

Bányász - Váczi Kincső Boróka

Feladó: ESSACHESS JCS <essachess@gmail.com>
Küldve: 2025.március.26. 18:22
Címzett: Bányász - Váczi Kincső Boróka
Tárgy: Re: Full paper submission

Kategóriák: Kutatás

Dear author,
Your full article has been received. We will contact you after the peer review process is finished.
Kind regards,

Essachess - Journal for Communication Studies
Parrainée par l'Agence Universitaire de la Francophonie
ECREA Associate Member
IARSIC-CORHIS EA 7400, Paul Valéry University of Montpellier
<http://www.essachess.com/index.php/jcs>
essachess@gmail.com

Le jeu. 20 mars 2025 à 13:44, Bányász - Váczi Kincső Boróka <Vaczi.Kincso.Boroka@uni-nke.hu> a écrit :

Dear Editors,

In agreement with Mónika Andok, the guest editor of the special issue, I am sending the full paper for the *Call for Papers for Volume 18, Issue 2(36) of 2025*.

I have attached the title page and the anonymized manuscript.

We look forward to your feedback and editorial review.

Yours sincerely,

Mrs. Kincső Boróka BÁNYÁSZ-VÁCZI

Ezen üzenet és annak bármely csatolt anyaga személyes adatokat és bizalmas információkat tartalmaz, így törvényi védelem alatt áll. Megismerése, nyilvános közlése, továbbítása csak erre külön feljegyzítés után engedélyezett. Az üzenetet kizárólag a címzett, illetve az általa meghatalmazottak ismerhetik meg, használhatják fel. Ha Ön nem az üzenet címzettje, úgy kérjük, értesítse erről az üzenet küldőjét és törölje az üzenetet, valamint annak összes csatolt mellékletét a rendszeréből.

This message and any files transmitted with it contain personal data and confidential information, therefore, it is under legal protection. Disclosing, copying, distributing or taking any action in reliance on the contents of this information is strictly prohibited unless one is authorized to take such action. The message is intended solely for the

use of the individual or entity to whom it is addressed. Please notify the sender immediately if you have received this message by mistake and delete this message and all of its attachments from your system.

The capability to identify images generated by artificial intelligence

Péter BÁNYÁSZ*

Associate Professor, PhD, Ludovika University of Public Service,
Department of Cybersecurity, HUNGARY
ORCID: 0000-0002-7308-9304, e-mail: banyasz.peter@uni-nke.hu

Ákos ORBÓK

Assistant Professor, PhD, Ludovika University of Public Service,
Department of Cybersecurity, HUNGARY
ORCID: 0000-0002-6085-0598, e-mail: orbok.akos@uni-nke.hu

Sándor MAGYAR

Associate Professor, PhD, Ludovika University of Public Service,
Department of Military National Security, HUNGARY
ORCID: 0009-0008-4824-0923, e-mail: magyar.sandor@uni-nke.hu

Rebeka TÓTH

PhD student, University of Oslo, NORWAY
ORCID: 0009-0000-9574-1896, e-mail: rebekat@ifi.uio.no

Kincső Boróka BÁNYÁSZ-VÁCZI

Researcher, Ludovika University of Public Service, HUNGARY
ORCID: 0009-0008-4824-0923, e-mail: vaczi.kincso.boroka@uni-nke.hu

Ágnes VESZELSZKI

Associate Professor, PhD, Ludovika University of Public Service,
Department of Digital Media and Communication, HUNGARY
ORCID: 0000-0003-2226-0607, e-mail: veszelszki.agnes@uni-nke.hu

Gábor István KOVÁCS

Associate Professor, PhD, Ludovika University of Public Service,
Department of Digital Media and Communication, HUNGARY
ORCID: 0000-0003-3680-8311, e-mail: kovacs.gabor.istvan@uni-nke.hu

* „Supported by the EKÖP-24-4-II-23 University Research Scholarship Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund.”

The capability to identify images generated by artificial intelligence

Abstract: Social media profoundly influences modern society, offering both opportunities and risks for communication. This paper examines the ability to identify AI-generated images through an online survey, investigating factors that affect participants' discernment between authentic photographs and AI-generated images. The findings highlight perceptual biases linked to generative AI technology, impacting perceptions of authenticity. The findings emphasize the importance of promoting media literacy and visual critical thinking, particularly regarding generative AI, as a key strategy in combating disinformation.

Keywords: generative artificial intelligence, survey, media literacy, cybersecurity education, critical thinking

La capacité d'identifier des images générées par l'intelligence artificielle

Résumé: Les médias sociaux influencent profondément la société moderne, offrant à la fois des opportunités et des risques pour la communication. Cet article examine la capacité à identifier les images générées par l'IA au moyen d'une enquête en ligne, en étudiant les facteurs qui affectent le discernement des participants entre les visuels authentiques et ceux générés par l'IA. Les résultats mettent en évidence les biais perceptifs liés à la technologie générative de l'IA, qui ont un impact sur la perception de l'authenticité. L'éducation aux médias et la pensée critique visuelle sont essentielles pour lutter contre la désinformation générée par l'IA.

Mots-clés : intelligence artificielle générative, enquête, éducation aux médias, éducation à la cybersécurité, pensée critique

The capability to identify images generated by artificial intelligence

Abstract: Social media profoundly influences modern society, offering both opportunities and risks for communication. This paper examines the ability to identify AI-generated images through an online survey, investigating factors that affect participants' discernment between authentic photographs and AI-generated images. The findings highlight perceptual biases linked to generative AI technology, impacting perceptions of authenticity. The findings emphasize the importance of promoting media literacy and visual critical thinking, particularly regarding generative AI, as a key strategy in combating disinformation.

Keywords: generative artificial intelligence, survey, media literacy, cybersecurity education, critical thinking

La capacité d'identifier des images générées par l'intelligence artificielle

Résumé: Les médias sociaux influencent profondément la société moderne, offrant à la fois des opportunités et des risques pour la communication. Cet article examine la capacité à identifier les images générées par l'IA au moyen d'une enquête en ligne, en étudiant les facteurs qui affectent le discernement des participants entre les visuels authentiques et ceux générés par l'IA. Les résultats mettent en évidence les biais perceptifs liés à la technologie générative de l'IA, qui ont un impact sur la perception de l'authenticité. L'éducation aux médias et la pensée critique visuelle sont essentielles pour lutter contre la désinformation générée par l'IA.

Mots-clés : intelligence artificielle générative, enquête, éducation aux médias, éducation à la cybersécurité, pensée critique

Introduction

Social media has become essential to modern life, significantly influencing our homes, workplaces, and social interactions. It has transformed communication, information sharing, and discourse, leading to unavoidable social changes. Social media serves various purposes, such as connecting with others, expanding business opportunities, advocating for issues, and managing crises through quick information dissemination. However, it also presents challenges and risks for defense, law enforcement, and national security organizations. While social media threats to information security exist, it offers new avenues to enhance organizational effectiveness. Disseminating disinformation on these platforms can disrupt societal dynamics and individual decision-making. A key concern stems from algorithms prioritizing attention-grabbing content, which is often polarizing and promotes the spread of misleading information, especially during socio-political or health crises. (Cinelli et al., 2020) The threat of disinformation is especially acute in contexts where accurate and trustworthy information is paramount, such as during electoral cycles or public health emergencies. In these instances, the rapid spread of fake news and distorted narratives can skew public perceptions and shape community attitudes and behavioral patterns. Social media thus functions as a conduit for the swift transmission of emotionally charged information, frequently evoking fear or anger, which significantly aids in accepting disinformation as a form of reality. (Vosoughi et al., 2018)

The digitization of the information landscape, coupled with the proliferation of social media, has led to an increasingly prevalent phenomenon known as information disorder. Claire Wardle and Hossein Derakhshan articulated this concept in their 2017 study, wherein they critique the term fake news frequently encountered in public discourse as inaccurate and misleading. (Wardle & Derakhshan, 2017) They advocate for a more nuanced framework to categorize online misinformation, delineating three principal types of information disruption: misinformation, disinformation, and malicious information. (COE) A comparable methodology has been adopted by the European Union's StratCom working group, which directs its strategic communication efforts towards mitigating misinformation.

