ACIG
APPLIED CYBERSECURITY & INTERNET GOVERNANCE

NASK

Deepfake Influence Tactics through the Lens of Cialdini's Principles: Case Studies and the DEEP FRAME Tool Proposal

Pawel Zegarow | Strategy and Development of Cyberspace Security Team, NASK – National Research Institute, Poland | ORCID: 0009-0006-5421-8656

Ewelina Bartuzi | Audiovisual Analysis and Biometric Systems Department, NASK – National Research Institute, Poland | ORCID: 0000-0001-6245-2908

Abstract

The advancement of artificial intelligence (AI) has introduced both significant opportunities and challenges, with deepfake technology exemplifying the dual nature of AI's impact. On the one hand, it enables innovative applications; on the other, it poses severe ethical and security risks. Deepfakes exploit human psychological vulnerabilities to manipulate perceptions, emotions, and behaviours, raising concerns about the public's ability to distinguish authentic content from manipulated material. This study examines the methods of influence embedded in deepfake content through the lens of Robert Cialdini's six principles of persuasion. By systematically analysing how these mechanisms are employed in deepfakes, the research highlights their persuasive impact on human behaviour, particularly in scenarios such as financial fraud. To address the challenges posed by deepfake technology, this study introduces DEEP FRAME, an original tool for systematically recording and analysing deepfake content. DEEP FRAME integrates technical and psychological analysis, enabling the identification of technological characteristics and manipulation strategies embedded within deepfakes. The findings underscore the need for a holistic and interdisciplinary approach that combines technological

Received: 17.09.2024

Accepted: 13.12.2024

Published: 30.12.2024

Cite this article as:
P. Zegarow, E. Bartuzi, "Deepfake influence tactics through the lens of Cialdini's principles: Case studies and the DEEP FRAME tool proposal," ACIG, vol. 3, no. 2, 2024, pp. 286–302. DOI: 10.60097/ACIG/201147

Corresponding author:
Pawel Zegarow, Strategy and Development of Cyberspace Security Team, NASK – National Research Institute, Poland; E-mail: pawel.zegarow@nask.pl
0009-0006-5421-8656

Copyright:
Some rights reserved (CC-BY):
Pawel Zegarow, Ewelina Bartuzi
Publisher NASK

OPEN ACCESS

CC BY

innovation, psychological insights, and legal frameworks to counter the growing threat of deepfakes.

Keywords

social engineering, cybersecurity, cyberpsychology, deepfake, DEEP FRAME tool

1. Introduction

While artificial intelligence (AI) has the potential to revolutionise key aspects of human life – from work and education to health and security – its development also introduces significant ethical and societal risks. One of these is the emergence of deepfake technology, identified by researchers at University College London in 2020 as one of the most dangerous AI-enabled crimes due to its potential for malicious use [1].

Deepfakes actually pose a significant challenge to researchers and practitioners seeking effective strategies to protect individuals and societies from manipulation. The potential of deepfakes to influence perception and decision-making raises serious concerns about information integrity across a variety of sectors, including politics, media, and the private domain. Despite the critical nature of this problem, existing research has mainly focused on technical aspects, such as detecting fake content. Therefore, our work focuses on an interdisciplinary approach, combining technological and psychological knowledge about the mechanisms of persuasion used in deepfakes. Additionally, we propose the DEEP FRAME tool for the systematic analysis of deepfakes, which takes into account both technical and psychological aspects.

The term 'deepfake' is derived from English and combines two words: 'deep learning' and 'fake'. It refers to a technology that uses AI to generate videos, images, and sounds so realistic that distinguishing it from authentic material becomes challenging. The term 'deepfake' was first used at the end of 2017 by an anonymous Reddit user operating under the pseudonym 'deepfakes'. This individual utilised deep learning methods to create video content in which the faces of performers in adult films were replaced with those of recognisable public figures [1–3].

