UDC 343.98:004.89
DOI https://doi.org/10.32849/2663-5313/2024.3.09

Oleksii Shamov,
Intelligent systems researcher, head of Human Rights Educational Guild, 40/28, Rizdvyana Street, Cherkasy, Ukraine, postal code 18001, yursprava@gmail.com
ORCID: *orcid.org/0009-0009-5001-0526*

Shamov, Oleksii (2024). From Synthetic Reality to Judicial Reality: Procedural Challenges of "Deepfakes" and Their Solutions in the Age of Artificial Intelligence. *Entrepreneurship, Economy and Law*, 3, 51–55, doi https://doi.org/10.32849/2663-5313/2024.3.09

FROM SYNTHETIC REALITY TO JUDICIAL REALITY: PROCEDURAL CHALLENGES OF "DEEPFAKES" AND THEIR SOLUTIONS IN THE AGE OF ARTIFICIAL INTELLIGENCE

Abstract. ***Purpose.*** The modern justice system is facing an existential challenge posed by the rapid development of generative artificial intelligence (AI) and its ability to create hyper-realistic audio and video materials, known as "deepfakes." The capability to fabricate compelling yet entirely false evidence, such as video confessions, audio recordings of conversations, or photographs from a crime scene, undermines the fundamental principle of evidence law the reliability and authenticity of proof. This threat is twofold. On one hand, deepfakes can be used to wrongfully accuse innocent individuals or to help the guilty evade responsibility. On the other hand, a mirror phenomenon arises the "Liar's Dividend," where a party to a proceeding can discredit entirely authentic digital evidence by baselessly claiming it is an AI-generated forgery. This creates informational chaos, erodes trust in the evidentiary basis and the justice system as a whole, and presents the legal community with an urgent task of developing adequate countermeasures. The purpose of this article is to propose a comprehensive procedural and evidentiary model aimed at enhancing the justice system's resilience to digital evidence manipulation, based on an analysis of the technical, legal, and ethical aspects of the deepfake problem in litigation. This purpose is achieved by addressing the following objectives: analyzing the limitations of existing rules for evidence authentication; examining the relevant provisions of the EU AI Act and their insufficiency for resolving the problem in judicial proceedings; and formulating an innovative hypothesis for a combined countermeasure mechanism that balances the rights of the parties and ensures access to justice.

Research methods. The methodological basis of the article comprises general scientific and special legal methods. The formal-legal method was applied to analyze the norms of evidence law, particularly the U.S. Federal Rules of Evidence and the EU AI Act. The comparative-legal method allowed for the juxtaposition of regulatory approaches in different legal systems. The system-structural method was used to develop a comprehensive countermeasure mechanism that combines procedural, evidentiary, and financial elements. The methods of analysis and synthesis were employed to process scholarly publications and formulate coherent conclusions.

Results. The article proves that neither traditional rules of evidence authentication nor new legislation like the EU AI Act can fully address the problem of deepfakes in judicial proceedings. The EU Act, while establishing important transparency obligations, relies on the good faith of content creators and fails to counter the malicious use of deepfakes to falsify evidence. The principal outcome of the research is the development of an original three-component model:

1) a two-tiered standard of authentication with a burden-shifting framework, which is activated only after the challenging party provides a minimal, good-faith basis for its doubts; 2) strengthening the role of the judge as an active "gatekeeper of evidence" who decides the issue of authenticity at a preliminary stage; 3) implementing a flexible mechanism for allocating the high costs of digital forensics to ensure equal access to justice.

Conclusion. The current evidentiary paradigm requires urgent adaptation.

The proposed comprehensive model, which combines evidentiary, procedural, and financial incentives, is more effective than attempts to solve the problem solely through technology or piecemeal legislative changes. The author proposes: 1) at the EU and national levels, to initiate the development of harmonized procedural norms that implement the proposed model; 2) for professional legal associations, such as the Council of Bars and Law Societies of Europe (CCBE), to develop ethical guidelines on countering the "Liar's Dividend"; 3) for judicial training institutions, such as the European Judicial Training Network (EJTN), to introduce specialized training for judges and prosecutors on handling digital evidence in the age of AI.

Key words: deepfakes, artificial intelligence, evidence authentication, admissibility of evidence, Liar's Dividend, EU AI Act, judicial proof.