Misinformation denotes the distribution of inaccurate or misleading information without the intent to inflict harm. In such instances, the individual disseminating the information may inadvertently fall victim to an error; for example, during a natural disaster, someone might share exaggerated yet false reports in good faith due to insufficient verification of the source's reliability. In contrast, disinformation entails intentionally disseminating fabricated or manipulated information maliciously. This communication is typically pursued to achieve political, economic, or ideological objectives; for instance, a state-sponsored propaganda outlet may intentionally propagate conspiracy theories or false narratives to manipulate public sentiment and erode trust in societal institutions. The third classification, malinformation, involves intentionally releasing authentic, factual information within a context intended to cause harm. Such scenarios may include the unauthorized leaking of confidential data or the contextual misrepresentation of personal information to inflict reputational damage or harm to the individuals concerned. (Konieczniak, 2023)

It is vital to emphasize that these three categories embody distinct motivations and pose varying threats, necessitating tailored responses. Misinformation frequently arises from individuals failing to adequately verify the source or credibility of information, resulting in the inadvertent dissemination of inaccuracies. In contrast, disinformation is characterized by the purposeful propagation of false information, which may serve as a component of governmental propaganda or as a mechanism for economic profit, exemplified by the circulation of clickbait articles to generate financial revenue. Malinformation often raises significant ethical and legal considerations, as its foundation may be factual; however, its distribution can infringe upon privacy rights or pose national security risks, as illustrated by the leakage of sensitive intelligence. While the term "fake news" is often employed as a catch-all descriptor for these three phenomena, experts from the Council of Europe advocate for its avoidance due to its inaccuracy and use by specific political figures as a derogatory tool to undermine legitimate news that contradicts their agendas. Consequently, contemporary academic and policy discussions increasingly favor the utilization of the more precise classifications outlined above.

The phenomenon of information disorder is intricately connected to several related concepts, including fake news, propaganda, and psychological operations. (Tandoc Jr. et al., 2018) Propaganda is traditionally understood as one-sided, manipulative communication that distorts reality—often utilizing half-truths or outright disinformation—to persuade the public for political or ideological ends. A shared characteristic between the propaganda employed by totalitarian regimes in the 20th century and contemporary extremist online campaigns is the selective presentation or distortion of reality to sway mass opinions. (McLarin, 1982) Psychological operations, and more broadly, information operations, constitute components of this propaganda arsenal; in military and geopolitical conflicts, information—regardless of its veracity—can function as a weapon intended to deceive, demoralize, or influence adversaries. In the context of hybrid warfare strategies, such as those implemented by Russian operations, cyber warfare, disinformation campaigns, and propaganda initiatives collectively present a multifaceted threat. Within this framework, disseminating fake news and conspiracy theories can be interpreted as psychological operations designed to fracture societal cohesion and erode trust in democratic institutions. It is not coincidental that fake news, propaganda, and misinformation have the potential to polarize public opinion, reinforce extremist ideologies, and ultimately jeopardize the stability of democratic systems.

While this paper primarily examines the implications of AI-generated images in the realm of disinformation, it is crucial to acknowledge that these images present various additional cybersecurity threats. Notably, they can be utilized in social engineering and phishing attacks, particularly through the use of deepfake technologies, which can either compromise real individuals or fabricate misleading narratives. Furthermore, such images can undermine authentication security by bypassing biometric identifiers and deceiving facial recognition systems. The generation of counterfeit profiles may facilitate online identity theft and industrial espionage. Additionally, AI-driven image generation holds the capacity for steganographic attacks, wherein malicious code may be stealthily integrated into image files, as well as the potential to produce disruptive patterns that impair machine vision systems. These multifaceted risks underscore the necessity of

implementing advanced detection mechanisms and enhancing cybersecurity awareness to mitigate the threats posed by visual manipulation.

Recent technological advancements, particularly in generative artificial intelligence (GenAI), have heightened our vulnerability to threats in cyberspace, thereby exacerbating the cybersecurity risks associated with social media platforms. Among the most pressing concerns is the capacity of GenAI technologies to facilitate the creation of increasingly sophisticated and credible phishing campaigns. These generative models can produce personalized messages in natural language or fabricate identities to establish trust with users, thereby enhancing the probability of illicit access to sensitive data. (Hancock et al., 2020)

GenAI also has been a great help to software development in code completion, code generation, or creating software specifications, and is a widely used technique—however, Toth et. Al. Evaluates LLM-Generated PHP small web applications that unveil a high percentage of vulnerabilities and highlight limitations in the complexity of LLM-created content. (Tóth et al., 2024) Thianyi et al. introduce the FormAI dataset, comprising 112,000 AI-generated, compilable C programs labeled for vulnerabilities. Using GPT-3.5-turbo with a dynamic zero-shot prompting technique, the dataset includes programs of varying complexity, from simple string manipulations to complex tasks like encryption. Vulnerabilities were identified using the Efficient SMT-based Bounded Model Checker (ESBMC), which employs formal verification methods to detect issues without false positives. Notably, 51.24% of the generated programs contained vulnerabilities, highlighting significant risks in AI-generated code. (Tihanyi et al., 2023) Chen et al. Their DiverseVul research creates a comprehensive dataset to enhance deep learning models for vulnerability detection in source code. The authors curated this dataset by crawling security issue websites and extracting vulnerability-fixing commits, resulting in 18,945 vulnerable functions spanning 150 Common Weakness Enumerations (CWEs) and 330,492 non-vulnerable functions from 7,514 commits. This dataset covers 295 more projects than all previous datasets, offering a broader scope for training and evaluating deep learning models. Their analysis indicates that, despite the dataset's diversity, deep learning models still face challenges in accurately detecting vulnerabilities, mainly due to high false positive rates and difficulties in identifying complex CWEs. (Chen et al., 2023)

In addition, GenAI technologies can be employed to disseminate disinformation. By automatically generating and circulating fake news and propaganda, these technologies enable attackers to manipulate public opinion, which poses significant risks during election cycles or periods of social unrest. Another notable risk is the utilization of deepfake technology, which permits the generation of highly realistic images, audio clips, or videos. This capability can lead to various forms of abuse, including creating fraudulent videos for blackmail or defamation purposes. (Mirsky & Lee, 2021)

The Dynamic Intelligence Assessment (DIA) framework evaluates large language models (LLMs) by introducing dynamic question templates across various domains, including mathematics, cybersecurity, and computer science. It proposes four key metrics—Reliability Score, Task Success Rate, Confidence Index, and Near Miss Score—to measure models' reliability and confidence beyond traditional benchmarks. The DIA-Bench dataset, consisting of 150 diverse and challenging task templates, was used to evaluate eight state-of-the-art LLMs,

revealing that most struggle with consistency, especially in mathematical reasoning. Results indicate that ChatGPT-4o, which has tool-using capabilities, outperformed GPT-4o API and other models, achieving a Reliability Score of -64 compared to -208 for GPT-4o API. It also had a higher Confidence Index (38.67%), meaning it consistently solved more problems correctly across variations. Other models, such as Gemini-1.5-pro and Mistral 7B, performed significantly worse, with near-zero Confidence Index values and high rates of incorrect answers. The study highlights that models without tool-using abilities tend to hallucinate more frequently, leading to unreliable performance. Ultimately, the DIA framework provides a more comprehensive evaluation of LLMs, emphasizing the need for adaptive intelligence and self-assessment in future AI development. (Tihanyi et al., 2024)

A critical challenge in developing generative AI models lies in the quality and integrity of the data utilized. If the reliability of the underlying data is not assured, there exists a risk that artificially generated content may be used in training models, leading to distortions in the learning process and a decline in model performance. The adoption of blockchain technology presents a potential solution to ensure data integrity by allowing information to be stored decentralized and transparently, thus mitigating the risk of manipulation. The convergence of blockchain and artificial intelligence offers promising avenues for the secure management of data and the reliability of these models. (Eszteri, 2020)

The ability to differentiate between images generated by artificial intelligence and those produced by humans presents a multifaceted challenge, shaped by advancements in technology and psychological, cognitive, and ethical considerations. Accurate identification of AI-generated images is crucial for mitigating misinformation and addressing ethical dilemmas. (Gupta et al., 2024) In professional applications, in which the distinction between authentic photographs and AI-generated images are of crucial importance, a range of technological solutions are available which have been shown to perform this task at a considerable level of reliability. For instance, research has employed a variety of datasets and convolutional neural network architectures to enhance the detection of AI-generated imagery, achieving accuracy rates as high as 97.29%. (Gupta et al., 2024) (Chinta et al., 2024) (Bird & Lotfi, 2024) (Nayim et al., 2024) Moreover, the emotional content of images, particularly the presence of positive emotions, can significantly influence individuals' ability to discern AI-generated images. (Park et al., 2024)

Recent studies have introduced large-scale datasets to improve the detection of AI-generated images. For instance, the GenImage dataset comprises over one million pairs of AI-generated images and authentic photographs synthesized using advanced diffusion models and GANs, providing a comprehensive benchmark for detection methods. (Zhu et al., 2023) Similarly, the RU-AI dataset offers a multimodal collection across text, image, and voice, facilitating robust detection of machine-generated content. (Huang et al., 2025) However, challenges persist due to biases in existing datasets, mainly related to JPEG compression and image size, which can affect detector performance. Addressing these biases is crucial for developing more robust detection systems. (Grommelt et al., 2024) In addition, recent advancements in AI also have greatly improved the ability to detect AI-generated images, addressing the challenges of increasingly sophisticated synthetic content. One notable development is Google's introduction of SynthID, a digital

watermarking system integrated into images edited using the Magic Editor's generative AI feature in Google Photos. (Weatherbed, 2025) Created by Google's DeepMind team, SynthID embeds metadata within images, allowing AI-generated or altered content to be identified without affecting the image's appearance. This system aims to help users distinguish between authentic and AI-manipulated visuals, combating the risk of misleading imagery. The Content Authenticity Initiative (CAI), a coalition of over 3,000 tech and media companies, has introduced Content Credentials. This system uses watermarks and metadata to verify digital image authenticity and flag AI edits. For example, a participating digital camera can embed metadata into an image, and software like Adobe Photoshop can track any AI modifications. Content Credentials serve as a "nutritional label" for digital content, solving the growing problem of fake images and misinformation. (Wilser, 2024) These technological innovations mark important progress in maintaining the integrity of visual content. By adopting such systems, the industry aims to provide users with tools to identify authentic media, reducing the impact of AI-generated misinformation in the digital world.