1.1. The Unexpected Challenges of Detecting Deepfakes

Until recently, content generated by AI was characterised by relatively low quality and contained easily identifiable errors,

such as distorted facial features, unnaturally rendered skin, or hands depicted with more than five fingers. However, advancements in machine learning techniques and the increasing computational power of modern systems have made contemporary deepfakes nearly indistinguishable from authentic content, exhibiting an exceptionally high level of realism.

It is important to note that not all deepfakes produced today achieve hyperrealistic quality, but many of these videos are consumed on mobile devices, where smaller screen sizes and lower resolution settings can mask imperfections and make distinguishing them from real content significantly more challenging.

Detection of deepfakes is a complex issue. Research shows that human ability to recognise deepfakes varies considerably, with detection accuracy ranging from 57% to 89%. This suggests that even in the most optimistic scenario, individuals fail to identify 11 out of every 100 deepfakes, while in the most pessimistic case, as many as 43 out of 100 deepfakes go undetected [4].

1.2. Deepfake Technology as a Tool for Cybercriminal Activity

Although the online dissemination of false content is not a novel phenomenon, the advent of deepfake technology has enabled cybercriminals to engage in malicious activities on an unprecedented scale. Deepfakes are employed in a range of criminal endeavours, including extortion, reputational damage, the manipulation of political processes, disinformation, and financial frauds. According to the assessment conducted by the authors of the publication 'AI-enabled future crime', deepfake technology may inflict the most substantial harm and yield the greatest potential profits for cybercriminals [5].

Cybercriminals frequently integrate deepfake technology with social engineering strategies to enhance the effectiveness of their activities. They exploit various human psychological vulnerabilities, such as the ease of eliciting trust, the propensity to act under time pressure, and susceptibility to suggestion. By doing so, cybercriminals can construct deceptive narratives and influence the behaviours of potential victims, ultimately aiming to gain access to sensitive information.

Deepfake technology poses threats not only to individual users but also to businesses, government institutions, and international organisations. By exploiting human beings' natural tendency to

trust the authenticity of perceived images and sounds, cybercriminals destabilise cognitive processes, affecting emotions, attitudes, and perceptions of reality. As a result, victims of such manipulation often make irrational decisions and engage in behaviours based on false premises, potentially leading to severe personal and professional repercussions.

In March 2022, as reported by the *Reuters* news agency, a deepfake of Ukrainian President Volodymyr Zelensky appeared on the social media platform X, calling on Ukrainian soldiers to lay down their weapons. It can be inferred that the purpose of this deepfake was to undermine morale and create confusion during an active armed conflict. Although the footage was almost immediately identified as inauthentic, its deceptive nature was not immediately apparent to all viewers – particularly older people or those less familiar with technological advances. This example illustrates how dangerous the spread of false content on social media can become in times of armed conflict and highlights the serious social and political repercussions that can result.

In recent times, the Russian Federation's production and large-scale dissemination of deepfake content has become increasingly intense. Merely one day following the terrorist attack of 22 March 2024, a fabricated video emerged in which the image of Oleksiy Danilov, Secretary of the National Security and Defense Council of Ukraine, was integrated into a format resembling a professional television interview. This footage served as part of a broader disinformation campaign aimed at attributing responsibility for the incident to Ukraine. To create this material, archived footage from 16 March 2024 – originally depicting Kyrylo Budanov, the head of the Main Directorate of Intelligence of the Ministry of Defense – was repurposed by substituting his likeness. Despite propagandists' efforts, the quality of the fabricated material remained low. Advanced voice cloning tools and lip sync techniques were employed to synchronise mouth movements with the manipulated statements; however, the final result was far from seamless. Visual distortions, unnatural synchronisation of lip movements with speech, and notable blurring – especially around the neck and mouth – remained discernible to vigilant observers. It must be acknowledged, however, that not all audiences are capable of detecting such subtle signs of interference. A previous attempt at using a similar disinformation strategy was recorded in October 2022, involving voice cloning to gain access to confidential information from the drone manufacturer Bayraktar. In this instance, perpetrators impersonated Ukraine's Prime Minister, Denys Shmyhal,

but their attempt was thwarted. This underscores the need for continuous improvement in methods of detection and countering such threats.

