https://doi.org/10.70265/SEGD3729

DEEPFAKE AND THE COLLAPSE OF TRUST: CHALLENGES TO CYBERSECURITY AND THE RULE OF LAW

Gergana Varbanova

Summary: *The advancement of artificial intelligence systems and the increasing accessibility of technologies capable of rapidly and inexpensively generating synthetic content – deepfakes – with a high degree of illusory credibility pose serious challenges to cybersecurity. These developments underscore the urgent need for an adequate legal framework to protect both individual citizens' rights and national security. This protection must be achieved through the enhancement of societal and individual critical awareness in response to the growing volume of synthetic content. Deepfakes undermine trust in the authenticity of information, create conditions for reputational damage, manipulate public opinion, erode institutional authority, and complicate evidentiary procedures in judicial practice.*

This study aims to analyze the cybersecurity challenges and the impact of deepfake technologies on the resilience of cyberspace. It highlights the necessity of establishing a comprehensive legal framework that reflects the dynamic nature of the increasing use of synthetic content as a tool for conducting cyberattacks.

In conclusion, the study asserts that an effective legal response should be not only restrictive and reactive but should also foster a culture of critical thinking in order to restore trust in the digital environment.

Keywords: *Deepfake, Cybersecurity, Synthetic Content, Legal Framework, Epistemic Vigilance*

INTRODUCTION

Following the global pandemic of 2019–2020, a significant portion of public relations and social interaction shifted from the analog to the digital environment. This transition introduced a number of challenges for law enforcement and judicial authorities. Cyberspace has become a more convenient virtual space for perpetrators of crimes involving social engineering, often combined with the use of software applications that generate synthetic content. The generation of synthetic content has revealed new forms of criminal behavior previously unknown to legal systems. This article examines the current national and supranational regulatory frameworks applicable to artificial intelligence systems, with a particular focus on the regulation of applications used for creating synthetic content.

Timely legal regulation is a key instrument for safeguarding the rights of individuals in cyberspace, as well as for the prevention and counteraction of cybercrime. *Social media is a powerful tool of influence; people can be convinced that they see or hear things that never actually happened* (Chaushev, 2024). This necessitates the establishment of specific rules governing the generation and use of synthetic content. Without such regulations, individuals or entire societal groups may be misled, misinformed, or even become victims of criminal acts.

This study aims to analyse the legal, social, and ethical challenges posed by synthetic content and deepfake technologies in the context of cybersecurity, the protection of individual rights and fundamental freedoms of citizens, as well as the pursuit of balance in the legal regulation of the information environment.

This study examines the social relations that arise from the creation, dissemination, and perception of synthetic content, while the subject of the study encompasses the applicable legal norms, regulatory mechanisms, and the accompanying issues related to the use of synthetic content in various spheres of social relations.

The author's thesis defended in the study is that addressing the risks arising from deepfake content must be based on an integrated approach – through the establishment of a clear legal framework, the implementation of technical solutions for labelling and identifying synthetic content, and the promotion of a culture of critical thinking aimed at reducing society's vulnerability to various forms of synthetic manipulation and disinformation.

THEORETICAL FOUNDATIONS AND DEFINITIONS

The past five years have shown that artificial intelligence systems are used not only to enhance quality of life and optimize the work processes of individual employees, but also to influence political leanings and decision-making in the context of political choices and electoral attitudes, thereby directly impacting political processes at both national and international levels. Through AI-generated deepfake content, a state with strategic interests in a given region may influence the electoral process in another state, thus gaining a strategic advantage in shaping regional policy. Synthetic content can be employed as a means of compromising political actors and public figures by attributing to them actions or statements they never made.

This study employs a comparative legal analysis methodology. It contrasts various legislative approaches – those of the European Union (the Artificial Intelligence Act), the United States (California State Bill AB 2839), and the People's Republic of China (PRC) – in order to assess their effectiveness in ensuring transparency, legal certainty, and the protection of both political processes and individual privacy. Empirical case studies are used to illustrate the legal, social, and ethical dimensions of synthetic content

use and its impact on individual rights, including the consequences of deploying such content for financial gain by perpetrators.

Synthetic content refers to deliberately created or manipulated text, images, audio, and/or video generated by artificial intelligence algorithms, which do not reflect actual events, behaviors of specific individuals, or authentic sources of information. Such content is frequently used to influence public opinion, discredit public figures, or obtain financial benefits through deception. According to Fisher (2024), *synthetic content is highly realistic and often indistinguishable from authentic material, highlighting the high degree of credibility achieved by AI-generated images, audio, and video content*.

The term deepfake is a portmanteau of deep learning and fake, used to denote synthetic multimedia content generated through deep learning algorithms. The term first appeared in 2017 on the internet forum Reddit, where a user under the pseudonym deepfakes posted pornographic videos created using generative neural networks. Among the manipulated figures featured was actress Gal Gadot, who is considered one of the first public victims of AI systems capable of generating synthetic content.