Additionally, psychological and cognitive factors are integral to this discernment process. Emotions, human behavior, and personal attributes—such as creative identity and openness—impact the perception and assessment of AI-generated artworks. (Park et al., 2024) (Grassini & Koivisto, 2024) In a specific study, participants struggled to distinguish between human-created and AI-generated images consistently; however, they tended to harbor negative biases toward the latter, particularly when evaluations were based on subjective source attributions. (Grassini & Koivisto, 2024) From an ethical perspective, the rapid advancement of AI-generated images raises significant concerns regarding the spread of disinformation and the potential for fraudulent schemes. This reality underscores the necessity for developing effective methodologies to detect AI-generated content. (Li et al., 2024) Understanding how advancements in AI technology impact human perception is particularly crucial in rapidly evolving domains such as digital food marketing. (Califano & Spence, 2024)

Technological progress in artificial intelligence has facilitated the generation of high-quality synthetic images nearly indistinguishable from authentic photographs, presenting formidable challenges in discerning between artificial and genuine stimuli. (Bird & Lotfi, 2024) Creating synthetic datasets and implementing sophisticated convolutional neural network architectures have markedly enhanced the recognition performance of AI-generated images, achieving accuracy rates of up to 98.49% on test datasets. (Nayim et al., 2024) Introducing new testing datasets, such as FakeGPT and PFake, in conjunction with the Meta Ensemble eXplainable Fake Image Classifier (MEXFIC), has contributed to substantial advancements in AI-generated image recognition. (Islam et al., 2025)

Deepfakes are becoming increasingly sophisticated, with their content often nearing indistinguishability from authentic sources. The generation of artificial intelligence images is becoming progressively accessible to a broader audience as various platforms and software proliferate. However, akin to many innovative technologies, the potential applications of generative AI are not limited to constructive uses; they also include the capacity to mislead, manipulate, and deceive, thereby presenting significant ethical and social dilemmas. Users may easily be deceived by a convincingly crafted fake video or image, mainly in a

strategic context, which can lead to the wrongful implication of innocent individuals, discrediting reputable figures, or manipulating public policy. Moreover, such technology can facilitate financial fraud, significantly eroding user trust in online platforms. The concurrent advancements in Deep Learning, Big Data, and image processing have unveiled new possibilities in artificial intelligence, yet this progress also risks exacerbating the dissemination of disinformation. (Gambín et al., 2024) The proliferation of increasingly advanced algorithms and extensive databases for image manipulation has made these tools more accessible to users. Consequently, the production of deepfake videos and images is no longer restricted to highly skilled professionals; instead, it is now within the reach of a broader audience, which significantly heightens the potential for malicious misuse. Furthermore, the realism of altered content has advanced so that it frequently becomes nearly indistinguishable for the average viewer from genuine footage. As AI-driven generative models evolve in complexity, the probability of disseminating entirely fabricated content that convincingly resembles reality also escalates. (Abbas & Taeiagh, 2024) These technologies enable malicious users to alter images and videos with a level of credibility that was previously inconceivable. The proliferation of deepfake images, audio, and videos is occurring at an alarming rate; however, their primary intent is frequently to mislead, manipulate, or undermine the credibility of individuals, thereby presenting a significant risk within the digital landscape.

Social media platforms significantly contribute to the dissemination of deepfake images. A diverse array of users can effortlessly access and share such content with a single click. Furthermore, algorithm-driven user experience systems frequently prioritize and promote materials likely to capture attention or elicit emotional responses, thereby facilitating manipulated images' rapid and intense reach to their intended audiences. Since individuals often utilize social media for entertainment or information, they may not invest substantial effort in verifying the authenticity of the content they encounter, resulting in a proportionate increase in the risk of deception. Consequently, content generated through deepfake technology can potentially mislead the online community on a large scale. The swift sharing capabilities inherent to social media platforms contribute to the widespread and often uncontrolled propagation of fake news and misinformation across the internet. Additionally, these technologies are intentionally employed for malicious objectives—such as harvesting personal data, compromising public figures, or fomenting societal divisions—resulting in potentially serious long-term repercussions. In the current landscape, distinguishing between authentic and artificially created content is exceedingly challenging, and disinformation can proliferate rapidly across online platforms.

The quality of deepfakes is continuously improving; however, detection methods remain nascent (Singh et al., 2023), making it exceedingly challenging to accurately identify such content and swiftly filter out misleading manipulations within the digital media landscape. This limitation exacerbates the rapid dissemination of disinformation and undermines social trust. The domains of deepfake creation and detection encompass a broad spectrum of challenges. (Chauhan et al., 2022) While significant advancements have been made, further progress is imperative. Both developers and malicious actors are employing increasingly sophisticated techniques to enhance the realism of manipulations or to expedite their detection. The swift evolution of technology perpetually introduces new tools and

methodologies for perpetrators and those seeking to implement defensive measures. This creates a competitive landscape in which increasingly advanced deepfake content emerges, complicating detection efforts. Nonetheless, widely accessible and effective detection technologies remain far from realization.

The proliferation of disinformation in the digital realm—encompassing the manipulation of public opinion, exerting political influence, incitement of public panic, facilitation of cybercriminal activities, or the mere pursuit of attention—often occurs insidiously. Furthermore, such fabricated information can be disseminated rapidly via social media platforms, where shares and reposts often accumulate without drawing sufficient scrutiny. Consequently, these narratives can profoundly influence individuals' decisions and behaviors. Strategically timed or intricately crafted instances of fake news, including deceptive imagery, can disrupt social discourse, erode public trust, and yield enduring adverse effects in both political and economic contexts.

Images exert a particularly profound emotional influence. A thoughtfully selected visual element can markedly alter the perception of a news item or narrative, affecting individuals' emotional responses. Thus, it is critically important to prioritize the identification of fake or manipulated visual content and to understand the motivations behind such deception from the outset. Individuals may inadvertently mislead themselves and others without a deliberate effort to verify this content and seek credible sources, as partial truths and manipulations can rapidly integrate into public consciousness. As the volume of information escalates, analyzing visual materials with the same rigor applied to textual content becomes imperative, employing increasingly advanced technologies to detect image manipulation. This approach will aid in mitigating the proliferation of false content on social media and other online platforms, particularly in the realms of misleading information and disinformation.

Individuals frequently believe they are less susceptible to deepfakes than others, wrongly assuming their ability to discern authentic content from fraudulent representations. However, such overconfidence can be perilous, as advancements in artificial intelligence have rendered manipulation techniques increasingly sophisticated, making it relatively easy for even the most discerning observer to be deceived. Deepfakes can propagate misinformation, presenting a considerable threat to public opinion, particularly within a political context. (Ahmed, 2023) A meticulously crafted fake video can influence electoral outcomes, diminish public trust, and potentially undermine the credibility of politicians. The characteristics of misinformation tend to evolve rapidly, rendering features effective in identifying such content today potentially obsolete tomorrow. Creators of deepfakes are continually innovating and refining their tools, necessitating practitioners to remain in a constant state of adaptation. (Pennycook & Rand, 2021) In a defensive capacity, it is imperative to develop new, adaptive techniques and strategies to respond to the dynamically shifting landscape of threats effectively. A significant limitation is that applications typically do not explicitly request information regarding the source when users upload an image, which could help clarify any ambiguities. (Grut, 2024) This lack of transparency complicates verifying the authenticity of uploaded content, as users often remain uncertain about whether an image originates from a legitimate source or is a manipulated, artificially generated version.