Recently, there has been a marked intensification in the use of fraudulent investment advertisements as tools for extracting personal data and financial resources. Such content appears on social media platforms, websites, and even as sponsored advertising materials. A defining characteristic of these schemes is the promise of rapid profits with minimal risk, often reinforced by the use of recognisable figures from the worlds of politics, business, entertainment, or finance.

Fraudsters increasingly produce complex audiovisual materials whose aesthetics and format resemble those of news programmes. They incorporate the likenesses of well-known presenters and journalists, who appear to endorse 'investment opportunities' or prompt viewers to take specific actions, such as clicking on a link or downloading an application.

According to information published in the British newspaper *The Guardian*, in early February 2024, the Hong Kong police reported that cybercriminals had employed deepfake technology to steal nearly £20 million. An investigation was launched following a report from an employee of a British company operating a branch in China, who informed the police that she had been coerced into transferring a significant sum of money into bank accounts designated by individuals posing as high-ranking company officials. Prior to executing the transfer, the employee had participated in a videoconference with the chief financial officer and other members of the management team. The investigation subsequently revealed that the individuals participating in this meeting were generated by AI [6].

Cybercriminals produce a diverse array of materials, each tailored to distinct target audiences. To promote fictitious investment ventures, they frequently use the likenesses of politicians and business leaders, intending to foster trust among individuals interested in traditional forms of investment. In contrast, content featuring celebrities and influencers are commonly deployed in advertising mobile applications – such as those simulating online casinos or games – primarily attracting younger audiences seeking entertainment. For older or unwell individuals, cybercriminals create deceptive 'miracle' drug or medical procedure advertisements, capitalising on the credibility associated with renowned physicians,

athletes, or religious figures. Among the widely publicised cases reported by the media are as follows:

- A 'gas pipeline investment' scam in which an elderly resident of Lower Silesia lost half a million PLN after believing promises of high returns on a purported investment project [7].
- A crime involving the misappropriation of funds under the guise of investing in Baltic Pipe gas pipeline shares. Victims, enticed by the reputation of this strategic energy project, ultimately lost their life savings [8].
- The case of a 35-year-old woman who invested approximately 150,000 PLN in a fraudulent crypto currency scheme. Exploiting her lack of experience and promising quick and guaranteed profits, the scammers induced significant financial losses [9].

All the above examples underscore the growing significance of fraudulent investment advertisements as tools employed by cybercriminals. Given the substantial social harm caused by such offenses, it is imperative to conduct in-depth research into the mechanisms governing the creation, distribution, and reception of this type of content, as well as to implement effective educational programmes and advanced technological measures. Such actions are crucial for reducing the scale of losses suffered by potential victims and for enhancing security within the digital environment.

It should be emphasised that deepfakes and social engineering are mutually reinforcing phenomena, giving rise to a new generation of threats characterised by a high level of technological sophistication and effectiveness in manipulating human behaviour. Unfortunately, conventional protection methods, which predominantly rely on technological safeguards, have proven insufficient. Consequently, countering deepfakes necessitates an interdisciplinary approach that integrates psychological, technological, and legal expertise.

2. Purpose

This study aimed to analyse the use of Cialdini's persuasion strategies in deepfake videos and present a DEEP FRAME – an original tool for recording and analysing deepfake content.

3. Methods

3.1. Deepfake Selection and Transcription

Given the objectives outlined in this study, a deepfake video observed on social media platforms in Poland between May

and July 2024 was selected to serve as the central case study. The sample was limited to Polish social media platforms to tailor the analysis to the local cultural and social context. The sample selection in our study was intentional. This video was purposefully chosen to ensure that it provided a diverse and comprehensive representation of deepfake-related content, encompassing a wide range of narrative structures, emotional triggers, and psychological attributes. Transcript analysis, therefore, incorporated both qualitative and quantitative methodologies, offering insights into the persuasive techniques at play, including mechanisms rooted in emotional appeal, authority, and social proof.

In the second stage, the researchers conducted the transcription of deepfake videos. The study employed automatic speech recognition (ASR) models to transcribe audio from deepfake videos. These models, often underpinned by large language models (LLMs), are specifically designed to accurately transcribe audio recordings into written text.

