crime of article 269I the increase was 19.14 %. However, the crime of article 269C decreased by 39.51 %; the crime of article 269D decreased by 34.98 %; the crime of article 269E decreased by 41.42 % and the crime of article 269J decreased by 1.66 %.

These figures show a mixed trend in the evolution of cybercrime in Colombia during the period 2023-2024. While some illicit activity show a sustained growth, which could be associated with the improvement of criminal techniques used by cybercriminals and the increase in the use of social engineering, other behaviors show a decrease, which may be related to the consequence of greater institutional and/or individual control, improvements in security systems or changes in criminal patterns by cybercriminals who find more interesting crimes because they obtain a

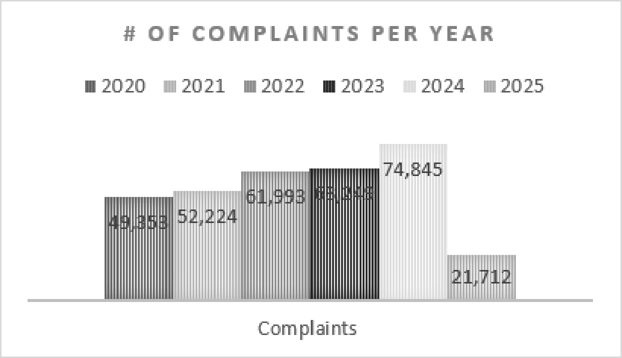

Figure 1. *No. of complaints per year*
Source: own elaboration based on Statistical Information System, Contravention and Operational Crime of the National Police, 2025.

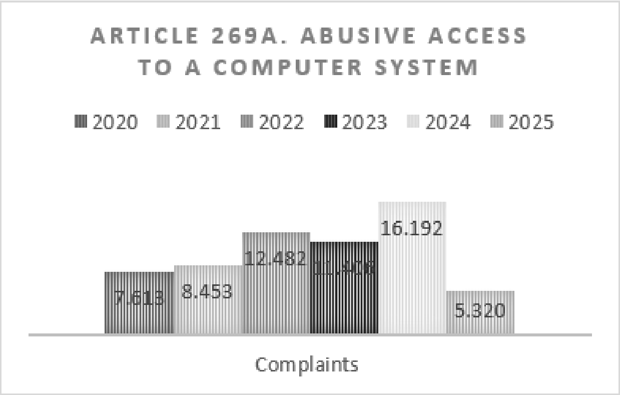

Figure 2. *Abusive access to a computer system*
Source: own elaboration based on Statistical Information System, Contravention and Operational Crime of the National Police, 2025.

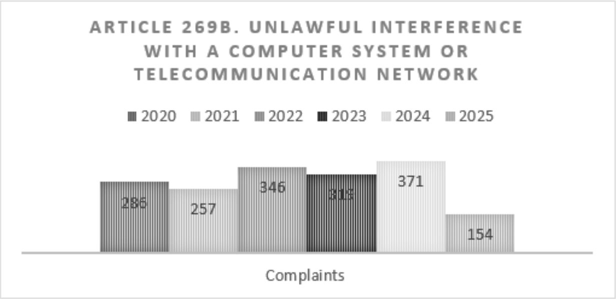

Figure 3. *Unlawful interference with a computer system or telecommunication network*
Source: own elaboration based on Statistical Information System, Contravention and Operational Crime of the National Police, 2025.

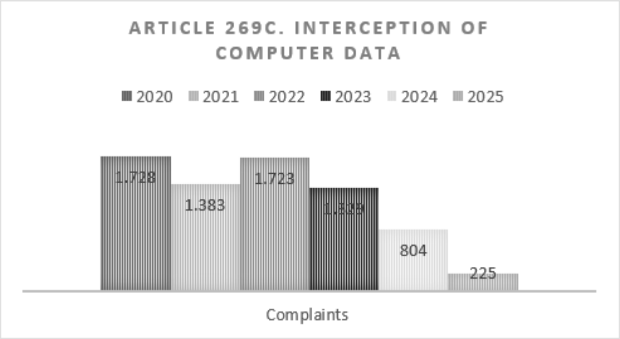

Figure 4. *Interception of computer data*
Source: own elaboration based on Statistical Information System, Contravention and Operational Crime of the National Police, 2025.

