Cognitive Security: Analysis of automated influence operations based on deepfake advertising

- DOCTORAL THESIS SUMMARY -

Scientific coordinator: Prof. univ. dr. Alina Bârgăoanu

PhD candidate: Mihaela Pană

This thesis explores the evolving interdependence between communication and technology, and how this has given rise to the research question of how online advertising platforms are being exploited as attack vectors in automated influence operations based on deepfake ads. The motivation behind this study is to shed light on this phenomenon and propose countermeasures against deepfake ads, which have become disruptive in an information climate dominated by overlapping crises. Utilising algorithmic instruments for the identification and examination of cases, the document investigates the trends and operational interdependencies between information attacks, manifesting as deceptive advertising through deepfake ads, and cyberattacks, exemplified by the malevolent utilisation of technology. The objective is to identify and depict the digital footprint associated with influence campaigns that lead to cognitive harm.

IMPACT AND ORIGINAL CONTRIBUTIONS

The theoretical framework provides an argument for the technological perspective of communication and outlines the cyber dimension of influence. It introduces into discussion the issue of deepfake advertisements from the perspective of the notion of false advertising as a hybrid security threat and draws attention to the need for developing the concept of cognitive security within the sphere of public policy.

The empirical research is original in its innovative approach to defining the research problem from two perspectives. From a communication perspective, the qualitative-quantitative analysis highlights the exploitation of advertising platforms and correlates the content promoted by deepfake ads with theories from the field of psychology, such as cognitive triggers reflected in the psychological factors behind social engineering (Gragg, 2003), but also the types of fears in Karl Albrecht's hierarchy (2007).

From a technological perspective, the innovative aspect of the research lies in the strategy of identifying misleading advertisements through the concept of an 'algorithmic honeypot'. This involves systematically accessing misleading advertisements and displaying behaviour that is of interest to this type of content, with the aim of influencing the algorithmic effect of the echo chamber and automatically displaying deepfake advertisements. The study also introduces the PART model for interpreting data resulting from OSINT checks that determine patterns in the modus operandi of influence operations carried out through deepfake ads. It also introduces an algorithmic statistical analysis model based on cross-checking results from public service platforms using artificial intelligence to generate and assist in statistical analysis.

In addition, the analysis corpus represented a relevant database with 400 recorded cases over the period 2023-2025, which was necessary for cybersecurity specialists to determine the digital infrastructure of the Doppelgänger operation in Romania and Europe. Immediate reporting to Google, Meta and the National Cyber Security Director allowed early identification of the threat for the development of countermeasures (e.g. the #NoFake governmental mechanisms for user reporting), and the results of the content analysis outlined the parameters of the cognitive threat generated by deepfake ads distributed in the Romanian information space.

RESEARCH CONTEXT

The choice of research topic was determined by the following four aspects:

- The emergence of new concepts in the field of information security in early 2022, such as FIMI (Foreign Information Manipulation and Interference), which defines foreign interference in the information environment, and the DISARM framework, which is designed to identify and counter such activities.
- The intensification of deepfake advertising on social media throughout 2023 targeting the Romanian audience in particular.
- The Digital Services Act (DSA), which includes transparency measures and requires technology companies to prioritise information security, will come into force in February 2024.
- In light of the upcoming European Parliament elections in June 2024 and the presidential elections in November 2024, there is an expectation of potential hostile actions.

Technological developments in the digital era have transformed communication at every level, while malicious influence operations have become increasingly sophisticated through the combination of diverse threats that exploit vulnerabilities across the technological usage chain—from technical flaws in information systems to human weaknesses shaped by emotions and biases.

In today's effective forms of communication, multimedia content follows the model of digital marketing strategies to deliver the message accurately, in the shortest possible time, to the appropriate target user categories. Conceptualizing fake news and the components of information disorder was a first step in researching the alteration of the information environment. The technological evolution of Artificial Intelligence (AI) allows for the generation of a new form of deepfake false content, which can be created very quickly, at low cost, and is much more difficult to identify. In addition, the new strategic approach to information warfare also involves the ability to use online marketing and advertising tools and mechanisms, evolving towards automated influence operations.