It should be emphasised that one of the primary challenges faced by ASR models is their limited ability to process noisy audio recordings. In cases of low-quality audio, where background noise, music, or other interferences are present, transcriptions may be incomplete or inaccurate. Under such conditions, models often generate extraneous elements in the text, such as phrases like 'subtitles sounds...' or other artifacts caused by misinterpretation of background noise. Another significant challenge arises when interpreting words that are mispronounced or articulated in an unusual manner. ASR systems tend to substitute such words with alternatives that better align with the surrounding context. This can result in distortions, particularly when proper nouns, technical terms, or slang expressions are replaced with incorrect equivalents.

Furthermore, transcriptions may contain repetitive words, sentences, or even entire speech segments. These repetitions often stem from the model's uncertainty regarding the interpretation of specific audio fragments or errors in the speech recognition algorithm. ASR models also struggle with recordings that feature shifts in accent or pronunciation. This issue becomes particularly pronounced when the speaker's intonation changes or when an accent characteristic of another language is introduced, such as Polish speech interspersed with Russian or English accents. In such scenarios, models may generate what are referred to as 'linguistic hallucinations,' introducing foreign language fragments into the transcription [10].

3.2. Cialdini's Persuasion Strategies

This study applied Robert Cialdini's six principles of social influence – reciprocity, commitment and consistency, social proof, authority, liking, and scarcity – to analyse deepfake [11]. The deepfake's transcript was analysed by a psychologist, who, drawing on their knowledge of influence techniques and professional experience, classified specific sections of the text as particularly persuasive. This assessment considered key persuasion mechanisms, such as authority, social proof, emotional appeal, and the principles of scarcity and reciprocity.

The principle of reciprocity refers to the innate human tendency to reciprocate benefits received from others, even in the absence of necessity. The analysis focused on identifying elements within deepfake videos that could evoke a sense of obligation or an inclination to reciprocate in the viewer. For instance, the videos may imply exclusivity in the presented information, potentially prompting viewers to reciprocate by further sharing the content.

The principle of commitment and consistency emphasises individuals' tendency to maintain alignment between their actions and decisions over time. The study examined whether the deepfakes employed techniques designed to prompt initial low-commitment actions, such as liking or sharing content, which could subsequently foster greater engagement.

The principle of social proof is based on the influence of others' behaviours on an individual's decision-making process, particularly in new or ambiguous situations, where people tend to follow the actions of others as a guide. The analysis assessed whether the deepfake materials incorporated elements, such as positive comments, or references to perceived broad social support, which could amplify the message by creating the impression that the stance presented is widely accepted and endorsed.

The authority principle is based on the tendency of individuals to place trust in and act upon information provided by perceived experts or leaders. The analysis examined whether the deepfake materials featured prominent figures, such as politicians, scientists, or opinion leaders, leveraging their perceived authority to enhance the credibility and impact of the message.

The liking principle refers to the idea that individuals are more likely to be persuaded by those they perceive as likable or who share similarities with them. The analysis investigated whether the deepfake

materials employed references to shared cultural and social values to strengthen their persuasive impact on recipients.

The principle of scarcity emphasises the perceived value of information by suggesting its limited availability. The study analysed whether the deepfake materials strategically employed techniques implying rarity, exclusivity, or urgency in the message, which could facilitate expedited decision-making or deepen viewer engagement through psychological pressure.

4. Results

4.1. Deepfake Analysis

4.1.1. Deepfake Transcript

You all know me. My name is Rafał Brzoska. I am a professional businessman and investor. Today is your lucky day. This page is available to only 100 people, and you are one of the few who will have the opportunity to make money and change your life. Only the most determined individuals will be able to achieve this. Of the 100 invited, only fifty of the most ambitious will take advantage of my offer. So let's get straight to the point. When I say this will change your life, I don't mean 2000 or 10,000 złotys. I mean an amount that will allow you to quit your job and go home. It's like early retirement or an additional pension – several times larger than your regular savings.