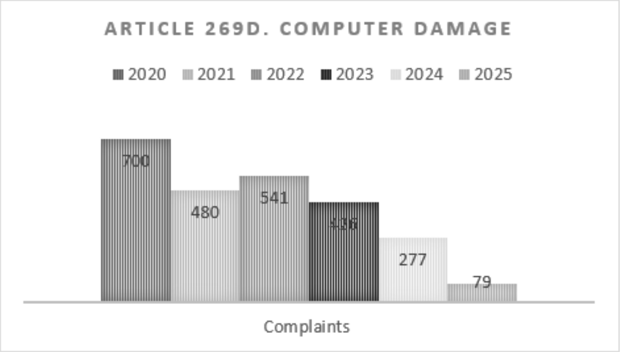

Figure 5. *Computer damage*
Source: own elaboration based on Statistical Information System, Contravention and Operational Crime of the National Police, 2025.

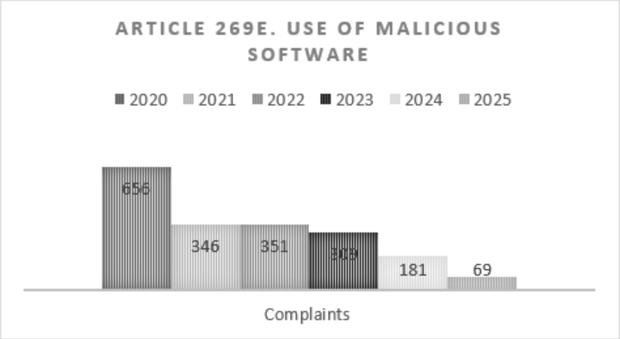

Figure 6. *Use of malicious software*
Source: own elaboration based on Statistical Information System, Contravention and Operational Crime of the National Police, 2025.

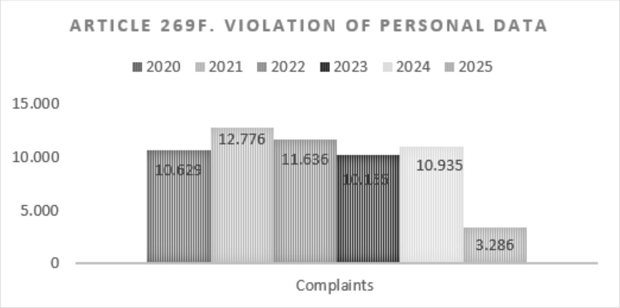

Figure 7. *Violation of personal data*
Source: own elaboration based on Statistical Information System, Contravention and Operational Crime of the National Police, 2025.

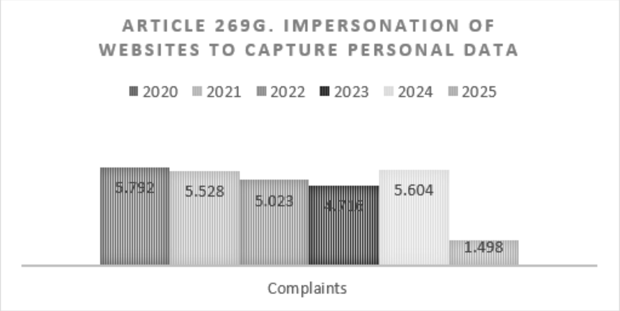

Figure 8. *Impersonation of websites to capture personal data*
Source: own elaboration based on Statistical Information System, Contravention and Operational Crime of the National Police, 2025.

