SYNTHETIC REALITY AS A STRATEGIC THREAT: FROM DISINFORMATION TO PREDICTIVE MANIPULATION

Vladimir Babanov, Ph.D.

Chief-assistant professor, South-West University Neofit Rilski

Scopus Author ID: 60021493300; ORCiD: 0000-0001-8596-6493

E-mail: v.babanov@law.swu.bg

Abstract: *The article examines the potential of artificial intelligence to transform disinformation into a sophisticated type of synthetic reality operating on hyper-personalized manipulation and cognitive exploitation. Synthetic reality is fundamentally different from traditional forms of propaganda as it is deeply immersive, far more persistent, and has the capacity to influence the decisions and behaviours of individuals and societies. The discussion explores how synthetic media undermine strategic trust, while creating a lasting threat bypassing conventional detection and defense mechanisms. The paper identifies significant gaps in preparedness in terms of law, technology and doctrine and concludes by identifying policy and resilience based solutions for dealing with these evolving challenges.*

Keywords: *Disinformation; Artificial Intelligence; Synthetic Reality; Security Policy; Cybersecurity;*

Introduction

In the digital era, adversaries have engaged in massive coordinated efforts to sway public opinion in certain directions through information campaigns. They are often described as disinformation, whose dangers become more evident (Helmus, 2022). The emergence of new technologies has created a different type of threat that shapes the relations between societies and factual knowledge. Synthetic reality is on the rise in cyberspace through algorithms and AI, generating and spreading fake content almost indistinguishable from authentic. The plethora of synthetic instances include AI-generated images, video, audio, AI-agents, automated systems for malicious actions in cyberspace and many more. It also spreads rapidly in the domain of simulated experiences such as augmented reality (AR) and virtual reality (VR). This illustrates the shift of

disinformation campaigns abandoning text-based form and replacing it predominantly with synthetic audio-visual media. For example, generative AI can create convincing deepfakes that even experts could not differentiate from real (Busch, E., & Ware, J., 2023). Experts also caution that technology is on the verge of reaching a synthetic reality threshold, at which artificially constructed media becomes indistinguishable from human-driven and ultimately replacing it.

As a result, the overtake of cyberspace by synthetic reality should be considered a separate security domain. Adversaries abandon the doctrine of simply altering human beliefs and focus on permanent control of perceptions and ultimately drive social collective decision-making through constant flow of synthetic content. This paper explores this phenomenon by examining the limits of the conceptualization of disinformation and assessing the emerging security risks associated with synthetic content, boosted by generative AI. It also evaluates how AI changes the paradigm of disinformation from influencing to direction-driven manipulation with challenges to strategy and governance.

Methodology

The research is based on a multidisciplinary approach, broadly based on aspects of security studies, media theory, and emerging technology analysis. Using primary sources, including policy documentation, academic literature, and official threat assessments from international organizations, this paper demonstrates the transformation of disinformation into AI-enabled synthetic reality. Through the application of conceptual analysis, the research redefines the implications of security regarding synthetic media and challenges the traditionally accepted framework of propaganda. Case examples demonstrate operational changes in influence tactics including, but not limited to, generation of deepfake media by AI and cognitive manipulation of individuals through hyper-personalized content.

Data from secondary sources, including cybersecurity incidents, government reports, and ethics discourse related to AI, provide additional evidence supporting the synthesis of the research. The methodology employed in it supports a critical-analytical perspective in order to reveal doctrinal, regulatory, and cognition asymmetries exploited by synthetic content. Ultimately, the research seeks to identify gaps in current defense paradigms and to provide a basis for arguing that synthetic reality should be viewed as a new, independent domain of security.

Discussion

Some other types of dangerous content originate from disinformation with the assist of generative AI. International bodies define three such content types. The disinformation is false information created and shared with malicious intent. In contrast, misinformation is false content shared unwittingly without the intent to deceive, and malinformation is truthful content used maliciously, for example, real data leaked with the intent to embarrass a target. These definitions highlight two features: falsehood and intent. While appealing to the concepts of falsehood and intent to be common in defining and judging disinformation, it can sometimes constrict the focus. Scholars note that concepts of truth and intent are themselves often legally and politically challenging (Appleman et al., 2022). Truth is often represented as argumentative and abstract, and in this context, proving false any controversial statement in polarized societies and a toxic information environment could be impossible. Likewise, suspecting a malicious intent behind a communication is difficult as doubtful claims are shared as truthful in a chaotic and random manner.