Before you leave this page thinking I'm a complete fool, wait a moment and listen to me. This isn't another video about someone trying to scam you out of your money, because I respect you and want to earn your trust. I won't make empty promises like everyone else. What's the difference between those scammers and me? First, I am Rafał Brzoska, and I don't need anything from you. I will provide you with proof that my project actually works. Promising you millions tomorrow is a lie, just like other empty promises, but four thousand PLN a day is absolutely achievable. Just do the math. Four thousand PLN a day equals 28,000 PLN a week or 1,96,000 PLN a month.

Now listen carefully. This video can only be viewed once. If you leave this page, you won't have another chance to return because your link will expire, as will your opportunity to make money. This has nothing to do with Forex stocks, financial pyramids, or any other nonsense you may see everywhere. I spend most of my life creating projects

that can improve the lives of every individual. The software we develop is better than all competitors thanks to its unique technologies and can be used on any computer or phone. Its unique AI-based analytical features allow it to stay ahead of market trends, ensuring exceptional success across all financial markets. The algorithm does everything for you. All you have to do is watch the results.

We tested our product on a small group of volunteers, each of whom earned over 24,000 PLN within the first week. I don't want to disappoint you, but you can only make 16,000 PLN, which is still a lot of money, isn't it? No, you don't need any special skills. If you're watching this video, it means your device supports this platform. You're probably wondering why I chose you and didn't keep such a unique algorithm for myself if it can generate such significant profits. The answer is simple. I decided to offer this program to 100 random users, but it turned out that only fifty of them would actually try to change their lives. I hope you are among those fifty people who want to become financially independent because you have the chance to try our algorithm for free and change your life.

In return, I ask you to write a short review so that people for whom our program will cost 4000 PLN a month know it really works. The more free clients I get now, the more paying customers I'll have once I start selling the program. But let's get back to the point. Imagine receiving 4000 PLN every day. You'll no longer have to worry about not having enough money to pay your rent. Taking a vacation several times a year? No problem. Buying a house and paying off all your debts? That's no issue. Providing your children with access to the best schools and travelling to the most luxurious destinations will certainly not be a problem. Imagine a life where you don't have to worry about anything.

The most important thing is time. If you start at the right moment, you can earn a lot of money. If you delay, you may end up like 99% of other people. Now the most important information. Pay close attention. To start using the program, you need to visit the website, enter the necessary details, and then your personal manager will contact you to answer all your questions and grant you access to the platform. From that moment, your life will change. You can call it a new life, and I'm sure you won't regret it.

The transcription demonstrates that persuasion mechanisms are systematically and deliberately employed, creating a cohesive and highly effective framework for persuasion. By integrating a nuanced combination of emotional and rational appeals, the deepfake strategically manipulates the audience, fostering a sense of trust, urgency, and a perceived necessity to act. This structured approach highlights the potential of deepfakes to exert significant influence on individual decision-making and behaviour, particularly within the context of digital environments. Table 1 provides a detailed analysis of the transcripts.

4.2. DEEP FRAME tool

In Appendix 1, we propose a DEEP FRAME tool designed for analysts that enables a comprehensive examination of deepfake content through an interdisciplinary approach. DEEP FRAME is a self-reported tool that includes a set of questions about both technical and psychological elements of deepfakes.

The proposed tool offers a wide range of benefits, serving as a valuable resource for interdisciplinary research on deepfake content by integrating both technological and psychological perspectives. The data and analyses it generates have the potential to advance significantly the development of more sophisticated algorithms for detecting manipulated content. Beyond its contributions to research and detection, the tool holds considerable promise for public education, as the insights it provides can support the design and implementation of effective awareness campaigns. Moreover, it strengthens our capacity to understand emerging threats and devise targeted strategies to mitigate their impact. By combining elements from computer science, psychology, and communication studies, it ensures a more thorough evaluation of the subtle and overt manipulations embedded in digital media.