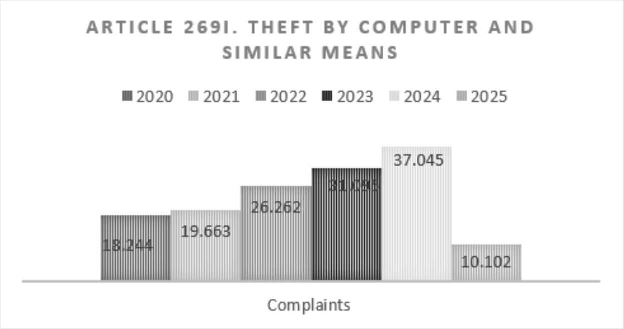

Figure 9. *Theft by computer and similar means*
Source: own elaboration based on Statistical Information System, Contravention and Operational Crime of the National Police, 2025.

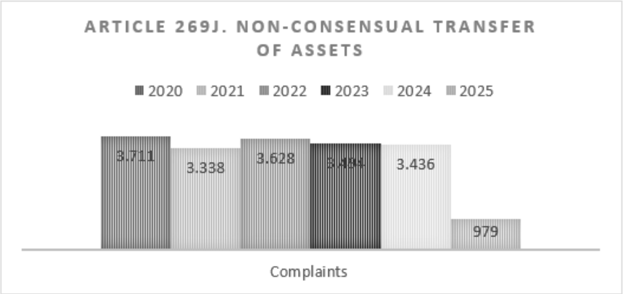

Figure 10. *Non-consensual transfer of assets*
Source: own elaboration based on Statistical Information System, Contravention and Operational Crime of the National Police, 2025.

greater economic benefit with reduced risk of detection.

What can be highlighted so far is that the crimes are on the rise, namely: abusive access to a computer system, illegitimate hindering of a computer system or telecommunication network, violation of personal data, impersonation of websites to capture personal data, and theft by computer and similar means. These are the crimes most likely to be committed in the form of cyber fraud through its different criminal modalities, such as phishing, ransomware, and cryptojacking. Considering the above, we are inclined to consider that one of the main factors that has influenced the increase in the number of illicit behaviors is the improvement of cybercriminal techniques with the rise of generative AI, including the so-called deepfakes.

Deepfakes work through computerized techniques in which generative AI is used to represent people saying or doing things that they never said or did (García-Ull, 2021). Several platforms and/or applications have been created to create deepfakes, among which we can highlight: (i) FaceApp; this application allows users to adjust a photograph with the use of AI to modify faces realistically, change gender, facial expressions, backgrounds, among others; (ii) DeepFaceLab; one of the applications most used by users to create realistic deepfakes, especially in videos. This application allows replacing faces in recordings using advanced generative AI techniques, in particular, convolutional neural networks (CNNs) and generative adversarial networks (GANs); (iii) Synthesia; this tool allows the creation of hyperrealistic human avatars that speak from text, among others.

Among the most relevant cases that have come to public light regarding the use of deepfakes are: (i) In March 2021 European deputies witnessed a live video call with a doctored image of Leonid Volkov, one of the opponents of the Russian government; they later discovered that it was a montage with deepfake filters (Campillo 2021). (ii) In 2019, a British energy company was scammed for US$243,000, in this case the scammers used voice-generating AI software to mimic the voice of the CEO of the company's parent company, based in Germany, to facilitate an illegal transfer of funds; (iii) In the current year (2025), it has been reported that calls are being made to victims using AI-generated voices, which sound like they belong to real people. During the call, they inform the victim that their email account has been "compromised" and that they can help them fix it, this to gain access to their personal information.

III. Routine activities theory: Detection and prevention of cyber fraud through deepfakes.

To try to understand modern criminal dynamics, especially cyber fraud through deepfakes, it is necessary to resort to criminological currents of thought. For this analysis, we will take as a starting point the theory of routine activities, which explains how daily routines in digital environments, such as the constant use of social networks, financial movements through banking applications, and the use of video call platforms, among others, generate opportunities for cybercriminals to act and to carry out their illicit activities.