The conceptual limits seem to point that the term disinformation may be too narrow to describe the scope of the contemporary phenomena in all communication channels. Information attacks often combine truth and falsehood as adversaries seek to inject true footage or statements in an attempt to gain credibility. Weaponized context and genuine but misleading materials are now mundane influence tools. The very notion of reality is under technological siege. When AI produces seemingly authentic events, it deepens even further the ambiguity between real and unreal content which accumulates in cyberspace. Synthetic reality evolves on this fake content and creates new truths by creating sequences of deception that remain in cyberspace and reach billions of users.

Synthetic reality as a security domain

Synthetic reality is a broad category that describes a flux of technologies, cognitive elements and information flows. The synthetic reality has spread almost to every technological aspect to reach even advanced concepts like extended reality (XR) with the subfields of virtual reality (VR) and augmented reality (AR). XR devices can layer digital imagery over a user's perceptions about the physical world through specialized equipment. An industry forecast predicted that the XR market would exceed $200 billion showcasing its significance (Happa et al.,

2019), meaning that immersive digital worlds are now commonplace, from enterprise VR meetings to AR used to navigate cities. To function properly, XR devices must connect to the Internet which appear to be a new attack surface with tremendous destructive potential. Researchers have notably demonstrated proof-of-concept attacks on AR systems. These instances show new attack methods where a hacked AR interface could manipulate real world objects, violate privacy norms, and misleadingly instruct users in ways that risk their physical safety. Attacks on VR pose a significant danger to security as military personnel is increasingly trained with the technology for real scenario operations and any fake information into the system means increased risk.

Beyond AR and VR devices, the essence of synthetic reality today lies in the synthetic media as audio-visual content, created by AI. This encompasses everything from deepfake images and videos used by chatbots to virtual influencers based on generative AI. The security ramifications on individual and societal levels are already present.

The prevalence of deepfake scams is growing in the corporate world too. Criminals are applying AI to mimic executives using video or audio, in remote meetings. Trusting their perceptions, the victim authorized a $25 million funds transfer to hackers, using synthetic voice and imagery (Romero-Moreno, F. (2025). This technology enhances social engineering and could evade conventional cybersecurity controls as no devices nor networks need to be compromised. Attackers need to hack just the human perception. Such incidents increase the likelihood that a virtual impersonation could compromise an entire security posture of an organization or a nation state. Even if exposed, the consequences of the breach of trust, cannot be undone easily as facts are lost under the flood of disinformation.

Synthetic reality attacks provide adversaries with a platform to integrate three types of operations such as traditional espionage, propaganda and social engineering in a single operation. Synthetic media, due to its ease of dissemination via social media platforms, can easily appear as a credible fact for a particular audience. Synthetic reality expands upon the existing arena of disinformation, because of the ability to merge generated AI content with authentic interaction. Because of these specifics, security experts would need to address both technical and cognitive vulnerabilities to defend against this kind of threat.

From influence to predictive manipulation

As the synthetic technologies mature, the malicious influence is transitioning from broad audience messaging to predictive manipulation of individuals through tailored artificial media. AI systems are learning to map how a person is likely to behave in response to stimuli and provide hackers suitable tools for maximum gain. In another way, attackers are moving along from flat propaganda to exploit their victims' psychological profile on a hyper-personalized basis. This critical difference is itself descended from developments in algorithmic profiling and behavioural social science. Machine learning models take as input a variety of data including a user's browsing habits, social media activity, purchase history, even biometric data, and highlight relationships often invisible to humans.

In finding patterns, a mental model is being built to understand the cognitive motives behind a human behaviour. Through this model, malicious actors or automated systems learn which content to serve to targeted victims. The information selection is updated in real time, progressively learning about the target the more it is exposed to it. This technique is genuinely different from influencing. Rather than writing an identical story for mass dissemination, predictive manipulators establish a feedback loop with each user.

Given enough data, an AI can predict someone's bias and consumer choices in future before they do (Madanchian, 2024). This means that propaganda and persuasion can be enacted on a national scale as well as in private life through cyberspace. AI could discover that viewing a particular news clip could indicate a voter's discontent with a certain political stance and exploit it further.

The phenomena, known as hypernudging, can stealthily overtake societies' decision making. Multiple international bodies have decided that algorithms that hypernudge, could undermine human rights such as the informed self-determination and freedom of thought (Morozovaite, 2023). The speed at which generative AI enables the breadth of mass influence allowing malicious actors to disseminate massive amounts of customized content at negligent cost. Further, malicious campaigns have become persistent as they can target both vast populations, and individuals through hyper-personalized messages and spam. The combination of hyperpersonalization and high-volume mass content pushed the information environment into a dangerous grey area. In such a context, defending societies collectively and individually becomes

extremely hard with the mass tuning-in of users into AI-boosted daily streams of disinformation and conspiracy theories.