The DEEP FRAME tool facilitates the systematic collection of knowledge about deepfakes. The collected data is categorised based on key parameters, such as the type of manipulation, the technologies used, and the context of publication. A technical module provides analysis of the quality of video and audio, potential artifacts of manipulation. A unique feature of the tool is its ability to conduct psychological evaluations of deepfake content. This module identifies persuasive techniques. Such analyses deepen our understanding of the manipulative mechanisms employed in deepfake content. The DEEP FRAME tool enables the monitoring of global trends in the use of deepfake technology, helping to identify emerging threats.

Table 1. Psychological analysis.

The principle of social influence	Examples	Frequency	Comments
The principle of reciprocity	'I respect you and want to earn your trust'. 'The more free clients I get now, the more paying customers I'll have once I start selling the program'. '...you have the chance to try our algorithm for free and change your life, in return, I ask you to write a short review...'	3	Building commitment by offering a free program and expressing appreciation. Offering something for free may create a sense of obligation in the recipient to write a review in return or take action.
The principle of commitment and consistency	'Before you leave this page thinking I'm a complete fool, wait a moment and listen to me'. 'Only the most determined individuals will be able to achieve this'. 'Of the 100 invited, only 50 of the most ambitious will take advantage of my offer'. 'I hope you are among those 50 people who want to become financially independent'.	4	Creating a sense of uniqueness and the need to act among the chosen ones. The recipient is gradually drawn into the process through initial commitments, such as qualifying for the 'most ambitious' and 'most determined' group. This principle emphasises the need for consistency in action – once the recipient has been selected, they should prove their determination by taking advantage of the offer.
The principle of social proof	We tested our product on a small group of volunteers, each of whom earned over 24,000 PLN within the first week. '... and you are one of the few who will have the opportunity to make money and change your life'.	2	Pointing to the success of other users and the elitism of the group. Social proof elements are visible through references to group tests and the success of other users. Emphasising the 'exceptionality' of the group of 100 people creates a sense of elitism, but also of universal support for this initiative.
The principle of authority	'You all know me. My name is Rafał Brzoska. I am a professional businessman and investor'. 'I am Rafał Brzoska, and I don't need anything from you'. 'The algorithm does everything for you. All you have to do is watch the results'.	3	Using Rafał Brzoska's image and technology as a source of credibility. The principle of authority is applied by referring to a person (businessman Rafał Brzoska) presented as an expert and emphasising the technological advantage of the algorithm.
The principle of liking	No clear manifestations of sympathy building	0	
The principle of scarcity	'This page is available to only 100 people'. 'This video can only be viewed once'. 'If you leave this page, you won't have another chance to return because your link will expire, as will your opportunity to make money'.	3	Time pressure and limited number of places as a motivator for action. The inaccessibility is strongly emphasised by the limited number of places, unique access to the site, and the need to make a decision immediately. This principle increases the pressure on the recipient to act quickly.

5. Conclusions

Artificial intelligence has the potential to revolutionise key aspects of human life; however, it also introduces profound ethical and societal challenges. Deepfake technology exemplifies this dual nature, offering innovative opportunities on one hand and unprecedented ethical and societal risks on the other. As highlighted in this study, deepfakes have become a sophisticated tool capable of undermining trust, spreading disinformation, and facilitating cybercrime on a global scale.

The authors of this study argue that it is crucial to reserve the term 'deepfake' exclusively for materials explicitly designed to mislead, propagate disinformation, manipulate, cause harm, or discredit individuals. This is particularly relevant to malicious applications, such as financial frauds, disinformation campaigns, blackmail, or criminal impersonation. We are aware that this definition of the term 'deepfake' is narrow, but we intentionally adopt it to emphasise the unethical nature of deepfakes. While broader definitions may encompass a wide range of applications, including creative and harmless uses, our approach facilitates a clear distinction between the ethical and innovative uses of AI – those that foster creativity and positively impact the society – and practices that violate individual rights and erode public trust.

Deepfakes, when combined with social engineering techniques, exploit human psychological vulnerabilities, such as trust and urgency, to manipulate perceptions, emotions, and behaviours. Leveraging established principles of persuasion, such as those identified by Cialdini, they amplify the perceived credibility of messages and raise significant concerns about the ability of individuals and organisations to distinguish truth from deception in critical scenarios, including political campaigns, armed conflicts, and financial frauds.