SOCIAL AND CRIMINAL ASPECTS OF DEEPFAKE

Several examples illustrate the capabilities of generative artificial intelligence models and the challenges they pose to legal science and practice – namely, the emergence of new forms of criminal behavior and the pressing need to regulate social relations involving the application of AI systems in civil and commercial contexts.

The first example demonstrates how these technologies can influence consumer decision-making when used in advertising campaigns. In 2021, the Spanish brand Cruzcampo employed deepfake tools to “resurrect” Lola Flores – one of the most iconic figures in Spanish flamenco and a cultural symbol – for its advertising campaign. The campaign showcased the advancements in synthetic content generation technologies demonstrating their ability to create content virtually indistinguishable from authentic material – producing illusions that can easily deceive consumers. Although such content may initially appear harmless, when created with the intent to misinform, manipulate public opinion, or generate financial gain, it becomes a powerful weapon in the hands of individuals or organizations pursuing criminal objectives – polluting the integrity of cyberspace and undermining public trust in technology.

The second example highlights a new form of criminal behavior involving synthetic audio content. In 2019, the United Kingdom recorded the first financial fraud case using AI-based voice generation software. Cybercriminals mimicked the voice of the CEO of the German energy company RWE AG in a phone call, instructing the managing director of its

UK subsidiary – RWE Npower Ltd. – to authorize an urgent transfer of funds. The amount, €220,000, was transferred to a bank account controlled by the perpetrators. The generated audio was so precise that it replicated not only the CEO's voice and tone but also his German accent and distinctive intonation effectively deceiving the managing director of RWE Npower. This case demonstrates how generative models can be exploited to create synthetic content for criminal purposes, resulting in significant economic damage. At the same time, it underscores the critical need to cultivate user-oriented critical thinking that questions all incoming information and actively seeks reliable sources to verify or compare any content encountered online.

These examples indicate the urgent need for legislative reform and the development of an effective and adaptive regulatory framework capable of adequately addressing the growing number of crimes committed through deepfake content. Such a framework should also aim to protect consumers from misleading information that may influence their decisions when choosing a product or service. *Law must reflect the evolving dynamics of social relations in a timely manner, and the functions of criminal law remain fundamental in ensuring societal security* (Lyubenov, 2024).

THE EUROPEAN LEGAL RESPONSE: THE ARTIFICIAL INTELLIGENCE ACT

With the adoption of Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 establishing harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), binding rules have been introduced concerning the development and use of artificial intelligence systems, including the generation of synthetic content – deepfakes. The Artificial Intelligence Act represents the *legislative response to the growing technological advancement of tools used in the creation of deepfake content* (Yoon, S., 2024). The average consumer is increasingly unable to distinguish between authentic and synthetic content, which necessitates the adoption of clear rules on transparency, labelling, and traceability of such content.

As EU regulations are directly applicable, the AI Act introduces a unified legal definition of the term "deepfake," which is binding upon the national legislation of Member States in order to avoid inconsistent interpretation. The preamble of the Act emphasizes that AI systems can generate large volumes of synthetic content that are difficult to distinguish from authentic material, thereby posing risks of manipulation, fraud, and erosion of trust in the information environment. The Regulation sets out legal obligations for providers and deployers of generative AI systems to

implement technical solutions for the traceability and labelling of synthetic content – such as watermarks, metadata, or cryptographic tools – in order to ensure that users are informed about the synthetic nature and origin of such content. Any synthetic content that resembles real people, events, or objects– i.e., deepfakes – must be clearly labeled as AI-generated. This disclosure requirement is closely tied to the broader principle of transparency and the user's right to be informed when encountering "persuasive forgeries" or deepfake materials.

However, the Regulation allows exceptions to these rules in cases where the synthetic content has a creative, satirical, or artistic purpose or has undergone human editing. In such instances, the obligation of transparency remains: the creator of the work must disclose the use of an AI system in the generation process, without restricting the display or creative sharing of the work. Pursuant to Article 50 of the Regulation, providers or deployers of AI systems must always indicate when users interact with an artificial intelligence system and must label deepfake content as synthetically generated, specifying that it does not reflect real events, facts, and/or individuals. It is a common misconception that deepfake content includes only synthetic faces or voices, while in fact it may also comprise images of objects, events, or even simulate attacks or military conflicts – each potentially used for manipulation.

In light of these considerations and the potential for cyberspace to be flooded with false, manipulative synthetic content, the Regulation provides a legal definition of "deepfake" as: "synthetically generated or manipulated image, audio or video content that resembles existing persons, objects, places, entities, or events and creates a misleading impression to a person that it is authentic or truthful" (Article 3, paragraph 60 of the Artificial Intelligence Act). This definition is intended to ensure interpretive clarity and uniform application of the concept across the EU.