Modern democracies are thus facing an avalanche of threats that implicitly affect the cognitive domain, combining information attacks with cyber operations that exploit technology consumption habits in order to shape perceptions and erode the societal cohesion that underpins national security. Through these hybrid techniques, adversaries seek to influence how people think and act, with their mode of operation conceptualized under the notion of *cognitive warfare*, in which the human mind has become the battlefield. Defending against these complex threats, identified in both electoral interference and disinformation campaigns related to medical pandemics and armed conflicts, requires an interdisciplinary approach commensurate with the sophistication of the dangers, combining expertise in communication, psychology, sociology, and cybersecurity. This approach is encapsulated in the emerging concept of *cognitive security*.

The exploitation of advertising platforms for hostile purposes is directly linked to acts of cybercrime, whether involving financially motivated offenders who engage in scams to deceive users less aware of the existence of cyber threats, or state actors who sponsor such influence operations in pursuit of strategic advantages through interference in the politics of adversary nations. Consequently, the analysis considers the polymorphic nature of the motivations of the actors involved, which may range from objectives specific to cybercriminal activity to strategic agendas characteristic of state-sponsored actions, the latter being, in some cases, deliberately camouflaged under the guise of autonomous criminal attacks.

The findings of the research help clarify the dilemma regarding the classification of fake advertising as cyber fraud or an influence operation with cognitive damage. In addition, the disclosure of new techniques used in cognitive hacking helps to understand the effectiveness of advertising reports in the context of the Digital Services Act (DSA), and the data obtained serves as a basis for in-depth research to counteract operations based on information manipulation and foreign interference (FIMI) in an approach that involves expertise in communication and cybersecurity.

RESEARCH QUESTIONS AND OBJECTIVES

The main research questions concern how advertising platforms become attack vectors in automated influence operations and aim to determine the parameters by which we can identify and counter deepfake ads. To achieve these objectives, the empirical research pursues a series of specific objectives in the analysis of hybrid threats that combine elements such as fake news, deepfakes and malvertising in content distributed as advertisements through online advertising platforms.

QUESTION	OBJECTIVE
How do influence campaigns based on online ads with deepfake content operate?	Identifying tactics and techniques used in advertising campaigns with deepfake ads;
How can cyber influence operations be identified using OSINT tools and FIMI analysis frameworks?	Creating an analysis model that leverages observations extracted with OSINT tools and FIMI analysis frameworks;
What are the current patterns and trends regarding the use of online advertising platforms in automated influence operations?	Identifying hybrid threats affecting the informational, cyber, and cognitive space through online advertising platforms;
How do psychological mechanisms based on triggers of fears and emotions work through deepfake ads?	Determining the parameters of cognitive attacks in deepfake advertisements.

THESIS STRUCTURE

In order to answer the research questions and achieve the objectives of this thesis, I have structured the paper into two main parts: conceptual frameworks and empirical research. The chapter dedicated to theoretical concepts deals with the technological perspectives of communication, theories of falsity in the online environment, and how online disinformation currently positions itself between fake news and deepfakes. In this chapter, I also argue how

current studies approach online advertising as a marketing tool and cognitive attack vector in the cyber dimension of influence. Another subchapter deals with the link between cyber threats and online scams, as well as the automation of cyber influence operations and high-tech conflict as part of the evolution of hybrid warfare. Another theoretical representation refers to the framing of the research problem in the field of cognitive warfare, with acceptance of the role of false advertising in attack tactics. There is also a defensive approach to these hostile phenomena included in the synthesis of models, frameworks, and tools used in the analysis of influence operations, as well as in the foreshadowing of the concept of cognitive security as a necessary approach in security policies.

The empirical research chapter encompasses two stages. The first is a case study of deepfake advertisements on YouTube, which demonstrates the exploitation of programmatic advertising. The second is a quantitative–qualitative analysis of deepfake advertisements distributed via Google's and Meta's online advertising platforms. The behavioural profile that emerged from correlating the observed tactics with the DISARM framework was applied to the content analysis in the second stage of the research. Analysis of the recorded data revealed patterns in the exploitation of advertising platforms and trends in misleading content.