The growing sophistication of deepfake technology, often rendering content indistinguishable from authentic material, poses severe challenges for detection. Even in cases where imperfections are present, the prevalent use of mobile devices for content consumption obscures subtle indicators of manipulation, further complicating detection efforts.

Countering deepfakes effectively necessitates an interdisciplinary approach that integrates psychological insights, advanced technological tools, and robust legal frameworks. Psychological research plays a pivotal role in elucidating how deepfakes influence human

behaviour while technological advancements improve detection capabilities. Simultaneously, legal measures must address regulatory gaps to ensure accountability for the misuse of such technologies.

This study introduces DEEP FRAME, an innovative tool designed to systematically record and analyse deepfake content. By integrating technical and psychological analysis, DEEP FRAME enables the collection of critical data, including technological characteristics, emotional impact, and manipulation patterns, fostering the development of more effective countermeasures. Additionally, the tool supports interdisciplinary collaboration by creating a comprehensive database to inform educational initiatives, policy-making, and technological advancements. The DEEP FRAME tool facilitates the systematic collection of knowledge about deepfakes. The collected data is categorised based on key parameters, such as the type of manipulation, the technologies used, and the context of publication. These categorisations help to reveal patterns and correlations, providing a structured basis for further analysis and tailored countermeasures. This tool enables the monitoring of global trends in the use of deepfake technology, helping to identify emerging threats.

Deepfakes represent a new generation of threats, combining technological sophistication with manipulative effectiveness. The findings of this study underscore the urgent need for a collaborative response that integrates technological innovation, psychological research, and legal regulation. Tools such as DEEP FRAME play a critical role in advancing these efforts, offering a comprehensive platform to analyse and mitigate the risks posed by this rapidly evolving technology. By addressing these challenges holistically, society can navigate the ethical and security dilemmas associated with AI and safeguard trust in the digital age.

5.1. Limitations

The study was limited to a single case, which may not reflect the full diversity of situations occurring in the deepfake phenomenon. The results may be difficult to generalise to a larger population or other contexts. The results depend on the quality of the data available on the case being studied. Short deepfakes that last, for example, for 1 minute will have significantly fewer persuasive fragments compared to longer materials that last, for example, for 3 minutes. However, the authors of the study plan to conduct

further in-depth research in the future on the phenomenon of using influence methods in deepfakes.

References

[1] M. Caldwell, J. T. Andrews, T. Tanay, L. D. Griffin, "AI-enabled future crime," *Crime Science*, vol. 9, no. 1, pp. 1–13, 2020, doi: 10.1186/s40163-020-00123-8.

[2] M. S. Rana, M. N. Nobi, B. Murali, A. H. Sung, "Deepfake detection: A systematic literature review," *IEEE Access*, vol. 10, pp. 25494–25513. doi: 10.1109/ACCESS.2022.3154404.

[3] R. Chesney, D. Citron, "Deepfakes and the new disinformation war: The coming age of post-truth geopolitics," *Foreign Affairs*, vol. 98, p. 147, 2019. Available at: https://scholarship.law.bu.edu/shorter_works/76.

[4] N. C. Köbis, B. Doležalová, I. Soraperra, "Fooled twice: People cannot detect deepfakes but think they can," *Iscience*, vol. 24, no. 11, p. 103364, doi: 10.1016/j.isci.2021.103364.

[5] M. Caldwell, J. T. Andrews, T. Tanay, L. D. Griffin, "AI-enabled future crime," *Crime Science*, vol. 9, no. 1, pp. 1–13, 2020, doi: 10.1186/s40163-020-00123-8.

[6] D. Milmo. (Feb 5, 2024). Company worker in Hong Kong pays out £20m in deepfake video call scam, The Guardian [Online]. Available: https://www.theguardian.com/world/2024/feb/05/hong-kong-company-deepfake-video-conference-call-scam. [Accessed: Aug. 05, 2024].