The criminological theory of routine activities proposed by Cohen and Felson in 1979 argues that the occurrence of a crime does not depend exclusively on social or psychological factors of the offender, but that three elements must coincide for it to befall: i) a motivated potential offender; ii) a suitable "target" or objective, and iii) the absence of possible guardians (Cohen & Felson, 1979). The first element refers to the fact that the offender has motivation and the necessary skills to commit the crime. The second element refers to the target or objective of the offender, which can be a person or an object. The availability of the target depends on four characteristics: value, inertia, visible and accessible, which have been synthesized with the acronym VIVA (Aloisio & Trajtenberg, 2009).

The first characteristic indicates that the target must have an attractive *value* for the offender; the value depends on the offender, since it is not always economic, and may depend on the offender's tastes. The second characteristic of the target must be *inertia*, i.e., if it is an object, the size has an impact, since the easier it is to move, the more attractive it is for the offender, and if it is a person, the inertia is related to the victim's lack of physical capacity to resist the offender. As for the third characteristic, the target must be *visible*, so that the offender will be able to determine whether the target is present or not. Finally, it is recognized as *accessible*, which implies that the offender can reach the target, but is also able to retreat or escape if necessary (Felson, 1986).

The third element refers to the lack of capable guardians, a person or element whose presence causes the offender to desist from committing the crime. This criminological current of thought is based on the central idea that daily or routine activities contribute to the creation of the propitious scenario for the crime to take place. This criminological theory focuses on the situational context of the crime and not exclusively on the individual characteristics of the offender as other theories have done (Norza *et al.*, 2011).

Once the offender's modus operandi or the situational context that he or she uses to achieve his or her objective has been identified, the theory of routine activities helps prevent crime by modifying behavioral patterns, reducing the opportunities or scenarios conducive to the commission of crime. However, despite its usefulness, the theory of routine activities has been criticized, among the main criticisms are: (i) it focuses its analysis on the situational aspect and leaves aside the structural and socioeconomic causes of crime, and (ii) it holds victims indirectly responsible for the commission of the crime, since it suggests that their routines are the cause of the crime.

However, its advocates argue that it is not a matter of blaming the victim or re-victimizing him or her for allowing the crime to take place, but of understanding how daily routines related to the likelihood of crime occurrence. Therefore, by knowing which scenarios or activities the offender takes advantage of to commit a crime, more effective preventive strategies can be implemented to put an end to or at least reduce the number of victims (Felson & Clarke, 1998). This contributes to the development of a more proactive criminal policy oriented to the prevention and management of the different risks in a crime, instead of an exclusively reactive or punitive criminal policy, which only seeks to compensate for the damage caused through the imposition of a penalty or sanction.

According to data from Signicat's *The Battle Against AI-Driven Identity Fraud* report, cyber fraud has increased by 80 % while identity fraud has increased by 74 % in the last three years. But what is most striking about this study is that deepfakes already account for 6.5 % of fraud attempts in the last three years (Signicat, 2024). This demonstrates how new technologies based on generative AI are being exploited by cybercriminals to perfect their social engineering methods to achieve identity spoofing and deceive victims, generating a new modern risk scenario to which we are exposed daily.

According to the report, it was concluded that 42.5 % of the detected cyber fraud attempts use AI, but what is most worrying is that 29 % of them are successful (Signicat, 2024). However, with respect to the use of deepfakes, two types of attacks were identified: (i) presentation attacks and (ii) injection attacks. The first occurs when the camera records another screen showing a deepfake. The second type occurs when the cybercriminal deliberately inserts a malicious program, including deepfakes or manipulated pre-recorded videos. According to the study, injection and presentation attacks using deepfakes have fluctuated as follows over the past four years:

The formula described above will be used to identify the increase in attacks by presentation and injection. I will begin the analysis of the attacks by presentation, which in 2021 were 7.6 % compared with 8.6 % in 2022, so in this interval they increased by 13.16 %; in 2023 they were 12.6 %, which compared with the previous year represents an increase of 46.51 %.