A potential issue arises in the interpretation of permissible exceptions, which may be exploited as a means of circumventing the transparency and disclosure requirements. Can satirical content cross the boundaries of what is acceptable and socially appropriate, or does freedom of expression inherently permit such transgressions, including the unrestricted creation of synthetic content? The Artificial Intelligence Act is built upon the principle of information authenticity; however, the potential for its compromise – through the generation and dissemination of synthetic content – undermines users' trust in the integrity of cyberspace and in the advances of information technologies.

In this context, permitted exceptions may become a tool for manipulation and shaping public opinion, disguised under the label of "satire" or a particular form of art. One illustrative social experiment is the deepfake video created by American filmmaker Jordan Peele, who used this

technology to deliver a message to internet users by generating synthetic content in the image and voice of former President Barack Obama. In the video, "Obama" expresses opinions on controversial policies and publicly insults President Donald Trump. The purpose of the video was to raise awareness about the potential of deepfake technologies to manipulate public opinion and influence electoral attitudes. While seemingly harmless, the experiment starkly demonstrates that we now *live in a world in which we cannot unconditionally trust what we see or hear, and in which our perception of reality must be critically re-evaluated* (Whittaker, Kietzmann, Kietzmann, & Dabirian, 2020).

CRITICAL THINKING AND MEDIA LITERACY

Within the framework of critical thinking, Staab (2023) emphasizes the necessity for citizens to develop new media literacy skills, including the ability to analyze and verify information flows when confronted with synthetic content. Critical thinking emerges as a key competence that enables individuals to identify manipulated visual and audio materials. According to Staab (2023), *deepfake content has the potential to trigger what is referred to as epistemic chaos, whereby users may either begin to question their entire perception of reality or, conversely, accept synthetic content as true without scrutiny* – especially when it aligns with their pre-existing beliefs and attitudes. The creation of deepfake materials can foster a sense of insecurity within segments of society, which in turn may generate risks of undermining public order, societal security, and political stability (Varbanova, G., 2024).

The use of synthetic content in election campaigns and the unethical conduct of political opponents have brought to the forefront the necessity of legislative regulation of public relations concerning the deployment of deepfake content during electoral processes and political campaigns. Various countries have undertaken legislative initiatives aimed at restricting and preventing the use of synthetic content generated for the purpose of distorting electoral processes through artificially created political messages and altering voters' perceptions of candidates, thereby manipulating their attitudes toward the election process.

One example of such reactive legislation is California State Bill AB 2839, adopted in 2024, which criminalizes the creation and dissemination of deepfake content intended to influence electoral attitudes during election campaigns. The law seeks to protect voters from misleading information that could affect their decision-making processes while simultaneously reinforcing the principle of informed choice within a framework of digital democracy. The ban on the use of misleading synthetic content applies to the 120-day period preceding an election. When such content targets electoral officials or the electoral process itself, the restriction is extended to cover an

additional 60 days after the election day. The objective of this measure is to safeguard the integrity of cyberspace during the particularly sensitive period before and immediately after an election, when the information environment plays a crucial role in shaping public trust and voter decision-making.

One weakness of this legislative solution, however, lies in the fact that outside the specified time frame, the restrictions of AB 2839 do not apply. This creates a window for potential manipulation through synthetic content during the earlier phases of the political cycle or pre-campaign period. Some scholars note that synthetic content may also have positive effects, such as increasing voter engagement and encouraging participation in the electoral process (Vinogradova, E. A., 2024).

LEGISLATIVE APPROACHES

The legislative approach and the efforts to preserve the purity of information flows in cyberspace must also be understood in the context of certain sociodemographic and even cultural-civilizational factors (Lutzkanova, S., 2021) specific to individual nations or social groups. *The degree of democratic development and cultural particularities of these groups or societies are directly linked to the way individuals cultivate critical thinking and apply analytical scrutiny to the media content they consume in the digital environment*.

The People's Republic of China has adopted the Provisions on the Administration of Deep Synthesis Internet Information Services, which entered into force on 10 January 2023. The legislation employs the term deep synthesis technology, referring to any technology that uses deep learning and virtual reality combined with generative algorithms to create or modify text, images, audio, video, virtual scenes, and other forms of content. This technology encompasses the generation and/or editing of textual, vocal, and audio content, biometric and non-biometric characteristics in images and/or video, as well as digital symbols or virtual scenes. The definition adopted in Article 23 of the Provisions is not exhaustive, but rather illustrative – listing the most common technologies and forms of synthetic content.

The Provisions are based on the concept of clear labelling of synthetic content, requiring providers of digital information services and platforms to implement mechanisms for the identification and traceability of such content.