The advancement of technological tools and filtering solutions is crucial for the prompt detection of deepfake images. Regarding protective measures, the greater the sophistication of the technologies and methodologies employed, the more challenging it becomes for individuals intent on disseminating manipulated material. Nevertheless, these innovative technologies alone will not suffice if users lack the knowledge and critical judgment to evaluate digital content effectively. Therefore, it is imperative to equip content consumers with information that elucidates the risks associated with deepfake images and aids in identifying misleading tactics. Educational initiatives and awareness campaigns can play a vital role in preventing or mitigating the probability of individuals being misled by fabricated images.

Automated fact-checking of images necessitates a comprehensive, interdisciplinary approach that parallels the complexity of the technologies employed to create manipulated visual content. (Kavtaradze, 2025) On the one hand, it involves leveraging the latest advancements in machine learning, computer vision, and artificial intelligence to ensure the system can identify fake or altered image components with high accuracy and speed. On the other hand, it is crucial to engage psychologists, communication experts, and ethics researchers who examine users' media consumption behaviors, methods of deception, and the societal impact of such content. An interdisciplinary approach can facilitate the development of automated image fact-checking as a technological endeavor and as a practical tool that effectively aids the community in detecting and understanding misleading, fabricated news or manipulative material.

The research thus highlights the challenges associated with recognizing images generated by artificial intelligence. The data and analyses provided are essential for gaining insights into users' reactions and formulating strategies to enhance the detection of AI-generated images and bolster defenses against misleading content. Moreover, the study will examine not only the perception of the authenticity of images and the ability to identify manipulated content but also how variables such as gender and age of individuals depicted in the images influence perception. By investigating these aspects in depth, we can advance our understanding of the factors that increase susceptibility to deception and develop more effective methodologies for recognizing manipulated or artificially created content.

This paper presents the findings from a questionnaire survey conducted among students at the Ludovika University of Public Service (LUPS). The questionnaire addressed two interrelated topics: first, it explored individuals' capabilities and willingness to identify and respond to online threats within the context of social media, framed by the motivational background of active defensive behaviours online; second, it examined the skills required to recognize images generated by generative artificial intelligence.

The research objectives were:

RO1: To evaluate the capacity of non-specialists to accurately distinguish between genuine photographs and photo-realistic images produced by artificial intelligence, as well as to identify the factors that influence their decision-making processes.

RO2: To investigate whether the accuracy in distinguishing between authentic photographs and those generated by artificial intelligence is influenced by the age

and gender of the subjects depicted in the images, and to analyze potential biases in perceptual judgment.

RO3: To identify and analyze the primary visual cues that viewers typically employ when intuitively evaluating the authenticity of an image, and to ascertain which features most significantly influence their decision-making processes.

Methodology

The questionnaire was informed by three foundational studies, adapting the methodological insights from the works of Nurse, Maples-Keller et al., and Boehmer et al. to offer a novel perspective on the interplay between social media, generative AI, and cybersecurity. (Nurse, 2019) (Maples-Keller et al., 2019) (Boehmer et al., 2015) Due to constraints in scope, this paper reports exclusively on the results of recognizing generative AI images. The analysis concerning the ability to identify images produced by generative AI contributes valuable insights for cybersecurity education and efforts to combat disinformation. This emphasis is particularly pertinent in light of recent technological advancements in AI and their implications for society. We organized our questionnaire survey under the title "Misinformation and Abuse on Social Media Platforms," targeting students from the LUPS, given the institution's specific mission as outlined previously. While we distributed the online survey exclusively among students at LUPS, the sample of respondents encompassed a broader range of demographic groups due to participants sharing the questionnaire with family members and other acquaintances

The questionnaire encompassed five primary sections. The time allocation was crucial, as we aimed to assess the ability to recognize generative AI without imposing time constraints, allowing respondents to examine the images at their own pace. The questionnaire was conducted anonymously, ensuring no personal data was collected from participants following GDPR. Responses were analyzed using the statistical software SPSS, which securely stores the data in a non-personally identifiable format, preventing any linkage to individual respondents. Participants were given a four-month window, from June to September 2024, to complete the questionnaire. During this period, 216 responses were recorded; however, one submission was excluded from the dataset due to an implausible age of 120 years, leading us to infer that the remainder of the responses may have been answered with a similar lack of seriousness.

Our research investigated the factors that affect the recognition of authentic images produced by artificial intelligence. The authentic images were curated from a complimentary stock photo database featuring portraits of individuals aged 20, 40, and 60. The AI-generated images were created using the Playground AI platform in the spring of 2024, adhering to the technological standards prevalent at that time and maintaining identical technical parameters to those of the original images. Utilization of the premium version facilitated more intricate adjustments. A consensus was reached to select the most realistic images presented in the questionnaire in a randomized order based on established criteria. The algorithm employed ensured that images of the same gender, age group, or category were not excessively repeated, thereby minimizing the potential bias introduced by habituation and serial effects. The complete array of visual stimuli utilized in the study is summarized in Figure 1, presented in the form of thumbnails. Respondents were given unlimited time to review the images before assessing their authenticity

using a six-point Likert scale ranging from "Almost certain this is an authentic photographs" to "Almost certain this is an AI-generated image." The responses indicated the participants' skepticism regarding the authenticity of authentic photographs and AI-generated images. If respondents refrained from selecting the "Almost certain this is an authentic photographs" option, they were prompted to briefly explain the specific elements within the image that they found suspicious or indicative of being generated by AI.