KEY CONCEPTS

The conceptual framework starts from the technological perspective of communication and proposes a version of mapping theories, from technological determinism to influence operations. This establishes a link with studies on the cybernetic dimension of influence and forms of advertising based on new technologies, which is the focus of the research. In addition, typologies of hybrid threats and perspectives on the relationship between cyber attacks and information warfare support the theoretical framework. Empirical research is anchored in chapters dedicated to cognitive warfare and tools for analyzing influence operations targeting the cognitive domain.

Deepfake content is defined as audiovisual disinformation (Farid et al., 2019) or synthetic media (Monteith et al., 2024), a product of generative artificial intelligence (McKinsey & Co, 2023), used in various versions of emerging threats in automated influence operations (Goldstein et al., 2023).

Fake advertising is the concept that defines false advertising, associated with intentionally false, misleading, or inauthentic advertising content, disseminated with the aim of deceiving or manipulating the public (Dan et al., 2022). The use of AI-generated deepfake

content in advertising is known as deepfake advertising (Agarwal & Nath, 2023; McCreary, 2024).

Cyber influence operations are defined as activities that affect the logical layer of cyberspace with the intention of influencing the attitudes, behaviours or decisions of the target audience (Brangetto & Veenendaal, 2016; Libick, 2007). These operations are related to algorithmic systems and marketing mechanisms, as well as algorithmic manipulation (Hacker, 2023) and automated influence operations (Goldstein et al., 2023).

Cognitive warfare is defined as the arming of public opinion by an external entity for the purpose of influencing public and government policies and destabilising public institutions (Bernal et al., 2020).

Cognitive security is a new concept currently being researched for a comprehensive definition. One of the defining approaches captures the situation in which technology escapes human control and becomes a weapon used for hostile purposes, in a duel of algorithms that requires intervention to protect users and ensure the technology operates with limited risks (Andrade & Yoo, 2019a; Michael Krigsman, 2019; Terp & Lesser, 2022).

METHODOLOGY

The research methodology comprises multiple forms of content analysis and case studies, organised into successive stages to answer specific research questions.

The first stage of the research project involves a case study examining an advertiser on the Google Ads network who was promoting deepfake ads with the same message on YouTube in May 2023. The aim of this initial analysis is to answer the following research question: *How can the modus operandi of influence operations based on deepfake ads be determined?* To strengthen the practice of integrated anticipation and response to cognitive threats generated by false advertising in the online environment, the multiple analysis of the case study includes message and digital identity analysis, analysis facilitated by Open-Source Intelligence (OSINT) tools for collecting technical evidence, interpreted through the proposed model designed on ***Purposes, Actions, Results, Technique (PART) model*** and the application of the innovative strategic analysis framework ***Disinformation Analysis and Risk Management (DISARM)***, used to anticipate information manipulation and foreign interference (Foreign Information Manipulation and Interference – FIMI) operations, even when these operations do not appear to have a specific, immediately identifiable purpose.

The findings from the initial analysis underscore the need for systematic monitoring of online advertising in the second research phase to identify additional cases of cognitive

hacking via misleading advertisements. Insights from the first phase inform the design of an analytical framework for phase two, aimed at defining the parameters for detecting and countering deepfake advertisements. The variables examined focus on deepfake content linked to websites that, during active campaigns, redirect users to fabricated news feeds embedded with malware, a cyber threat known as *malvertising*.

The analysis employs two data collection approaches: **identification** – recording platform-specific information and details on advertising agents and **indexing** – cataloguing the tactics and techniques used, incorporating both content analysis and assessment of the cognitive impact of such advertisements.

The sample analysed includes 400 cases representing advertisers/promoters who managed ads on the Google Ads and Meta Ads platforms. These cases were identified through systematic monitoring carried out between November 2023 and January 2025, and were grouped into thematic clusters according to the promoted subject. The analysis of all cases revealed the common features of deepfake ads and the pattern of this persistent threat, which is carried out through fake accounts or compromised pages. These pages combine messages based on psychological mechanisms with deepfake content and fake domains that redirect to phishing sites containing fake news that impersonates media brands and celebrities. These sites are also potentially associated with malware.

ANALYSIS TOOLS

A series of OSINT analysis tools were used during the research. The Nvivo platform was employed to index and organise the collected data based on the analysis grid. OpenAI ChatGPT Deep Research was used for automated statistical analysis processing. Google AI Studio was useful in verifying the reasoning of OpenAI ChatGPT Deep Research, on the basis of which the statistical results were obtained.