Security and strategic implications

These trends have huge ramifications for national security and strategy. Broadly they're all part of the emerging notion of cognitive warfare. Military analysts define cognitive warfare as the use of technology to enable tactics that produce effects on the psyche of an adversary's population and leadership, including manipulation of information (Deppe & Schaal, 2024). Synthetic reality tooling is becoming an underlying enabler of successful cognitive attacks. That warning is not new, but the actual early evidence of its impact appeared a few days after the Russian invasion of Ukraine in 2022. Back then, a deepfake video of President Zelensky promising Ukrainian citizens that they would be spared if they laid down their weapons circulated widely on Telegram channels (Byman et al., 2023).

Though Ukrainian officials quickly reacted by posting official statements, the event represented a milestone. For the first time in a conflict anyone was able to create an AI deepfake which depicted a believable message from a head of state with an effect, although short-lived. Security analysts face the problem of adapting strategically under the assumption that militaries and intelligence services could accept any message to be a deepfake by an adversary. A Brookings Institution study highlighting this trend shows that rival state and nonstate actors now have free access to deepfake capabilities, generating compelling audio and video messages.

In the flux of synthetic reality, everything from false press statements to spurious orders in the decision-making command chain could be generated by AI. The risk in failing to fortify defenses therefore gives the adversary the ability to sow massive confusion, panic, or misdirection without the need to engage.

At a broader level, synthetic reality is changing the way global influence campaigns are waged. For decades, government and party operatives in numerous countries have employed various forms of computational propaganda. More recently, studies have found that government aligned entities in at least 70 different countries have engaged in social media manipulation campaigns and at least 45 countries have utilized AI-generated content in election campaigns (Perrigo, 2022). While Russia and China have been cited as leading the charge in terms of utilizing online propaganda, numerous other countries and political organizations are also participating in this arms race. The suggested creation of Deepfakes Equities Process in the USA, would provide

a mechanism for ensuring that any decision to utilize deepfake technology in a campaign against another entity is thoroughly reviewed by multiple government agencies. Similar to the existing Vulnerabilities Equities Process, which balances the offensive benefits of exploiting vulnerabilities with the ethical and diplomatic costs of doing so, such a review process would serve to ensure that any use of deepfake technology is done in a manner that reflects the values and policies of the government. The fact that such recommendations are currently being considered indicates how central synthetic content has become to national security planning.

Results

Governance limits and defensive mismatch

Although synthetic reality threats are increasing, defending against them is extremely challenging. The ability to create, distribute and attack is inexpensive and quick, and most detection and attribution technologies lag behind. Most proposed legal and technological defenses are severely limited by their practicality. For example, the European Union's proposed Artificial Intelligence Act provides that generated media shall be labeled artificially generated (Garcia Luengo, 2025). In theory, this establishes a transparency requirement as voters should be able to identify whether a political message was produced by AI. Similarly, users should be able to recognize a deepfake watermark on audio calls, for example. However, the significant challenge to enforcement of the act is that unfriendly actors can create and host the content outside of EU jurisdiction or remove any identifiable markers. Additionally, watermarking algorithms are inherently imperfect and can be removed or spoofed. Labeling has no impact upon preventing the widespread dissemination of a false image or video once it has been created and distributed.

The existence of defensive imbalance is evident as it is often easier for an attacker to create a believable synthetic content than it is for a security expert to determine whether it is real or artificial. When a fake is introduced, fact checks and takedown attempts are typically late responses. While social media platforms may attempt to establish forensic tools or hash databases to alert users to known fakes, these comprise reactionary measures. On the other hand, research and development to automate detection continues to lag behind the advances in generative models.

Policy-wise, the effectiveness of efforts to reduce synthetic reality threats varies widely. A number of jurisdictions have enacted legislation prohibiting malicious deepfakes with debatable effectiveness. Internationally, there is no consensus with respect to labels for authenticity on shared

standards. Industry associations have established voluntary options for labeling AI-content but this measure is simple to evade. As a result of the lack of effective legal remedies, experts emphasize the importance of human education and cognitive resilience. Education through media literacy programs can assist individuals to be aware of sensationalized content. Similarly, organizations are encouraged to educate their employees about detecting or validating suspicious communications. Governments and companies are also advised to develop incident response protocols for synthetic reality attacks.