1. Introduction

The Fourth Industrial Revolution, unfolding before our eyes, has brought not only technological progress but also new, previously unimaginable threats to fundamental societal institutions. One of the most vulnerable has been the institution of justice, the foundation of which is trust in facts and evidence. The emergence and rapid proliferation of generative artificial intelligence (AI) and, in particular, "deepfake" technology, have heralded a new era where the line between reality and its synthetic imitation is becoming increasingly blurred (Vig, 2024). The ability of AI to generate photos, audio, and video that are virtually indistinguishable from real ones creates a perfect tool for manipulation and falsification in judicial proceedings.

The problem is not limited to the risk of fabricated evidence appearing in a case file. Equally dangerous is the reverse effect, aptly termed by scholars as the "Liar's Dividen". This term describes a situation where the very existence of deepfake technology allows unscrupulous litigants to cast doubt on any genuine digital evidence by baselessly claiming it is artificial (Citron & Chesney, 2019). As a result, judges and juries find themselves in a state of profound uncertainty, which can lead to a "reverse CSI effect" a total skepticism towards all digital evidence, undermining trust in the justice system itself. This dual challenge presents legal science and practice with the important task of developing a new evidentiary paradigm capable of adapting to the realities of the digital age and protecting the truth-finding process from manipulation.

2. Analysis of Recent Research and Publications

The issue of deepfakes' impact on the judiciary is actively discussed in foreign, predominantly American, legal doctrine. Researchers have broadly divided into two camps.

Proponents of the first, such as Riana Pfefferkorn, believe that existing procedural norms, particularly the U.S. Federal Rules of Evidence (FRE), are flexible enough to counter new threats. They emphasize that the combination of authentication rules (FRE 901), judicial oversight (FRE 104, 403), and the adversarial process (cross-examination, expert witnesses) creates a sufficient filter (Pfefferkorn, 2021).

In contrast, another group of scholars, including Jim Hilbert, argues that the unique nature of deepfakes, which perfectly imitate reality, requires the creation of new, specialized rules.

Their proposals range from raising the standard of proof for authentication to transferring the authority to determine authenticity from the jury to the judge (Sohrawardi & others, 2020; Hilbert, 2019).

Of particular note is the so-called “Deepfake Defence”, which poses a serious ethical and evidentiary dilemma for judges (Dixon, 2024).

Rebecca Delfino's work deserves special mention, as he focuses on the problem of access to justice, arguing that the high cost of digital forensics creates a "pay-to-play" system, where only wealthy parties can afford to effectively prove or disprove the authenticity of evidence (Delfino, 2024).

At the same time, despite the depth of analysis of individual aspects, the problem of creating a comprehensive mechanism that would simultaneously counter the direct use of deepfakes, the abuse of the "Liar's Dividend," and the problem of unequal access to justice remains unresolved. Furthermore, most research focuses on the U.S. legal system, while the analysis of the latest European legislation, particularly the EU AI Act, and its relationship with evidence law remains fragmented. This article is dedicated to addressing this complex problem and filling the indicated gap.

3. Purpose and Objectives of the Scientific Research

The purpose of this research is to develop and substantiate a comprehensive procedural and evidentiary model aimed at neutralizing the threats posed by "deepfakes" and AI-generated disinformation to the justice system.

To achieve this purpose, the following objectives were set:

1. To analyze the technical nature of deepfakes and identify the key challenges they pose to traditional methods of evidence authentication.
2. To examine existing legal norms and judicial practice (using the U.S. as an example regarding the handling of challenged digital evidence and to identify their weaknesses.
3. To analyze the provisions of the EU Artificial Intelligence Act concerning deepfakes and determine the limits of its applicability in the context of judicial proof.
4. To formulate and substantiate an innovative three-component model of procedural response that combines evidentiary, procedural, and financial elements.
5. To develop specific recommendations for European and national legislators, professional legal associations, and judicial training institutions on the implementation of the proposed model.

4. Scientific Methods Used

The methodological basis of the research consists of a system of philosophical-worldview, general scientific, and special legal methods of cognition. The dialectical method allowed for the examination of the deepfake problem in its development and the interconnection of its technical, legal, and ethical aspects. The for-

mal-legal method was used to analyze the content of legal norms, particularly the U.S. Federal Rules of Evidence and the EU AI Act. The comparative-legal method was useful in contrasting regulatory approaches and judicial practices in different jurisdictions. The system-structural method formed the basis for developing a holistic model for countering deepfakes, consisting of interconnected elements. The methods of analysis and synthesis were employed to process scholarly sources, identify key ideas, and generalize them in the conclusions and proposals.