TECHNICAL VALIDATION

The research was technically validated by checking the case identification process and verifying the results of the algorithmic analysis.

The case identification process was validated by cross-checking results using alternative methods. The results of direct observation or crowdsourcing reporting were verified by querying the advertising library of the advertising platform where they were identified, and validated by official reporting. The indexed information was reported to both the platform where the ads were detected for blocking, and to the National Director for Cyber

Security in Romania, for threat assessment and the elimination of cyber threats from the information space.

Cross-algorithmic validation involved verifying the results obtained using OpenAI ChatGPT Deep Research by conducting a similar analysis using a system such as Google AI Studio.

SELECTION CRITERIA AND VARIABLES ANALYSED

Identifying the behavioral profile and mode of operation in the first stage of research determined the conditions for selecting cases—the accounts of advertisers included in the corpus: (1) false identity: accounts use false names, generic names, or pseudonyms; (2) techniques for redirecting promoted links to compromised/hijacked fake sites; (3) symbolic communication elements regarding national identity/state authority; (4) fake news content - fabricated or copied text from original sources; (5) deepfake content - photo, audio, or video; (6) warning about cyber threats: phishing and/or malware; (7) techniques for hiding and/or deleting digital footprints.

The analysis grid is structured around 14 case-specific aspects identified in the first stage, focusing on variables with characteristics specific to the accounts of agents on advertising platforms, as well as descriptive elements of the promoted content. This includes examining who is promoting it, the origin of the accounts promoting the content, what is being promoted, the details of the promoted content, and how it is being promoted, as well as aspects of the mode of operation and associations with cognitive triggers and the exploitation of fears.

FINDINGS

Analysis of the identified cases revealed significant correlations and behavioural patterns in malicious campaigns involving deepfake advertisements. The study of these ads, which integrates data about the advertiser's account, details related to the moment of identification, and content analysis, reveals that they differ in size, duration, and language, but all essentially pursue the same goal: manipulating public opinion in favour of hidden agendas. Highlighting the links between the characteristics of insidious ads can contribute to the development of a proactive framework to combat these hybrid threats to the information society.

The results of the analysis show that key variables, such as campaign duration, ad volume, audience/impact, or languages communicated, are interdependent with the content

promoted, which in most cases is structured around a combination of psychological mechanisms. All these aspects reflect the strategy of the mode of operation and can determine the parameters of an early warning system in cases of cyber influence operations where advertising platforms are exploited.

The overwhelming majority of advertising campaigns identified were persistent over time, a sign that the administrators of these ads have an infrastructure that allows them to maintain a continuous presence in order to consolidate their influence. Long-term campaigns, particularly those involving a large number of ads and multilingual content, are likely to be high-level orchestrated operations associated with cybercrime groups linked to state actors. In contrast, short-term campaigns suggest experiments to test audiences or rapid responses to events that aim to shape the information agenda and gauge public response. Based on the evidence of their operational patterns, campaigns run by recently launched advertiser accounts with a small number of ads may indicate new narratives that are about to be propagated on a large scale.

The Romanian language dominates in deepfake advertisements, indicating a strong focus on influencing the Romanian audience. At the same time, the significant presence of the Russian language confirms the existence of active external influence (possibly Russian), consistent with the geopolitical context in which Romania and its neighbouring countries have been subjected to Russian propaganda campaigns. This is also supported by multilingual campaigns (Romanian-Russian-Ukrainian-English) targeting audiences segmented across several European countries, which signals concerted transnational efforts, representing the most complex cases, but also the most targeted by countermeasures, as suggested by the status of the advertisements at the end of the monitoring period.