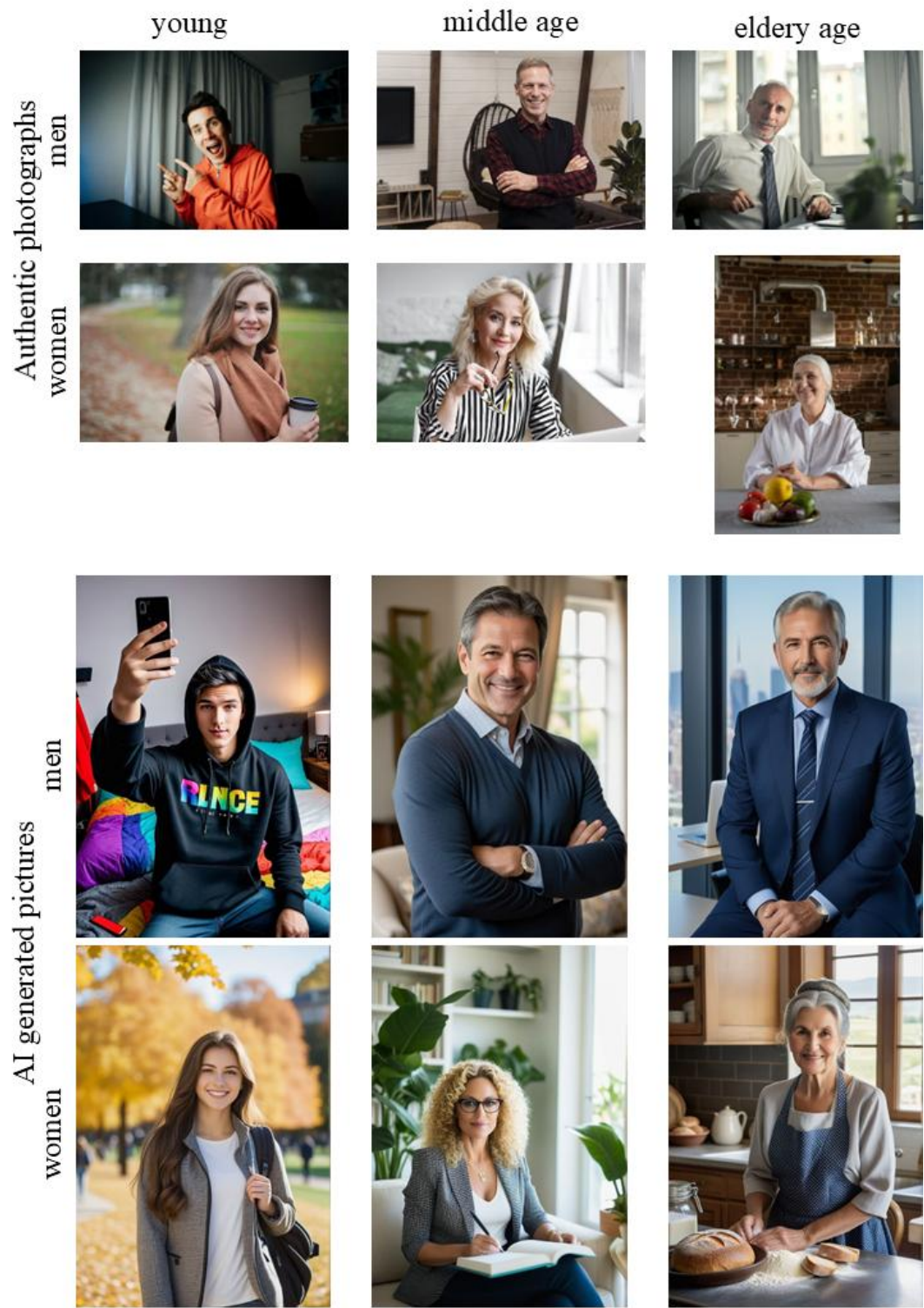


Figure 1. AI-generated images and authentic photographs

Results

The demographic analysis of the survey reveals a nearly balanced gender distribution among respondents, albeit with a slightly higher representation of males (119 men compared to 96 women). The distribution of age ($M = 29.0$, $SD = 10.2$) was considerably right-skewed, the sample predominantly comprising individuals aged 18–25 (52.6%), with a notable concentration within the 20–25 age range (44.2%), and a marked decline in participation from older demographics. Regarding educational qualifications, most respondents possess a school-leaving certificate (43.7%), while those holding a bachelor's degree constitute 40%, and those with a master's degree make up 14.4%. Regarding residency, nearly half of the participants (47.4%) reside in the capital, with the remainder primarily located in duchy towns, smaller urban centers, or rural areas. Concerning employment status, a substantial portion of respondents are higher education students (69.8%), and 50.2% are engaged in gainful employment. Within the public sector workforce, a predominant share is represented by individuals from law enforcement (16.7%), while smaller proportions emerge from the defense, public administration, and educational sectors. Furthermore, less than half of the respondents (43.7%) have previously participated in cybersecurity or data protection coursework, which is pertinent experience for this investigation.

Our findings indicate that respondents can reliably differentiate between authentic photographs and those generated by currently accessible AI tools, represented explicitly in our study by the subscription version of the Playground AI platform, which epitomizes the current state of the art for users. On a six-point scale, the average suspicion level toward AI-generated images was 4.74 ($SD: 0.98$), whereas the average level of unjustified suspicion for authentic images was significantly lower at 2.27 ($SD: 0.85$). Results from a matched-sample t-test revealed this difference to be statistically significant: $t(214) = 28.433$; $p < 0.001$. Furthermore, the correlation between suspicion levels for the two types of stimuli was found to be non-significant and negligible ($r = 0.034$), suggesting no overarching tendency among respondents to exhibit suspicion regardless of the image type. The considerable disparity in perception between the two stimulus categories was characterized by a large effect size (Cohen's $d = 1.939$) and visually confirmed by a lack of overlap between the 95% confidence intervals for mean suspicion under the two conditions. (Figure 2)

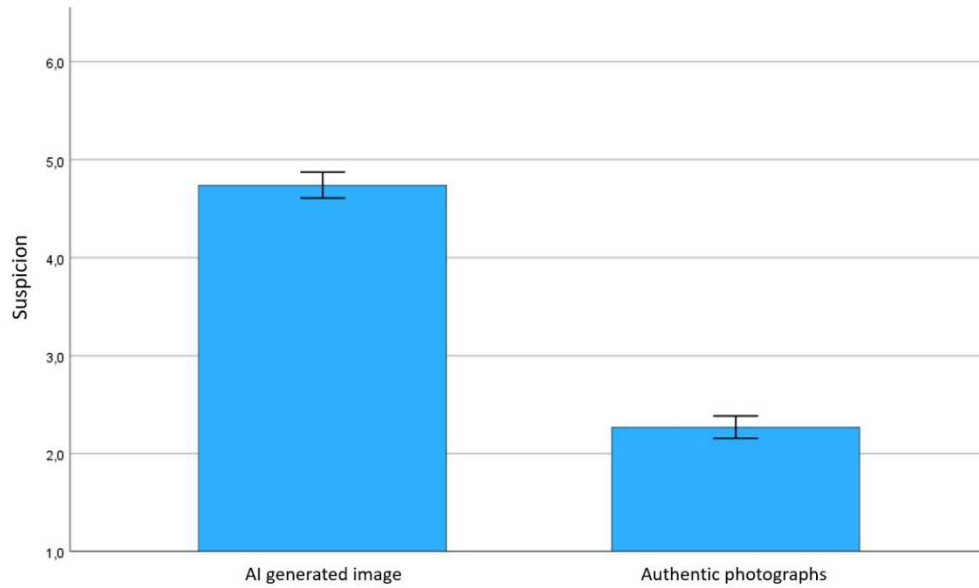


Figure 2. *The difference between the perception of AI-generated images and authentic photographs*

The disparity in responses to the two categories of stimuli is distinctly evident in the distribution of perceived suspicion. (Figure 3) The distribution associated with AI-generated stimuli exhibited a leftward skew, whereas that corresponding to authentic images displayed a rightward skew.

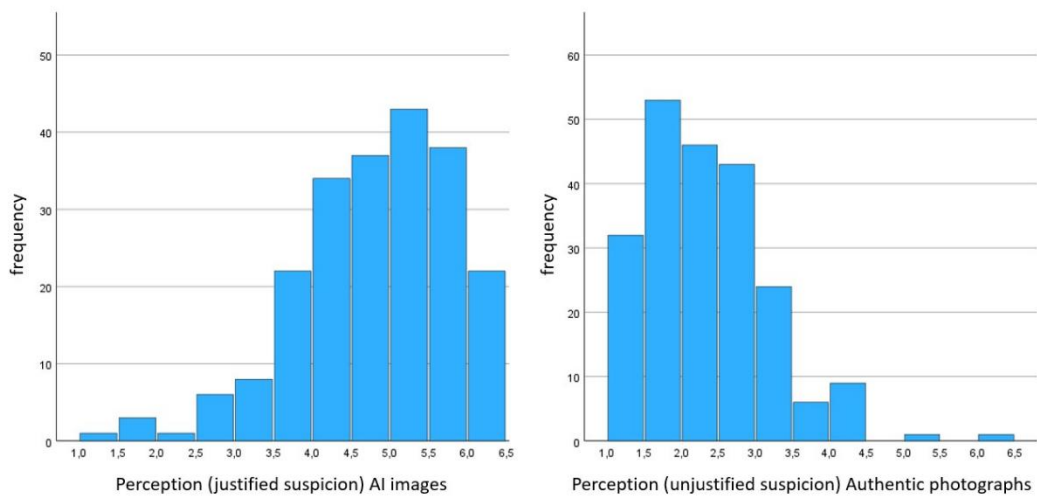


Figure 3. *Distribution of suspicion levels for AI-generated images and authentic photographs*

The efficacy of detecting machine-generated images, or the extent of justified suspicion, varied across different stimuli. Participants were most inclined to regard the image of an elderly woman as suspicious, with a mean suspicion rating of 5.44. In addition, the probability of identifying machine-generated features in the images of a young man and an elderly man was relatively comparable, with mean suspicion scores of 5.10 and 5.07, respectively. Notably, the image of a middle-aged woman posed the most significant challenge, yielding an average suspicion score slightly above the midpoint of 3.5 on the rating scale (mean: 4.01). Overall, however, it can

be observed that a considerable proportion of respondents opined that the images were likely not authentic. The average suspicion levels associated with the six stimuli can be quantified using 95% confidence intervals. (Figure 4)