[7] P. Kmiecik. (Dec. 21, 2023). Oszustwo 'na inwestycję w gazociąg'. Seniorka straciła pół miliona złotych, RMF24 [Online]. Available: https://www.rmf24.pl/regiony/wroclaw/news-oszustwo-na-inwestycje-w-gazociag-seniorka-stracila-pol-mili,nId,7223202#crp_state=1. [Accessed: Dec. 01, 2024].

[8] M. Pawłowska. Oszustwo podczas inwestycji w akcje 'Baltic Pipe'. Policja Tomaszów Lubelski [Online]. Available: https://tomaszow-lubelski.policja.gov.pl/ltl/informacje/aktualnosci/141018,Oszustwo-podczas-inwestycji-w-akcje-Baltic-Pipe.html. [Accessed: Dec. 01, 2024].

[9] https://policja.pl/pol/aktualnosci/242330,35-latka-stracila-blisko-150-000-zl-inwestujac-w-kryptowaluty.html. [Accessed: Dec. 01 2024].

[10] Y. Weng, S. S. Miryala, C. Khatri, R. Wang, H. Zheng, P. Molino, G. Tur, "Joint contextual modeling for ASR correction and language understanding," in *ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* May 4–8, 2020. New York, NY: IEEE, 2020, pp. 6349–6353. doi: 10.1109/ICASSP40776.2020.

[11] R. Cialdini, *Wywieranie Wpływu na Ludzi. Teoria i Praktyka* (*Influencing people: Theory and practice*). Gdańsk: Gdańskie Wydawnictwo Psychologiczne, 1993.

Appendix 1. DEEP FRAME tool.

Module 1 – Technical analysis

Are there visible artifacts in the material?	☐ Yes ☐ No
Is the audio in sync with the video lip movement?	☐ Yes ☐ No
Is the area around the mouth and teeth more blurred or sharper than the rest of the image?	☐ Yes ☐ No
Does the voice sound natural?	☐ Yes ☐ No
Are there audio artifacts, such as clicks, distortions, or unnatural breaks?	☐ Yes ☐ No
Is there a shift in accent or tone, suggesting a foreign language influence?	☐ Yes ☐ No
How do you rate the level of realism of the material?	☐ Low quality (easy to detect) ☐ High quality (difficult to detect) ☐ Hyper-realistic (virtually indistinguishable)
Are there signs of editing or manipulation (cuts, shifts)?	☐ Yes ☐ No
Are there background interferences, such as static noise, patterns, watermarks, halftones, or visual masks?	☐ Yes ☐ No
Does the content include logical errors or inconsistencies?	☐ Yes ☐ No
Are there grammatical errors or incorrectly pronounced words in the content?	☐ Yes ☐ No

Module 2 – Psychological analysis

The principle of social influence	**Examples**	**Frequency**	**Comments**
The principle of reciprocity			
The principle of commitment and consistency			
The principle of social proof			
The principle of authority			
The principle of liking			
The principle of scarcity			

Module 3 – Scope and context

Does the context of the publication suggest a specific target audience?	☐ Children ☐ Eldery people ☐ People with chronic diseases ☐ Adutls ☐ Men ☐ Woman ☐ Non-binary people ☐ The voters ☐ Believers of conspiracy theory ☐ Others...

What is the purpose of the material?	☐ Parody ☐ Disinformation ☐ Financial fraud ☐ Discredit ☐ Blackmail ☐ Violation of privacy and dignity
Does the material use the image of a leader, expert, or celebrity?	☐ Yes ☐ No ☐ Provide the name and surname of the leader, expert, or celebrity
Did the material appear during the election campaign?	☐ Yes ☐ No
What is the risk level?	☐ Material may influence election results or political decisions ☐ Material may generate hate speech or escalate social conflicts ☐ Material may damage the reputation of a public or private person ☐ Material may lead to fraud ☐ Material may influence public opinion
Module 4 – Recomendations	
Final steps	☐ Necessary confirmation (fact-checking) from trusted sources ☐ Reporting to the platform administration or services ☐ Educating recipients: publishing warnings and guides on manipulation