However, between 2023 and 2024, there is a decrease of 4.76 % in presentation attacks through deepfakes. On the other hand, injection attacks for the year 2021 represented 1.5 % compared to 2022 with 1.7 %, which represents an increase of 13.33 %; in 2023 they were 3.5 %, which compared to the previous year represents an increase of 105.88 %. Finally, in 2024, they were 7.5 %, which,

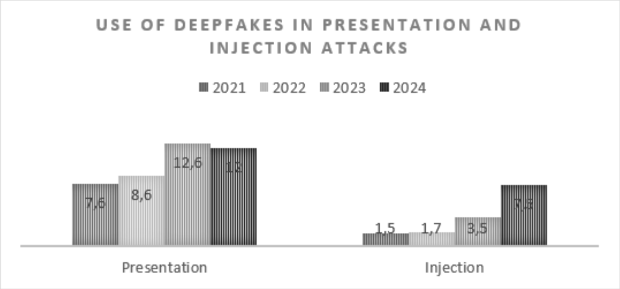

Figure 11. *Use of deepfakes in presentation and injection attacks*
Source: own elaboration based on Statistical Information System, Contravention and Operational Crime of the National Police, 2025.

compared to the year 2023, represents an increase of 114.29 %.

The above figures show that there has been a sharp increase in recent years in the number of injection attacks using deepfakes. There has also been a relative stabilization of presentation attacks, but this highlights the fact that deepfakes are not only more common but also more sophisticated and difficult to detect. This situation poses a significant challenge for identity verification systems, which must begin a process of updating to adapt to modern technological advances used by cybercriminals to carry out cyber fraud using deepfakes.

This implies that we must be more cautious, avoiding predictable password routines, such as the use of birth dates, simple numerical sequences, or passwords reused on various platforms, which facilitates the work of cybercriminals. It is also important to verify the sender's e-mail address or telephone number, since cybercriminals commonly use addresses or numbers that appear legitimate to deceive their victims, but contain slight errors that can be detected by humans if they exercise caution and verify where the information comes from.

If you receive an unexpected and/or suspicious message, you should resort to several authentication factors, such as validating directly with the company and/or person who sent the message. If, on the other hand, the message contains a URL, it is not recommended to access it, because through these URLs, cybercriminals appropriate confidential information from the owner that will later be used to carry out various cyber frauds. Therefore, it is important to check for any warning signs, as it is common for cybercriminals to use bad grammar, misspellings, and unusual proposals.

That said, we must take a stance towards prevention of the use of deepfakes to commit cyber fraud; even if the specialized doctrine points out that every day they are more difficult to detect by the human eye, it is important to adopt verification strategies such as : (i) sudden movements, (ii) unnatural elements, (iii) excessive lighting, (iv) sharp color contrasts, (v) facial expressions, (vi) synchrony, (vii) naturalness of gesticulation, among other factors. These allow us to verify whether we are in the presence of a human being or a deepfake to prevent deception in identity verification processes, personal data protection, and security in digital environments.

Recently, anti-deepfake technologies have been developed, as in the case of Microsoft, who developed a program called Microsoft Video Authenticator to validate if a video is real or fake. Regarding the use of deepfakes to simulate voice, it is recommended that the receiver of the information can validate certain information with the transmitter, such as names of people, pets, special dates, among others. This validation significantly reduces the risk of being a potential victim of voice impersonation using voice deepfakes, since personal elements are difficult to replicate even by advanced AI systems.

It is important to note that nowadays there are multiple proposals to combat cyber fraud, including: (i) updated software, multifactor authentication, watermarking, anti-depository programs, among others. But what is common to all proposals is that they are aimed at self-protection and self-regulation; hence, it is important to apply the criminological theory of routine activities to prevent becoming a victim of cyber fraud with deepfakes. It is necessary to analyze which routine activities can be used by cybercriminals and adopt the respective changes to promote a culture of cybersecurity based on the reduction of criminal opportunities, thus minimizing exposure to risks in digital environments and strengthening personal vigilance over the information that is shared and consumed.