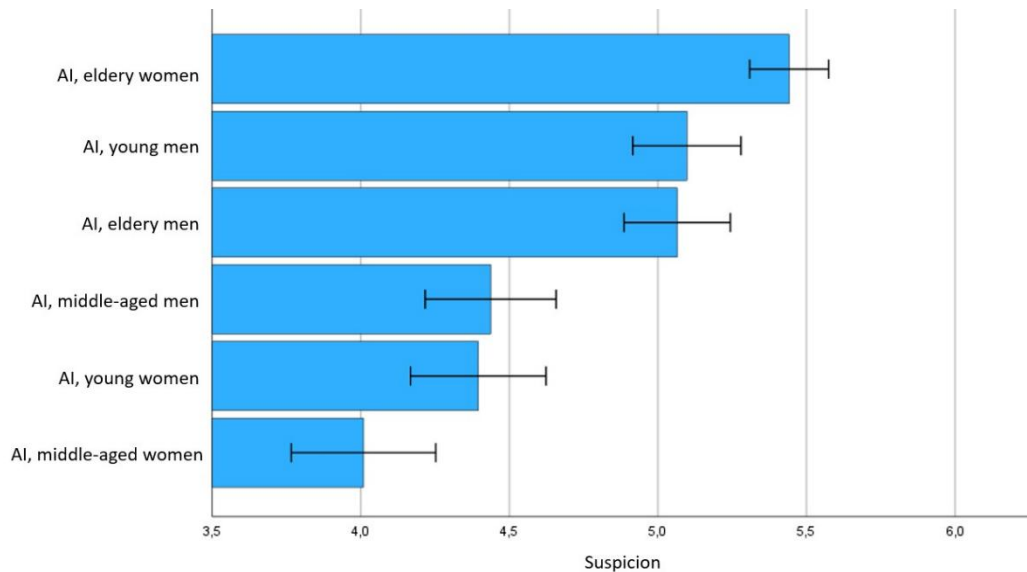


Figure 4. *The average level of suspicion for the six AI-generated images*

Nonetheless, suspicion is not solely an intriguing phenomenon associated with AI-generated stimuli; the emergence of this technology has also engendered a perceptual bias in the opposite direction. This bias leads some individuals to erroneously perceive typical characteristics of AI algorithms as inherent in authentic photographs, thus prompting skepticism regarding the authenticity of actual images. In our study, as previously discussed, the overall level of such unjustified suspicion is relatively low (average: 2.27). However, a closer examination of individual images reveals variability in this perception. Notably, respondents exhibited the strongest inclination to believe that the photograph of a authentic young man could have been generated by AI, with a mean suspicion rating of 3.53. In contrast, for the other images, participants expressed high confidence in their authenticity, with suspicion scores ranging from 1.79 to 2.29. A 95% confidence interval characterizes this pattern. (Figure 5)

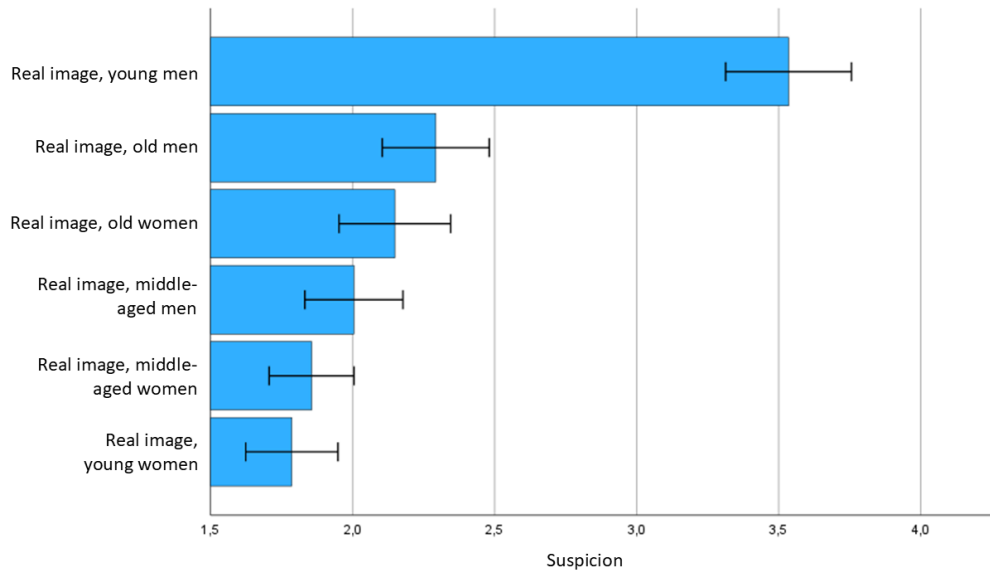


Figure 5. *The average level of suspicion for the six authentic photographs*

In addition to calculating the suspicion index for authentic photographs and AI-generated images independently, a comprehensive index was derived for each respondent to assess the accuracy of their perceptions. The value of this aggregate indicator was computed using the following formula:

$$A = \frac{S_{AI} + (7 - S_a)}{2}$$

In this context, A represents perceptual accuracy, S_{ai} denotes the average level of justified suspicion for AI-generated images, and S_a indicates the average level of unjustified suspicion for authentic photographs. Consequently, the indicator achieves its theoretical maximum value of 6 when a respondent selects "Almost certain that this image was generated by AI" for all AI-generated images and "Almost certain that this is an authentic photographs" for all authentic images. Conversely, the theoretical minimum value of 1 would be assigned to a respondent who consistently believes in the authenticity of the AI-generated images and the artificial nature of the authentic photographs, without exception.

This perceptual index exhibited a minimum value of 2.75 and a maximum value of 6.00, indicating that at least one respondent in our sample not only answered all 12 stimuli accurately but did so with complete confidence. The mean perceptual accuracy was recorded at 4.74, with a SD of 0.64. Further analysis revealed that the distribution of this indicator was mound-shaped and nearly symmetrical, although it exhibited a slight leftward skew due to the mean being positioned above the scale's midpoint. (Figure 6)

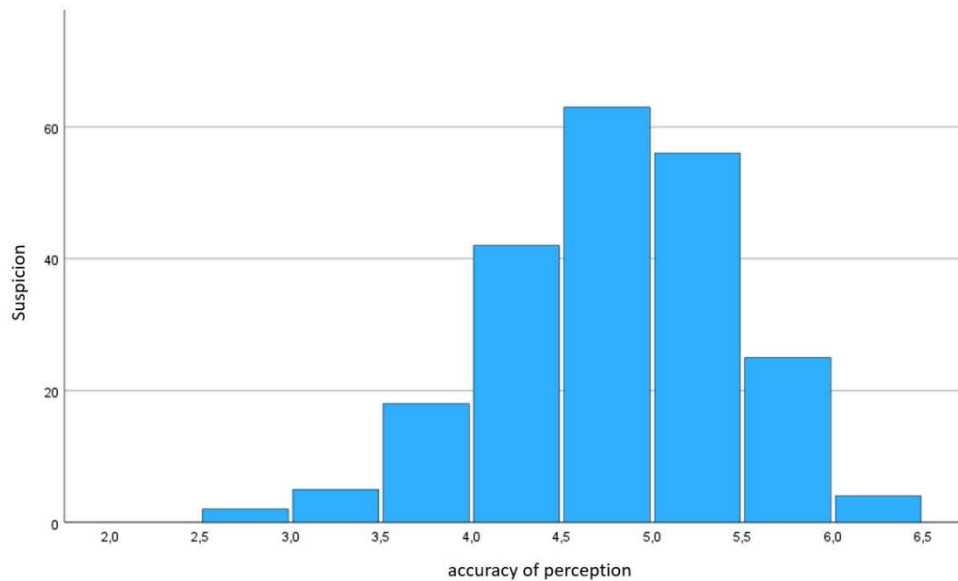


Figure 6. *Distribution of the aggregate indicator measuring the accuracy of perception*

This aggregate indicator also facilitated the investigation of whether significant differences in perceptual accuracy could be observed across various types of stimuli. The data indicate notable differences based on the gender and the age group of the individuals depicted.

The results from the applied t-test reveal that respondents a significantly higher degree of perceptual accuracy for images depicting women compared to those depicting men, with mean accuracy scores of 4.83 and 4.62, respectively. Although the magnitude of this difference is relatively small (Cohen's $d = 0.311$), it is statistically significant: $t(214) = 4.567$, $p < 0.001$. The disparity in perception between the two categories of stimuli is estimated using 95% confidence intervals. (Figure 7)

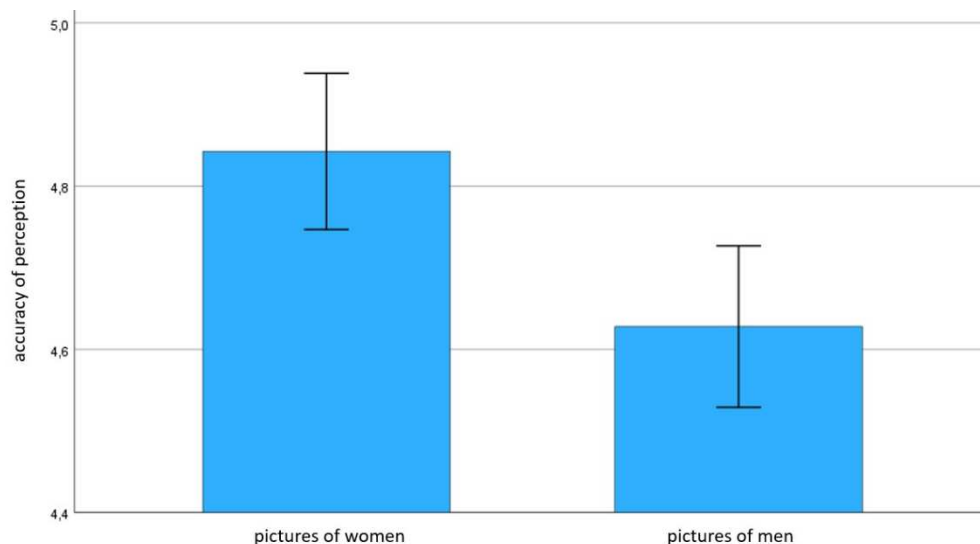


Figure 7. *Accuracy of perception according to the gender of the person in the picture*

Similarly, the age of the individuals depicted significantly influences perceptual accuracy. The results demonstrate that the critical assumption for repeated measures ANOVA, sphericity, is not violated (Mauchly's $W = 0.995$; $\chi^2(2) = 1.045$, n.s.). The repeated measures analysis of variance, conducted under the assumption of

sphericity, indicated significant differences in the correct recognition of images of individuals across different age groups: $F(2, 428) = 39.498, p < 0.001$. (Figure 8) These differences are attributed to the fact that participants were more accurate in discriminating between authentic photographs and AI-generated images of elderly individuals (mean accuracy: 5.02) compared to the stimuli depicting middle-aged individuals (mean: 4.65) and young individuals (mean: 4.54). In this instance, as in the previous analysis, 95% confidence intervals were also calculated.