Quantitative correlations can be used to determine parameters that contribute to early warning mechanisms for this threat. A large volume of advertisements leads to a large audience; longer durations of active advertisements allow for audience accumulation; and multilingualism expands the area of action. These correlations are important because they provide clues about the effectiveness of tactics. If an actor wants to reach a very large audience, the data shows that they should invest in lots of advertisements and keep them active for as long as possible (or combine the two options to attract as many visits as possible to the promoted website). Thus, the relationships obtained from correlation analysis can constitute early detection heuristics. For instance, a new promoter/advertiser account that broadcasts dozens of ads in several languages simultaneously in the first few weeks may indicate coordinated inauthentic behaviour warranting immediate investigation, before the

promoted message reaches millions of users. Such behaviour on the advertising network can also be associated with recurring influence tactics involving specific content and targeting the intended audience, such as imitating authentic local media to target groups segmented at the local/county level, regional propaganda integrated with content distributed through audience segmentation at the country level, exploiting key moments in the media agenda and diversifying channels for greater coverage. All these strategies have been recorded in global briefings on persuasion tactics; for instance, Facebook has noted the tendency of 'platform diversification', 'retail IO' (smaller, targeted initiatives), as well as the employment of intermediaries to manage such initiatives – developments that our information mirrors in the context of Romania.

The present study explores the correlations between the category, duration, volume, audience and language of misleading advertisements. This analysis highlights two predominant categories of influence operations active in the Romanian online space: sustained local campaigns, which appear to originate from within the country, and large-scale cross-border campaigns, likely associated with foreign influence given the discrepancy between the origin of the promoter accounts and the language used in the advertisements they sponsor. It is evident that both categories utilise paid advertisements to manipulate perceptions; however, they differ in terms of volume, language, and target audience. The identification of recurring patterns (e.g. persistent versus spontaneous, multilingual versus monolingual) has the potential to inform future endeavours aimed at automatically detecting influence operations. These characteristics, which comprise the digital fingerprint of inauthentic behaviour, are of particular relevance in this context.

The data gathered indicates the presence of substantial correlations and delineating trends in the modus operandi of influence campaigns, which are facilitated by the deployment of misleading advertisements. The employment of manipulated imagery in the context of advertisements represents a potent instrument of manipulation, particularly in instances where the subjects are prominent figures, such as politicians, media personalities, or esteemed experts. These individuals possess the capacity to exert a significant influence on public sentiment. This finding underscores the gravity of the phenomenon, as it demonstrates not only the hijacking of institutional brands but also the exploitation of individuals' identities. This constitutes an abuse that can be regarded as defamatory and may bear legal ramifications from the perspective of identity usage without consent.

This analysis has revealed a complex ecosystem of deepfake advertisements, which strategically combine elements of disinformation, social engineering and advanced cyber

techniques. From a quantitative perspective, it has been observed that the phenomenon involves a large number of fake entities (dozens of countries of origin, hundreds of fake identities, numerous topics covered). The preponderance of external sources, notably from Ukraine and Asian countries, in conjunction with the pervasiveness of fake identities, underscores the internationally coordinated nature of the operations, thereby suggesting the existence of well-established networks. From a qualitative perspective, deepfake advertisements are notable for their versatility and adaptability. They address current topics (financial, social, medical and political), utilise the image of the most trusted public figures and brands representative of the target audience, and press exactly the right emotional 'button' for the target audience (fear of poverty/illness, greed for gain, patriotism, vices, etc.). These entities employ a multifaceted approach, encompassing tactics such as the dissemination of false information through media channels and the manipulation of digital content, known as "deepfakes," in addition to sophisticated phishing techniques. This diversified array of strategies culminates in a multifaceted attack on the user, thereby enhancing the efficacy of the advertisements in deceiving the intended target audience.

Another significant aspect in determining the deceptive pattern is the synchronisation of the categories studied in the analysis grid. These categories include the alleged origin of the advertisers, the content promoted, and the tactics used to construct the message to activate the audience. The manner in which these elements support each other is indicative of an operation orchestrated through a complex infrastructure. The hidden origin enables the employment of aggressive tactics without the risk of repercussions, the misleading content is based on false information, and the tactics allow for the origin and false information to be covered up. When considered as a whole, these elements delineate a disinformation campaign of a sophisticated nature, designed to support an integrated fraud, which must not be confused with disparate acts of cyber fraud.