In China, huanlian content – face-swapping – has gained popularity, particularly through the use of generative models and reconstructed video. With the growing use of huanlian, various platforms have integrated mechanisms for the generation of synthetic images and audio-visual content, which are both commercialized and monetized within the formal digital market as well as through informal content-sharing channels (De Seta, 2021). Given the widespread use of huanlian, the Provisions impose a strict

requirement that, when generating synthetic content involving the likeness or voice of an individual, prior consent from the person whose image or voice is being processed is required by law.

A similar approach – requiring the consent of the individual who is the subject of synthetic content – has been adopted in the proposed Take It Down Act, introduced by Senator Ted Cruz and supported by former First Lady Melania Trump. The bill seeks to criminalize not only the publication but also the mere threat of publishing intimate or pornographic images created using deepfake technologies without the consent of the person depicted. The draft legislation establishes a protective mechanism for victims of synthetic content of an intimate or sexual nature, requiring platforms or content providers to remove the material within 48 hours of receiving notice from the affected individual.

The conducted legislative analysis reveals that even states with different political systems recognize the challenges posed by emerging technologies, as well as the ethical and legal risks arising from the use of synthetic content. Despite differences in scope and implementation, the examined legal instruments share a common objective – to preserve human dignity, to protect democratic processes, and to limit infringements on individual rights within cyberspace.

CONCLUSION

The rapid development of artificial intelligence and the growing accessibility of tools for generating synthetic content pose serious challenges to cybersecurity and the legal framework designed to safeguard public trust and democratic values. Deepfake technologies not only distort the perception of reality but also give rise to new forms of criminal behavior that existing legal systems often struggle to address effectively.

This study has demonstrated that the legal response should not be limited solely to restrictive and reactive measures. Alongside legislative reforms, it is crucial to foster a culture of critical thinking and epistemic vigilance within society – one that enhances resilience against manipulation, disinformation, and the increasing proliferation of synthetic content in cyberspace. The European Union's Artificial Intelligence Act serves as a valuable example of how transparency and accountability can be introduced into the creation and dissemination of synthetic content. However, the success of this regulatory framework will ultimately depend on its practical implementation by Member States and the manner in which its restrictive provisions are interpreted.

Diverse legal systems must unite around a common objective: to adapt to technological evolution not only by regulating emerging risks, but also by upholding the core principles of democracy, informed public discourse, and trust in digital communication. The fight against threats posed by deepfake

content is not solely a legal matter; it is a cultural and ethical imperative for the future of cybersecurity and the preservation of information integrity in cyberspace.

BIBLIOGRAPHY:

Чаушев, Х. (2024). Генеративен AI и съдържание в социалните медии – комуникационни предизвикателства в сферата на сигурността. *Сигурност и отбрана,* (2), 179-187. https://doi.org/10.70265/YVKC6923 // Chaushev, H. (2024). Generativen AI i sadarzhanie v sotsialnite medii – komunikatsionni predizvikatelstva v sferata na sigurnostta. *Sigurnost i otbrana*, (2), 179–187. https://doi.org/10.70265/YVKC6923

De Seta, G. (2021). Huanlian, or changing faces: Deepfakes on Chinese digital media platforms. *Convergence: The International Journal of Research into New Media Technologies, 27* (4), 935–953. https://doi.org/10.1177/13548565211030185

Fisher, S. A. (2024). Something AI should tell you – The case for labelling synthetic content. *Journal of Applied Philosophy*. https://doi.org/10.1111/japp.12758

Lyubenov, L. (2024). Domestic legal measures for increasing security in the Black Sea region. *Security and Defense,* (2), 9–18. https://doi.org/10.70265/VIRM1297

Lutzkanova, S. (2021). The regionalization of European security and defense policies in the maritime domain: The Black Sea case. *Pedagogical Almanac, 7*(S), 18. https://doi.org/10.53656/ped21-7s.18def

Staab, L. C. (2023). Making sense of deepfakes: Epistemic harms and the EU policy response [Master's thesis, University of Twente]. University of Twente Student Theses. https://essay.utwente.nl/97795

Varbanova, G. (2024). The significance of electronic evidence in the context of cybersecurity and national security. Varna: Print Master.

Vinogradova, E. A. (2024). Potential threats from unauthorized use of political deepfakes during elections: International experience. *World Politics, 3*, 44–60. https://doi.org/10.25136/2409-8671.2024.3.71519

Whittaker, L., Kietzmann, T., Kietzmann, J., & Dabirian, A. (2020). "All around me are synthetic faces": The mad world of AI-generated media. *IT Professional, 22*, 90–99. https://doi.org/10.1109/MITP.2020.2985492

Yoon, S. (2024). A study on civil liability in the so-called 'weak artificial intelligence' area. *The Korean Association of Civil Law. 106*, 171-201. https://doi.org/10.52554/kjcl.2024.106.171