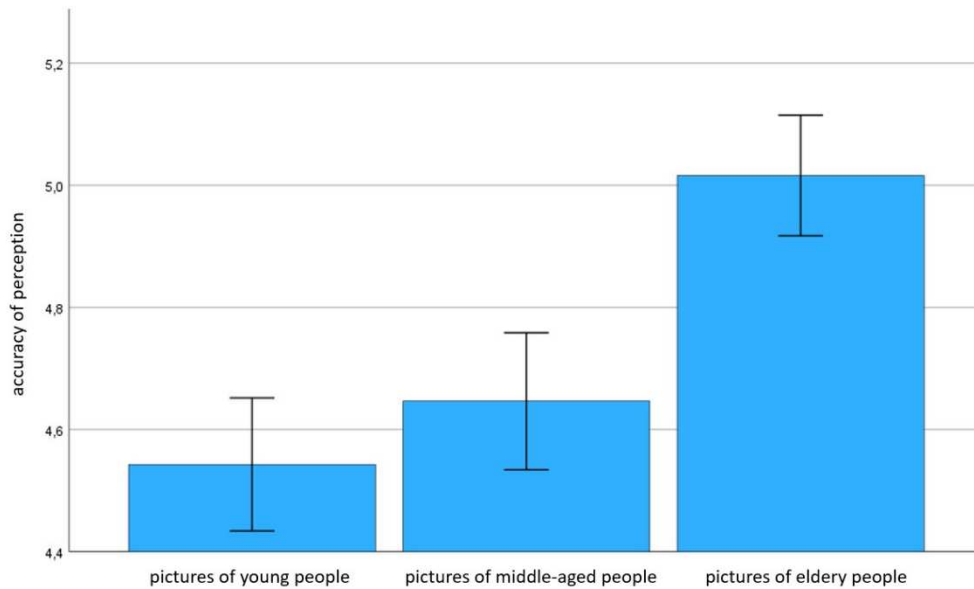


Figure 8. *Accuracy of perception according to the age of the person in the picture*

Based on an analysis of statistical findings and textual responses, suspicion regarding AI-generated images arises from specific characteristics that often appear "too perfect" or "too smooth," especially concerning skin, face, hair, and eyes. Issues like anatomical irregularities—such as disproportionate hands and fingers—and a lack of expressiveness in the gaze, along with excessively uniform facial features, are notable indicators of artificial generation. Regarding environmental context and composition, aspects such as unnatural backgrounds, defects in light and shadow, and questionable object positioning suggest an artificial origin. Distortions, blurriness, and nonsensical inscriptions on clothing and accessories also contribute to this suspicion. On the other hand, skepticism towards authentic images is often driven by overarching impressions perceived as "unnatural" or "unreal-looking," frequently influenced by environmental factors or specific details within those images. As AI technology advances, distinguishing between authentic photographs and AI generated images has become more complex. This underscores the importance of meticulous attention to detail, a solid understanding of anatomy, and a thorough analysis of image composition in recognizing generated images. Future technological advancements will require the development of new methodologies and tools for effective discrimination.

Discussion

In our study, we examined the ability of non-expert media users to identify images created by artificial intelligence. We compared lifelike portraits generated by the Stable Diffusion model on the Playground AI platform with authentic photographs in an online survey experiment. Our findings indicate that respondents could differentiate between the two types of images with notably high accuracy. The mean

level of suspicion for AI-generated images was significantly greater than for authentic photographs. This difference is substantial. It is essential to acknowledge that as technology advances, this disparity will likely diminish in the future. Utilizing our overall perceptual accuracy indicator, we determined that respondents were able to correctly identify the origin of the images in the majority of instances. Notably, the accuracy for images of women was slightly higher than those of men, while images of older individuals exhibited the highest level of perceptual accuracy. This finding aligns with the analysis of textual responses, which indicates that AI algorithms struggle with the realistic representation of human anatomy, particularly concerning hands and faces. Respondents frequently cited indicators such as unnatural smoothness of skin, absence of wrinkles, awkward positioning of fingers, and expressionless gazes as telltale signs of artificiality. Furthermore, the assessment of background and environmental context was also crucial to the perception of the images; characteristics such as blurriness, unnatural backgrounds, peculiar lighting and shadow distribution, and discordance between the subject and the surrounding environment were all perceived as potential indicators of an artificial origin.

The findings from our research underscore the necessity of fostering media literacy in the digital era. As images generated by artificial intelligence become increasingly realistic, distinguishing them from authentic photographs poses a growing challenge. Critical thinking, attention to detail, a foundational understanding of anatomy, and the ability to analyze image composition are all essential skills for recognizing manipulated images. While a trained observer may identify these alterations with relative ease, this raises a pertinent question for the future: should we expect the average media consumer to possess expertise in visual analysis? Further research is imperative to develop more effective detection methodologies and educate media users about the potential risks associated with deepfake technology. As algorithms continue to evolve, research must also advance to uphold public confidence in visual information's authenticity and mitigate the risks posed by manipulated images. Without fully reliable detection systems, our best approach may be to trust that enhanced individual awareness and performance will collectively contribute to overcoming these challenges.

The potential for further and more comprehensive research on this topic is significant. Users' responses and the myriad factors influencing their perception of visual information exhibit considerable complexity. Additional studies would be beneficial, where participants first complete a questionnaire and then undergo awareness training—such as exposure to specific examples and strategies for identifying manipulated or artificially generated images—before retaking the questionnaire. By administering the same or a modified version of the initial questionnaire post-training, researchers could acquire valuable comparative data regarding the extent to which increased skepticism or improved recognition rates may result from such awareness initiatives. This research could inform the development of more effective educational and awareness campaigns, ultimately equipping users with the skills needed to recognize and mitigate the impact of visual misinformation in the long term.

Acknowledgment

The authors would like to thank the Institute of Cybersecurity of the Ludovika University of Public Service for their valuable contribution and support, which

contributed significantly to the realization of the research. To improve the linguistic quality of the manuscript, we used the linguistic support of Grammarly, which helped us fine-tune the academic style.

References

- Abbas, F., & Taeihagh, A. (2024). Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. *Expert Systems with Applications*, 252, 124260. <https://doi.org/10.1016/j.eswa.2024.124260>
- Ahmed, S. (2023). Examining public perception and cognitive biases in the presumed influence of deepfakes threat: Empirical evidence of third person perception from three studies. *Asian Journal of Communication*, 33(3), 308–331. <https://doi.org/10.1080/01292986.2023.2194886>
- Bird, J. J., & Lotfi, A. (2024). CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images. *IEEE Access*, 12, 15642–15650. IEEE Access. <https://doi.org/10.1109/ACCESS.2024.3356122>
- Boehmer, J., LaRose, R., Rifon, N., Alhabash, S., & Cotten, S. (2015). Determinants of online safety behaviour: Towards an intervention strategy for college students. *Behaviour & Information Technology*, 34(10), 1022–1035. <https://doi.org/10.1080/0144929X.2015.1028448>
- Califano, G., & Spence, C. (2024). Assessing the visual appeal of real/AI-generated food images. *Food Quality and Preference*, 116, 105149. <https://doi.org/10.1016/j.foodqual.2024.105149>
- Chauhan, S. S., Jain, N., Pandey, S. C., & Chabaque, A. (2022). Deepfake Detection in Videos and Picture: Analysis of Deep Learning Models and Dataset. *2022 IEEE International Conference on Data Science and Information System (ICDSIS)*, 1–5. <https://doi.org/10.1109/ICDSIS55133.2022.9915885>
- Chen, Y., Ding, Z., Alowain, L., Chen, X., & Wagner, D. (2023). DiverseVul: A New Vulnerable Source Code Dataset for Deep Learning Based Vulnerability Detection. *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses*, 654–668. <https://doi.org/10.1145/3607199.3607242>
- Chinta, D. S., Kamineni, S., Chatragadda, R. P., & Kamepalli, S. (2024). Analyzing Image Classification on AI-Generated Art Vs Human Created Art Using Deep Learning Models. *2024 Third International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, 1–6. <https://doi.org/10.1109/ICEEICT61591.2024.10718485>
- Cinelli, M., Quattrocioni, W., Galeazzi, A., Valensise, C. M., Brugnoti, E., Schmidt, A. L., Zola, P., Zollo, F., & Scala, A. (2020). The COVID-19 social media infodemic. *Scientific Reports*, 10(1), 16598. <https://doi.org/10.1038/s41598-020-73510-5>
- COE. (n.d.). *Dealing with propaganda, misinformation and fake news—Democratic Schools for All—Www.coe.int*. Democratic Schools for All. Retrieved March 17, 2025, from <https://www.coe.int/en/web/campaign->