A fundamental characteristic of such deceptive advertisements is their capacity to manipulate human psychology by eliciting fears, desires, and cognitive biases, thereby influencing user behaviour. Qualitative coding reveals a series of cognitive triggers and systematically reveals the utilisation of fear in campaigns. The psychological mechanisms that were found to be most effective included authority, reciprocity, fear, and trust building. These factors were often identified in a complex approach that significantly increases persuasiveness by involving multiple triggers simultaneously. For instance, a fraudulent scheme that relies on the concept of investment employs authoritative figures (in the form of state endorsements) to establish trust, subsequently leveraging the abuse of power to prompt

immediate action. This fraudulent scheme instils fear (concerns regarding poverty, inflation, unemployment, and financial crises) and fosters the pursuit of expeditious and effortless gains through a reciprocal reward system (e.g., bonuses or time-limited access).

Cases of advertisers whose advertisements depict the government as a menacing entity suggest a concern for understanding audience sensitivities: some segments of the public fear the state more than they trust it, which is why they have been targeted with messages that promote feelings of revolt against the abuse of authority. Furthermore, cases that exploit the fears of war or territorial occupation have been observed to build their communication on the expansion of the conflict beyond Ukraine's borders, with the objective of scaring people into protecting their property or following the recommendations in nationalist messages.

Consequently, the correlations between fears and tactics highlighted the patterns linking psychological mechanisms to false messages. For instance, advertisements that exploited the fear of poverty employed authority triggers and appropriated the identity of dignitaries and financial institutions. The utilisation of nationalist rhetoric was also observed, with suggestions being made that investing was not solely an action for personal gain, but also an act of patriotism. In the case of fear of disease, the correlation was with the usurpation of medical authority and powerful emotional triggers to induce users to purchase the promoted treatments.

A further psychological complexity is exhibited in the manner in which authority triggers are utilised to either amplify or alleviate fear. When an authority figure issues a warning of potential danger, this is an example of the former, whereas when the authority figure conveys confidence and offers reassurance, this is an example of the latter. The high coefficient of authority (321 advertisements) and fear (particularly the fear of poverty, as evidenced by 289 advertisements) indicate that financial scams were perpetrated through the exploitation of the fear-authority nexus: the audience was initially intimidated, subsequently reassured by the prospect of authority endorsement, and ultimately deceived.

A further correlation has been identified between the utilisation of relational triggers and particular fears, a phenomenon that has been predominantly observed in medical advertisements. The 'deceptive relationships' triggers (116 advertisements) frequently manifested in clusters on health subjects and the recovery of financial losses incurred in cyber scams, based on trust-building scenarios.

In conclusion, the psychological profile of these deepfake advertisements has been shown to provoke fear and then offer a (false) saving solution, combining the appeal of fear

(especially economic and health fears) with cognitive biases (trust in authority, reciprocity, sense of urgency) to reduce scepticism and critical thinking. This synergy of fear and authority is a hallmark of effective influence operations, and is evident in the dataset.

OPERATIONAL MODELS OF COGNITIVE ATTACKS BASED ON DEEPFAKE ADVERTISING

This thesis explores the operational models of cognitive attacks based on deepfake advertising. The correlation of the analytical findings from the content analysis stage with the operational data from the identification stage has enabled the differentiation of several operational models of advertising campaigns that evolved during the monitoring period.

The exploitation of advertising platforms was a key finding of the study, with the majority of the analysed deepfake advertisements (89% of cases) being distributed through the Meta network (comprising Facebook and Instagram). This finding underscores the significance of advertising distributed on social networks as the primary vector for influence operations. The Meta network is favoured in hostile campaigns for a number of reasons, including its ability to target specific demographic categories through multimedia diversity. Furthermore, the presence on both Meta and Google indicates that some of the most prolific campaigns were conducted on both platforms. This multi-platform approach renders detection more arduous and is indicative of a coordinated operation.