5. Presentation of the Main Research Material

The essence of the technical challenge lies in the continuous improvement of the generative adversarial networks (GANs) underlying deepfakes. Each new iteration of the algorithms learns from the mistakes of the previous ones, making passive detection methods (searching for visual artifacts, anomalies) increasingly unreliable. This creates a situation of a constant "arms race," where the legal system, being inherently more inert, risks always being one step behind the technologies of falsification.

For the law, this means that relying solely on an expert's technical opinion regarding the authenticity of a recording becomes dangerous. If today an expert can detect a forgery by a barely noticeable flicker of the skin around the eyes, tomorrow a new algorithm will learn to imitate this aspect as well. Therefore, the legal response should focus not on finding a silver bullet" in the form of a perfect detector, but on creating a robust procedural framework that can withstand the pressure of technological uncertainty. This procedure must ask the right questions: not only "Is this recording a fake?" but also "What is the origin of this file?", "Is there an unbroken chain of its custody?", "What motives and opportunities did the party presenting it have to alter it?".

The EU Artificial Intelligence Act is a revolutionary step in regulating technology, but its architecture is primarily aimed at market regulation and consumer protection, not the reform of evidence law (European Parliament, 2024). The key provision, Article 50, which requires the labeling of deepfakes, is based on a presumption of good faith. It is effective against pranksters, marketers, or even some media outlets, but it is completely powerless against a person who purposely creates false evidence for use in a criminal proceeding to evade punishment for a serious crime. Such a malicious actor will never voluntarily label their creation as a deepfake.

The classification of AI systems for justice as "high-risk" is also an important but insufficient step. It regulates the use of AI by judicial authorities (e.g., for case analysis), but it does not address the situation where AI-generated content enters the process from the outside, as evidence from one of the parties.

Thus, the Act creates an important foundation but leaves a critical gap concerning the adversarial process of proving authenticity. This gap must be filled by special procedural rules.

To fill this gap and create an effective countermeasure system, a three-component model is proposed, which should be implemented through harmonized legislation.

The need to combine technological and legal approaches for the authentication of evidence is a central thesis of modern research (Goldstein & Lohn, 2024).

Component 1: A Two-Tiered Standard of Authentication with a Burden-Shifting Framework.

This mechanism is designed to strike a balance between preventing forgeries and protecting against baseless accusations of fakery.

First Tier (Basic Authentication): The party submitting digital evidence (video, audio) performs the standard authentication procedure. This could be the testimony of the person who made the recording or another person who can confirm that the recording is a fair and accurate representation of the events. At this stage, a presumption of authenticity applies.

– Activation of the Second Tier: The opposing party may challenge the evidence, claiming it is a deepfake. However, to prevent the abuse of the "Liar's Dividend", a simple assertion is insufficient. The challenging party is required to provide the court with a minimal, "good faith" evidentiary basis for its doubts. This does not require a full expert report but must be more than a mere allegation. Examples of such a basis could include:

– A preliminary report from a technical specialist pointing to specific visual or audio anomalies.

– Evidence of an alibi for the person depicted in the video, making their presence at the specified time and place highly improbable.

– An indication that the file shows signs of editing, has inconsistent metadata, or has breaks in the chain of custody.

– Second Tier (Heightened Authentication): Only after the court finds that such a good-faith basis exists does the burden of proving authenticity shift to the party that submitted the evidence. However, the standard of proof is now significantly higher. The party must not only provide a witness but also likely engage an expert to confirm the file's integrity and provide evidence of an unbroken chain of custody from the moment of its creation to its submission in court.

Component 2: Strengthening the Judge's Role as an Active "Gatekeeper of Evidence".

The question of the authenticity of challenged digital evidence, given its technical complexity and potential to confuse the jury, should be decided exclusively by the judge in a preliminary hearing. This will allow for the filtering out of baseless claims and prevent evidence of questionable authenticity from being considered by the jury. The judge must be given clear authority to assess not only the evidence itself but also the reliability of the methods and technologies used by experts for its analysis, guided by criteria of scientific validity.

Component 3: A Flexible Mechanism for Allocating the Costs of Expertise.

Digital forensic expertise is extremely expensive, creating significant inequality between parties with different financial capabilities. To ensure genuine access to justice, courts must be empowered to flexibly allocate the costs of such expertise. Instead of automatically imposing the costs on the losing party or the party that initiated the expert examination, the court should consider a range of factors:

– The financial situation of the parties.

– The reasonableness of the claim of forgery.

– The outcome of the authenticity determination.