free-to-speak-safe-to-learn/dealing-with-propaganda-misinformation-and-fake-news

- Eszteri D. (2020). Elosztott mesterséges intelligencia fejlesztés blokklánc alapon az adatvédelem érvényesülése érdekében. *Pro Futuro*, 10(1), 9–27. <https://doi.org/10.26521/Profuturo/2020/1/7554>
- Gambín, Á. F., Yazidi, A., Vasilakos, A., Haugerud, H., & Djenouri, Y. (2024). Deepfakes: Current and future trends. *Artificial Intelligence Review*, 57(3), 64. <https://doi.org/10.1007/s10462-023-10679-x>
- Grassini, S., & Koivisto, M. (2024). Understanding how personality traits, experiences, and attitudes shape negative bias toward AI-generated artworks. *Scientific Reports*, 14(1), 4113. <https://doi.org/10.1038/s41598-024-54294-4>
- Grommelt, P., Weiss, L., Pfreundt, F.-J., & Keuper, J. (2024). *Fake or JPEG? Revealing Common Biases in Generated Image Detection Datasets*. <https://arxiv.org/abs/2403.17608>
- Grut, S. (2024). Source-Critical Affordances in Social Media Apps. *International Journal of Communication*, 18(0), Article 0.
- Gupta, A. S., Shreneter, K. P., & Sehgal, S. (2024). Visual Veracity: Advancing AI-Generated Image Detection with Convolutional Neural Networks. 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 1–6. <https://doi.org/10.1109/ICRITO61523.2024.10522113>
- Hancock, J. T., Naaman, M., & Levy, K. (2020). AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication*, 25(1), 89–100. <https://doi.org/10.1093/jcmc/zmz022>
- Huang, L., Zhang, Z., Zhang, Y., Zhou, X., & Wang, S. (2025). *RU-AI: A Large Multimodal Dataset for Machine-Generated Content Detection*. <https://doi.org/10.1145/3701716.3715306>
- Islam, M. T., Lee, I. H., Alzahrani, A. I., & Muhammad, K. (2025). MEXFIC: A meta ensemble eXplainable approach for AI-synthesized fake image classification. *Alexandria Engineering Journal*, 116, 351–363. <https://doi.org/10.1016/j.aej.2024.12.031>
- Kavtaradze, L. (2025). Dominant Disciplinary and Thematic Approaches to Automated Fact-Checking: A Scoping Review and Reflection. *Digital Journalism*, 0(0), 1–26. <https://doi.org/10.1080/21670811.2024.2427036>
- Konieczniak, W. (2023, November 22). Disinformation, misinformation, malinformation and fake news: Cracking the code of information disorders by Gracia Sumariva Reyes. *CYBERSEC – European Cybersecurity Forum*. <https://cybersecforum.eu/2023/11/22/disinformation-misinformation-malinformation-and-fake-news-cracking-the-code-of-information-disorders-by-gracia-sumariva-reyes/>
- Li, Y., Liu, Z., Zhao, J., Ren, L., Li, F., Luo, J., & Luo, B. (2024). The Adversarial AI-Art: Understanding, Generation, Detection, and Benchmarking. In J. Garcia-Alfaro, R. Kozik, M. Choraś, & S. Katsikas (Eds.), *Computer*

Security – ESORICS 2024 (pp. 311–331). Springer Nature Switzerland.
https://doi.org/10.1007/978-3-031-70879-4_16

- Maples-Keller, J. L., Williamson, R. L., Sleep, C. E., Carter, N. T., Campbell, W. K., & Miller, J. D. (2019). Using Item Response Theory to develop a 60-item representation of the NEO PI–R using the International Personality Item Pool: Development of the IPIP–NEO–60. *Journal of Personality Assessment*, 101(1), 4–15.
<https://doi.org/10.1080/00223891.2017.1381968>
- McLarin, R. D. (1982). *Military propaganda: Psychological warfare and operations* (First Edition). Praeger, 1982.
- Mirsky, Y., & Lee, W. (2021). The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surv.*, 54(1), 7:1-7:41.
<https://doi.org/10.1145/3425780>
- Nayim, M., Mohan, V., Pandey, T. N., Dash, B. B., Dash, B. B., & Patra, S. S. (2024). Detection of Leading CNN Models for AI Image Accuracy and Efficiency. *2024 International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS)*, 1–7.
<https://doi.org/10.1109/IACIS61494.2024.10721936>
- Nurse, J. R. C. (2019). Cybercrime and you: How cybercriminals attack and the human factors that they seek to exploit. In A. Attrill-Smith Fullwood, C. ., Keep, M. ., & Kuss, D. (Ed.), *The Oxford handbook of cyberpsychology*. Oxford University Press.
- Park, H., Kim, G., Lee, D., & Kim, H. K. (2024). Can You Spot the AI-Generated Images? Distinguishing Fake Images Using Signal Detection Theory. In P.-L. P. Rau (Ed.), *Cross-Cultural Design* (pp. 299–313). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-60913-8_21
- Pennycook, G., & Rand, D. G. (2021). The Psychology of Fake News. *Trends in Cognitive Sciences*, 25(5), 388–402.
<https://doi.org/10.1016/j.tics.2021.02.007>
- Singh, J. N., Gautam, A., & Tomar, H. (2023). Deep Fake in picture using Convolutional Neural Network. *2023 5th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, 1104–1107.
<https://doi.org/10.1109/ICAC3N60023.2023.10541758>
- Tandoc Jr., E. C., Lim, Zheng Wei, & Ling, R. (2018). Defining “Fake News”: A typology of scholarly definitions. *Digital Journalism*, 6(2), 137–153.
<https://doi.org/10.1080/21670811.2017.1360143>
- Tihanyi, N., Bisztray, T., Dubniczky, R. A., Toth, R., Borsos, B., Cherif, B., Jain, R., Muzsai, L., Ferrag, M. A., Marinelli, R., Cordeiro, L. C., Debbah, M., Mavroeidis, V., & Jøsang, A. (2024). Dynamic Intelligence Assessment: Benchmarking LLMs on the Road to AGI with a Focus on Model Confidence. *2024 IEEE International Conference on Big Data (BigData)*, 3313–3321. <https://doi.org/10.1109/BigData62323.2024.10825051>
- Tihanyi, N., Bisztray, T., Jain, R., Ferrag, M. A., Cordeiro, L. C., & Mavroeidis, V. (2023). The FormAI Dataset: Generative AI in Software Security through

- the Lens of Formal Verification. *Proceedings of the 19th International Conference on Predictive Models and Data Analytics in Software Engineering*, 33–43. <https://doi.org/10.1145/3617555.3617874>
- Tóth, R., Bisztray, T., & Erdódi, L. (2024). LLMs in Web Development: Evaluating LLM-Generated PHP Code Unveiling Vulnerabilities and Limitations. In A. Ceccarelli, M. Trapp, A. Bondavalli, E. Schoitsch, B. Gallina, & F. Bitsch (Eds.), *Computer Safety, Reliability, and Security. SAFECOMP 2024 Workshops* (pp. 425–437). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-68738-9_34
- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146. <https://doi.org/10.1126/science.aap9559>
- Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making* (No. DGI(2017)09). Council of Europe. <https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html>
- Weatherbed, J. (2025, February 6). *Google is adding AI watermarks to photos manipulated by Magic Editor*. The Verge. <https://www.theverge.com/news/607515/google-photosynthid-ai-watermarks-magic-editor>
- Wilser, J. (2024, October 30). *Content Credentials: The 200 Best Inventions of 2024*. TIME. <https://time.com/7094554/content-credentials/>
- Zhu, M., Chen, H., Yan, Q., Huang, X., Lin, G., Li, W., Tu, Z., Hu, H., Hu, J., & Wang, Y. (2023). *GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image*. <https://arxiv.org/abs/2306.08571>