Campaign persistence: A substantial proportion of deepfake advertising operations are characterised by longevity, with approximately 77% of ad campaigns remaining active for a duration exceeding 30 days. The longevity of the malicious advertising activity suggests that these accounts have been successful in evading detection in order to maintain a presence in the audience's attention over an extended period. In order to evade detection, tactics were employed to modify accounts frequently, thus avoiding closure or the concealment of deepfake advertisements within multiple versions of an ostensibly authentic advertisement. The ongoing development of methods employed for the concealment of malevolent content has resulted in a considerable proportion (exceeding 60%) of identified accounts remaining undetected from the respective platforms, despite having been formally reported, by the conclusion of the monitoring period. In the majority of cases, deepfake advertisements were either removed or disabled; however, the accounts in question proved resilient. It was found that only 36% of accounts were 'deleted' by 2025, either through countermeasures by the platform or through removal by hostile operators as a precautionary measure. The operational conclusion suggests that enforcement by the advertising platform was

inconsistent or delayed, with many malicious accounts surviving long after reporting. However, this longevity enabled the malefactors to disseminate substantial quantities of advertisements and attain vast audiences (with some exceeding one million impressions) prior to any intervention.

Scale of the advertising campaigns: The distribution of the number of advertisements exhibited significant variation from account to account, with pronounced differences observed between accounts with a high number of advertisements and those with a low number. Nevertheless, certain observations have been made with regard to the manner in which they function. It has been observed that accounts exhibiting a minimal presence of advertisements tend to disseminate uniform content material over brief periods. Accounts containing a volume of advertisements exceeding 1,000 are indicative of industrial-scale influence operations, frequently associated with foreign entities undertaking multilingual campaigns aimed at a broad regional or global audience in the Russian Federation (RU), Ukraine (UA) and the European Union (EU). It is highly probable that these extensive campaigns have utilised internet platforms to disseminate a variety of deepfake advertisements, experimenting with diverse messages and optimising their reach. This pattern suggests the existence of coordinated inauthentic behaviour, in which a small number of groups with sufficient resources can produce content on a large scale.

Audience segmentation: a significant proportion of the accounts examined were found to have advertising campaigns targeting audiences of tens of millions of users at the time of identification. In 83 cases, the distribution of advertisements was estimated to have had an impact on more than one million people, and in a further 95 cases, between 500,000 and one million. It is noteworthy that a mere 25% of the campaigns retained a viewership of fewer than 50,000 viewers. Multilingual campaigns were especially effective in targeting users from multiple countries or ethnic communities. For instance, some advertisements were broadcast simultaneously in Romanian and Russian, targeting both Romanian citizens and the Romanian-speaking Moldovan or Ukrainian population, as well as the Russian-speaking minority. It is notable that 19 campaigns incorporated content in Romanian, Russian, Ukrainian and English, suggesting a strategic approach aimed at maximising reach within the region. This multilingual approach suggests that the actors were operating with full awareness of the facts on a global scale, repackaging the same scam for different localities. Furthermore, the multilingual nature of the campaign could be implemented by either an international team or a single actor with a range of language skills.

Another observation is that gender segmentation is relevant to the promoted topic. Therefore, financial investment advertisements were observed to be targeted at males aged between 35 and 64, whilst those for miracle treatments were found to be targeted at females within this age range.

Legitimacy tactics: According to the available identification data, a significant proportion of advertiser accounts were identified as 'unverified' identities within advertising networks. These accounts were often newly created Facebook accounts or pages. However, it is noteworthy that 14 cases involved 'verified' identities. This suggests that the administrators of these accounts either compromised legitimate Facebook pages/accounts or manipulated the identity verification process by the advertising platform. It is evident that a further pattern that has been observed is that a number of fraudulent accounts have been found to adopt the guise of companies or non-governmental organisations (NGOs). Furthermore, it has been determined that these accounts have been able to elude detection by creating sock-puppet accounts over an extended period of time (a tactic employed by operators to ensure that accounts remain inactive and avoid arousing suspicion regarding newly created accounts in a campaign scheduled at a distant point in time). This operational patience is indicative of coordinated influence efforts.

Detection and removal: the process of identifying and eradicating fake accounts disseminating deepfake advertisements demonstrates the efficacy of direct observation by a human operator. This approach is more effective than automatic detection by the platform or by external or community supervisory bodies. The platform transparency tools stipulated by European regulations were instrumental in facilitating the implementation of the analysis grid. However, it is evident that these measures are inadequate in preventing the propagation of misleading advertisements. The operational model in this case demonstrates a tendency towards reactive removal rather than proactive prevention.