For example, if a party's claim that a piece of evidence is a deepfake proves to be justified, the court may impose the costs of the expertise on the opposing party, which attempted to mislead the court. Conversely, if the claim is found to be baseless and shows signs of abuse of rights, all related costs are imposed on the initiator of such a claim. This will create a powerful economic incentive for the parties to act in good faith.

6. Conclusions

This research concludes that the threat posed by "deepfakes" to the justice system is systemic and therefore requires an equally systemic response. Piecemeal technological solutions or minor legislative changes cannot fully solve the problem. Only a comprehensive approach that combines procedural rules, evidentiary standards, and economic incentives can create a resilient and adaptive justice system.

The proposed three-component model (a two-tiered standard of authentication, an enhanced role for the judge, and flexible cost allocation) constitutes such a comprehensive solution. It is aimed not at a complete ban or ignorance of digital evidence but at creating a reliable procedure for its verification that protects against both forgeries and baseless accusations, while ensuring equal access to justice.

Prospects for further research in this area are multifaceted. First, empirical research is needed to examine the impact of deepfakes and the "Liar's Dividend" on how professional judges and juries in different jurisdictions perceive and evaluate evidence. Second, it is essential to develop specific methodological recommendations and training programs for judges, prosecutors, and lawyers based on the proposed model. Third, an important area is the further analysis of international cooperation in the exchange and mutual recognition of digital evidence, particularly in a context where the risk of forgery is transnational. Finally, the development of authentication technologies, such as standardized digital watermarks and blockchain systems for recording provenance, requires constant legal monitoring for their possible integration into evidence law.

References:

Delfino, R. (2024). Pay-to-Play: Access to Justice in the Era of AI and Deepfakes, *Loyola Law School, Los Angeles Legal Studies Research Paper No. 2024-08*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4722364

Sohrawardi, S., Seng, S., Chintha, A., Thái, B., Ptucha, R., Wright, M. & Hickerson, A. (2020). DeFaking Deepfakes: Understanding Journalists' Needs for Deepfake Detection. *ResearchGate Website*. Retrieved from https://www.researchgate.net/publication/353523710_DeFaking_Deepfakes_Understanding_Journalists'_Needs_for_Deepfake_Detection

European Parliament. (2024). EU AI Act: first regulation on artificial intelligence. *Official Website of the European Parliament*. Retrieved from https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

Hilbert, J. (2019). The Disappointing History of Science in the Courtroom: Frye, Daubert, and the Ongoing Crisis of "Junk Science" in Criminal Trials. *Oklahoma Law Review*, 2019, Volume 71, Number 3. Retrieved from https://digitalcommons.law.ou.edu/cgi/

viewcontent.cgi?article=1360&context=olr

Dixon, H. (2024). The "Deepfake Defense": An Evidentiary Conundrum. *ABA Judges Journal*, 2024, Vol. 63, No. 2, pp. 38-40. Retrieved from https://www.americanbar.org/content/dam/aba/publications/judges_journal/vol63no2-jj2024-tech.pdf

Pfefferkorn, R. (2020). "Deepfakes" in the Courtroom. Silicon Flatirons Center Report, *University of Colorado Law School*, pp.245-276. Retrieved from https://siliconflatirons.org/wp-content/uploads/2021/02/Pfefferkorn.pdf

Citron, D. & Chesney, B. (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California law review*, 2019, Volume 107. Retrieved from https://www.californialawreview.org/print/deep-fakes-a-looming-challenge-for-privacy-democracy-and-national-security

Goldstein, J. & Lohn, A. (2024). Deepfakes, Elections, and Shrinking the Liar's Dividend. *Brennan Center For Justice Website*. Retrieved from https://www.brennancenter.org/our-work/research-reports/deepfakes-elections-and-shrinking-liars-dividend

Vig, S. (2024). Regulating Deepfakes: An Indian perspective. *Journal of Strategic Security*, 2024, Vol. 17, No. 3, pp. 70-93. Retrieved from https://www.jstor.org/stable/48793910

Олексій Шамов,
дослідник інтелектуальних систем, голова ГО "Просвітня фундація з прав людини", вулиця Різдвяна, 40/28, Черкаси, Україна, індекс 18001, yursprava@gmail.com
ORCID: *orcid.org/0009-0009-5001-0526*

ВІД СИНТЕТИЧНОЇ РЕАЛЬНОСТІ ДО СУДОВОЇ ДІЙСНОСТІ: ПРОЦЕСУАЛЬНІ ВИКЛИКИ «ГЛИБИННИХ ФЕЙКІВ» ТА ШЛЯХИ ЇХ ВИРІШЕННЯ В ЕПОХУ ШТУЧНОГО ІНТЕЛЕКТУ