Likewise, through what has been called the "capable guardian" doctrine, it is important to promote safe practices in the digital environment that make it difficult to manipulate, alter or disseminate false content, thus reducing opportunities for victimization. We must all be watchdogs and verify information before sharing it, verify sources, analyze the authenticity of content, and use technological tools to verify its veracity. This active role of the "capable guardian" represents a stance on digital self-protection that contributes significantly to preventing cyber fraud through deepfakes and to building a safer, more trustworthy digital environment for all.

Conclusions

The use of GAI has radically transformed the current cyber fraud landscape, especially in the field of social engineering. In particular, the use of deepfakes such as videos, images, or audios manipulated by GAI is being used by cybercriminals to achieve hyper-realistic identity impersonation in a sophisticated way, getting victims to make critical decisions such as making money transfers, delivery of sensitive data, among others, without questioning the authenticity of the message, voice, or video. This technological advance generates a shift from textual to audiovisual cyber fraud, which reduces the ability to detect deception.

Deepfakes have managed to amplify the power of social engineering, as they have become a more credible, personalized, and difficult for humans to detect. The high use of deepfakes is reflected in the figures presented above, which show an increase in the number of cases of abusive access to a computer system, illegitimate hindering of a computer system or telecommunication network, violation of personal data, impersonation of websites to capture personal data, and theft by computer and similar means. Against this backdrop, defensive strategies must also evolve. While it is true that measures have been taken, a more active and multifactorial verification of identity is required today, accompanied by technological tools that can detect manipulations in real time.

In the current context, I believe that it is important to apply the criminological theory of routine activities for the prevention of cyber fraud in which deepfakes are used, because it is important to analyze how the everyday behavior of people in digital environments can increase or reduce their vulnerability to sophisticated attacks such as those with GAI. The theory of routine activities provides a preventive approach based on reducing opportunities for crime, fully applicable to cyber fraud. In this context, the key is not only in the technology, but also in educating users to modify their digital routines, identify threats, and adopt safe practices, i.e., becoming a "capable guardian" that can limit the effectiveness of social engineering strategies used by cybercriminals.

Referencias

Aloisio, C., & Trajtenberg, N. (2009). Rationality in contemporary criminological theories: Uruguay from sociology VII. Department of Sociology of Uruguay, Faculty of Social Sciences, UDELAR.

Alvarez, F. (2020). Machine learning in e-commerce fraud detection applied to banking services. *Science and Technology, 20*, 81-95.

Amerini, I., Baldini, G., & Leotta, F. (Eds.). (2022). *Image and video forensics.* MDPI. https://doi.org/10.3390/books978-3-0365-2807-6

Benavides, E., Fuertes, W., Sanchez, S., & Nuñez-Agurto, D. (2020). Characterization of phishing attacks and techniques to mitigate these attacks: a systematic review of the literature. *Science and Technology, 13(1)*, 97-104.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *arXiv.* https://doi.org/10.48550/arXiv.2005.14165

Campillo, Beatriz. (2021). From fake news to deepfakes: New challenges for cyberethics. *Anuario Colombiano de Filosofía, 2*(2), 89-106. https://tinyurl.com/yc3wcw45

Cerdán Martínez, V. M., & Padilla Castillo, G. (2019). History of audiovisual fake: Deepfake and women in a falsified and perverse imaginary. *History and Social Communication, 24*(2), 505-520. https://doi.org/10.5209/hics.66293

Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review, 44*, 588-608. https://doi.org/10.2307/2094589

Corredera, J. C. (2023). Generative artificial intelligence. *Anales de la Real Academia de Doctores, 8*(3), 475–489.

Deressa, D. W., Lambert, P., Van Wallendael, G., Atnafu, S., & Mareen, H. (2024). Improved deepfake video detection using convolutional vision transformer. In 2024 IEEE Gaming, Entertainment, and Media Conference (GEM) (pp. 492–497). IEEE. https://doi.org/10.1109/GEM61861.2024.10585593

Domínguez Arteaga, R. A., & Vázquez, R. V. (2022). Spatial analysis of e-commerce cyberfraud: Considerations in the Tamaulipeca political agenda. *Podium, (41)*, 21-40.