In light of the findings, a correlation between content analysis and operational data has been established, providing evidence of a meticulously organised and persistent influence operation. This operation has been shown to involve the dissemination of sophisticated deepfake content on pertinent platforms, with the objective of targeting substantial audiences. The operators functioned on a substantial scale, frequently on an international level, and adapted their strategies (in terms of language, theme and technique) to diverse contexts. The employment of deepfakes and psychological persuasion constituted an element of a sophisticated framework of tactics, techniques and procedures, encompassing social engineering (impersonation, authority), technical exploits (phishing, fake domains) and

operational security (covering tracks). This demonstrates how misleading deepfake advertisements combine elements of disinformation campaigns and cybercrime, blurring the line between disinformation and fraud.

The findings of this research highlight the necessitation of countermeasures at that level of sophistication, in a manner that combines the enforcement of platform policies, the education of users about deepfake information threats, and international cooperation to detect cross-border actors involved.

In essence, the deepfake advertisements examined serve as a compelling illustration of an emerging threat to the information space, with the potential to inflict both financial and cognitive harm upon society. It is evident that, in terms of their structural design, these entities effectively amalgamate the most deleterious elements of both the domains of fake news and online scams. This amalgamation is further compounded by the utilisation of contemporary digital propagation instruments. The observed trend is one of increasing sophistication, and it is the duty of information security to anticipate the threat that these advertisements represent, especially when technology makes it difficult to distinguish legitimate communications from false ones (as deepfake technology advances and becomes accessible to the masses). In this context, it is imperative to possess a comprehensive understanding of the tactics emphasised in this analysis to facilitate the development of effective countermeasures and the education of the public to avoid falling into cyber traps.

The findings of this research demonstrate that deepfake advertisements should not be regarded as mere 'false advertisements'; rather, they constitute a pernicious amalgamation of misinformation, cyber fraud and image abuse, necessitating a suitable response. A comprehensive understanding of the operators responsible for producing the promoted content, in addition to the mechanisms through which it reaches individuals likely to access it, is essential for the development of more effective countermeasures. This analysis has exposed patterns of action, thereby enabling the proactive utilisation of this knowledge in efforts to protect the information space.

FUTURE RESEARCH AGENDA

The present research can serve as a basis for the development of **early warning systems based on operational parameters** resulting from correlations between the variables studied. The database obtained has the potential to contribute to **the expansion of the analysis of shared infrastructure** (which includes social media IDs, IPs, and promoted domain names) by creating a **threat stream in the Open CTI** cyber threat intelligence

platform (contributing to the integration of data in cybersecurity community investigations). The findings of this study can also assist in the formulation of a **public awareness policy** on the threat posed by fake advertising, thereby supporting entities responsible for enhancing security culture. The classification of the threat posed by deepfake advertisements as a tool of cognitive warfare necessitates **the popularisation of the concept of cognitive security** at the level of the National Defence System, through the implementation of a programme of internal protection measures.

RECOMMENDATIONS FOR PUBLIC COMMUNICATION

The conclusions of this research are particularly relevant for specialists in the field of public communication, providing arguments for the necessity of an integrated approach to cybersecurity elements in the development of media culture and digital education. An adequate response to cognitive threats based on false advertising can be built on a three-dimensional perspective—human, technical, and institutional—reflected in the concept of cognitive security.

From the human perspective, both the legislative and regulatory dimension and that of public communication can be further developed. Legislative adjustments could involve the introduction of **clear rules regarding platform** accountability for the dissemination of deepfake ads, **mandatory labelling** of synthetic content, and **rapid reporting mechanisms** for suspicious ads by users, with legally established response times. Optimizing public communication would involve conducting **awareness campaigns** about the dangers posed by deepfake ads, so that they can be more easily recognized by users, as well as encouraging public rebuttals with **prompt reactions** from brands and public figures whose image is being abusively used.

Technical cooperation on combating deepfake advertising would require **international standardization for the automatic detection of deepfake content** (similar to cybersecurity standards for spam or malware), **joint databases** between states and tech companies to flag fake accounts and domains, and **public–private partnerships to develop open-source tools** for identifying deepfakes and networks of fake accounts on digital platforms. Being a borderless criminal phenomenon, international cooperation in combating cyber fraud should also focus on the indicators provided by online advertising in order to halt the activity of the human networks behind virtual ones.