The world is still developing a defensive posture on the synthetic reality risks and several promising concepts do exist. However, they still remain far from being universally adopted. Mitigating synthetic reality risks would resemble an arms race, because even increasing the costs and risks associated with attacking, complete prevention is impossible. The ultimate goal may be preservation of strategic trust and societal resilience. To accomplish this, societies and institutions need to build frameworks and norms that allow for the swift identification of truth versus fiction during major events or incidents, and for the maintenance of trust in official channels.

Conclusion

Synthetic reality reveals a new area of information security. Though it utilizes an old technique of disinformation, it goes further as today's AI creates completely believable audio-visual events and individualized delusions. It causes a blurring of the lines between real and created. Therefore, defenders need to think about new ways to protect themselves against this type of threat. Content policing is not going to be enough for that purpose. A defense strategy needs to include legal structures to support it, for example the EU has already started to put regulations in place for AI labeling, and there are proposals for a process called the Deepfakes Equities Process for the use of deep fakes in the military by the United States.

In addition to legal and technical defenses, other forms of defense are also necessary. For example, creating technologies improving detection of deep fakes, refining the authentication of official media outlets, and using digital watermarking to prevent attacks. Social and educational programs, such as media literacy programs, could teach users how to recognize manipulative deep fakes. Rapid response fact checking networks, and public communications programs to prepare preventively for the distortion of messages could also help improve the defenses. International

coordination would be critical because of the ability of synthetic media to flow across borders instantly.

There are a number of significant risks associated with allowing synthetic reality campaigns to continue without some form of regulation. The potential risk includes the erosion of trust in the election process, the degradation of the effectiveness of public health, and the destabilization of national institutions due to the exploitation of uncertainty by those who wish to create doubt. However, recognizing synthetic reality as a distinct security domain, and understanding that AI generated manipulation should be viewed as a strategic threat, provides the opportunity for societies to develop resilience and protect the confidence in democracy. Though it is unlikely to completely eliminate deception, the ultimate goal is to ensure that truth and trust are resilient in the dangers of synthetic reality.

References

Helmus, T. C. (2022). *Artificial intelligence, deepfakes, and disinformation: A primer*. RAND Corporation. DOI: 10.7249/PEA1043-1

Busch, E., & Ware, J. (2023). *The weaponisation of deepfakes: Digital deception by the far-right*. International Centre for Counter-Terrorism – The Hague. https://static1.squarespace.com/static/5b7ea2794cde7a79e7c00582/t/65c284a6b0bb5b2623054030/1707246758762/The+Weaponisation+of+Deepfakes.pdf

Morozovaite, V. (2023). Hypernudging in the changing European regulatory landscape for digital markets. *Policy & Internet, 15*(1), 78–99. https://doi.org/10.1002/poi3.329

Appelman, N., Dreyer, S., Bidare, P. M., & Potthast, K. C. (2022). Truth, intention and harm: Conceptual challenges for disinformation-targeted governance. *Internet Policy Review*. https://policyreview.info/articles/news/truth-intention-and-harm-conceptual-challenges-disinformation-targeted-governance/1668

Happa, J., Glencross, M., & Steed, A. (2019). Cyber security threats and challenges in collaborative mixed-reality. *Frontiers in ICT*, *6*, 5., https://doi.org/10.3389/fict.2019.00005

Madanchian, M. (2024). Generative AI for consumer behavior prediction: Techniques and applications. *Sustainability*, *16*(22), 9963. https://doi.org/10.3390/su16229963

Deppe, C., & Schaal, G. S. (2024). Cognitive warfare: a conceptual analysis of the NATO ACT cognitive warfare exploratory concept. *Frontiers in Big Data*, *7*, 1452129. doi: 10.3389/fdata.2024.1452129

Byman, D. L., Gao, C., Meserole, C., & Subrahmanian, V. S. (2023). Deepfakes and international conflict (Vol. 8). Washington, DC: Brookings Institution., https://www.brookings.edu/wp-content/uploads/2023/01/FP_20230105_deepfakes_international_conflict.pdf#:~:text=On%20March%202%2C%202022%2C%20shortly,instead%20implored%20them%20to%20lay

Perrigo, Billy "Inside Facebook's African Sweatshop," Time, February 14, 2022, https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/

García Luengo, J. (2025). Transparency Obligations for Providers and Deployers of certain AI Systems (Chapter IV)

https://digibuo.uniovi.es/dspace/bitstream/handle/10651/78956/TransparencyObligationsLuengo.pdf?sequence=1