Estupiñán, R. (2002). *Internal control and fraud.* ECOE Ediciones.

Felson, M. (1986). Linking criminal choices, routine activities, informal control, and criminal outcomes. In D. B. Cornish & R. V. Clarke (Eds.), *The reasoning criminal: Rational choice perspectives on offending* (pp. 119–128). Springer-Verlag. https://doi.org/10.1007/978-1-4613-8625-4_8

Felson, M., & Clarke, R. (1998). *Opportunity makes the thief: Practical theory for crime prevention.* Home Office Policing and Reducing Crime Unit, Research, Development and Statistics Directorate.

Franganillo, J. (2023). Generative artificial intelligence and its impact on media content creation. *Methods: Journal of Social Sciences, 11*(2), 10.

García-Peñalvo, F. J., Llorens-Largo, F., & Vidal, J. (2024). The new reality of education in the face of advances in generative artificial intelligence. *RIED: Revista Iberoamericana de Educación a Distancia, 27*(1), 9-39.

Garcia-Ull, F. (2021). Deepfakes: The next challenge in fake news detection. *Anàlisi. Quaderns de Comunicació i Cultura, 64*, 103-20. https://doi.org/10.5565/rev/analisi.3378

Kara, I., & Aydos, M. (2020). Cyber fraud: Detection and analysis of the crypto-ransomware. In *2020 11th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*; IEEE. https://doi.org/10.1109/UEMCON51285.2020.9298128

Khonji, M., Iraqi, Y., & Jones, A. (2013). Phishing detection: A literature survey. *IEEE Communications Surveys & Tutorials, 15*(4), 2091-2121.

Komatwar, R., & Kokare, M. (2020). Survey on malware detection and classification. *Journal of Applied Security Research, 15*(1), 1–31.

Lara Guijarro, E. G., & Albán Silva, L. C. (2017). The risks of internet banking transactions. *Revista Publicando, 4*(10[1]), 62-74. https://revistapublicando.org/revista/index.php/crv/article/view/436

López Grande, C. E. (2015). Social engineering: The silent attack. *Technological Journal, 8*.

Norza, E., Ruiz, P. J., Rodríguez, M. L., & Useche, H. S. (2011). Theories and explanatory models of criminology. *Revista Investigación Criminológica, 2*(1).

O'Kane, P., Sezer, S., & Carlin, D. (2018). Evolution of ransomware. *IET Networks, 7*(5), 321-327.

Quevedo González, J. (2017). Investigation and proof of cybercrime [Doctoral dissertation, University of Barcelona]. http://hdl.handle.net/10803/665611

Rincón Nuñez, P. M. (2023). Social engineering based attacks in Colombia, good practices and recommendations to avoid risk. *InterSedes, 24*(49), 120-150. https://doi.org/10.15517/isucr.v24i49.50345

Tekiner, E., Acar, A., Uluagac, A. S., Kirda, E., & Selcuk, A. A. (2021). SoK: Cryptojacking malware. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)* (pp. 120-139). IEEE.

Varlıoğlu, S., Elsayed, N., ElSayed, Z., & Ozer, M. (2022). The dangerous combo: Fileless malware and cryptojacking. In *SoutheastCon 2022* (pp. 125–132). IEEE.

Velasco Sanchez, L. (2022). New dilemmas introduced due to SARS-CoV-2 (COVID-19) from a malware point of view. University of Malaga.

Yavuz, C. (2024). Criminalisation of the dissemination of non-consensual sexual deepfakes in the European Union: A comparative legal analysis. *Revue internationale de droit pénal, 95*(2), 419-457.

About the authors

[1] PhD candidate, Universidad de Lleida, Spain. Master's degree in Criminal Procedural Law, Universidad Militar Nueva Granada. Specialist in Criminal Law, Universidad del Rosario. Lawyer, Universidad La Gran Colombia. Correo: yonni.bermudez@campusucc.edu.co. ORCID: 0000-0001-8766-6953.