Анотація. Сучасна система правосуддя зіткнулася з екзистенційним викликом, зумовленим стрімким розвитком генеративного штучного інтелекту (ШІ) та його здатністю створювати гіперреалістичні аудіо- та відеоматеріали, відомі як «глибинні фейки» (deepfakes). Можливість сфабрикувати переконливі, проте цілком неправдиві докази, такі як відео зізнання, аудіозаписи розмов чи фотографії з місця злочину, руйнує фундаментальний принцип доказового права - достовірність та автентичність доказів. Ця загроза є подвійною. З одного боку, діпфейки можуть бути використані для безпідставного обвинувачення невинних осіб або для уникнення відповідальності винними. З іншого боку, виникає дзеркальний феномен — «дивіденд брехуна» (Liar's Dividend), коли сторона процесу може дискредитувати цілком автентичні цифрові докази, безпідставно заявляючи, що вони підробкою, створеною ШІ. Це створює інформаційний хаос, підриває довіру до доказової бази та системи правосуддя загалом, ставлячи перед юридичною спільнотою нагальне завдання з розробки адекватних механізмів протидії.

Мета цієї статті полягає в тому, щоб на основі аналізу технічних, правових та етичних аспектів проблеми діпфейків у судочинстві запропонувати комплексну процесуально-доказову модель, спрямовану на підвищення стійкості системи правосуддя до маніпуляцій із цифровими доказами. Мета досягається через вирішення завдань: аналіз обмеженості існуючих правил автентифікації доказів; дослідження релевантних положень Акту ЄС про штучний інтелект та їх недостатності для вирішення проблеми в судовому процесі; формулювання інноваційної гіпотези щодо комбінованого механізму протидії, який збалансовує права сторін та забезпечує доступ до правосуддя.

Методи. Методологічну основу статті склали загальнонаукові та спеціально-юридичні методи. Формально-юридичний метод застосовано для аналізу норм доказового права, зокрема Федеральних правил доказування США та Акту ЄС про ШІ. Порівняльно-правовий метод дозволив зіставити підходи до регулювання в різних правових системах. Системно-структурний метод використано для розробки комплексного механізму протидії, що поєднує процесуальні, доказові та фінансові елементи. Методи аналізу та синтезу застосовувались для опрацювання наукових публікацій та формування цілісних висновків.

Результати. У статті доведено, що ані традиційні правила автентифікації доказів, ані нове законодавство, як-от

Акт ЄС про ШІ, не здатні повною мірою вирішити проблему діпфейків у судовому процесі. Акт ЄС, хоч і встановлює важливі зобов'язання щодо прозорості, покладається на добросовісність творців контенту і не протидіє зловмисному використанню дінфейків для фальсифікації доказів.

Основний результат дослідження полягає у розробці авторської трикомпонентної моделі: 1) дворівневий стандарт автентифікації з механізмом перекладання тягаря доведення, що активується лише після надання стороною-опонентом мінімального добросовісного обґрунтування своїх сумнівів; 2) посилення ролі судді як активного «охоронця доказів», який вирішує питання автентичності на досудовому етапі; 3) впровадження гнучкого механізму розподілу високих витрат на цифрову експертизу для забезпечення рівного доступу до правосуддя.

Висновки. Сучасна доказова парадигма потребує термінової адаптації. Запропонована комплексна модель, що поєднує доказові, процесуальні та фінансові стимули, є більш ефективною, ніж спроби вирішити проблему виключно через технології чи точкові зміни до законодавства. Автором пропонується: 1) на рівні ЄС та національних юрисдикцій розпочати розробку гармонізованих процесуальних норм, що імплементують запропоновану модель; 2) професійним юридичним об'єднанням, таким як Рада адвокатських та правничих товариств Європи (CCBE), розробити етичні настанови щодо протидії «дивіденду брехуна»; 3) судовим навчальним закладам, таким як Європейська мережа підготовки суддів (EJTN), впровадити спеціалізовані тренінги для суддів та прокурорів щодо роботи з цифровими доказами в епоху ШІ.

Ключові слова: глибинні фейки, штучний інтелект, автентифікація доказів, допустимість доказів, дивіденд брехуна, Акт ЄС про ШІ, судове доказування.

The article was submitted 13.11.2024
The article was revised 04.12.2024
The article was accepted 24.12.2024