ESSACHESS – Journal for Communication Studies

The Capability to Identify AI-Generated Images

ESSACHESS –
Journal for Communication Studies
Volume 18 Issue 2(36), p. 165-191
© The Author(s) 2025
Reprints and Permission:
© ESSACHESS
https://www.essachess.com/
DOI: 10.21409/essachess.1775-352x

Cite : BÁNYÁSZ, P., ORBÓK, Á., MAGYAR, S., TÓTH, R., BÁNYÁSZ-VÁCZI, K. B., VESZELSZKI, Á., & KOVÁCS, G. (2025). The Capability to Identify AI-Generated Images. *ESSACHESS*. 18(36): 165-191. https://doi.org/10.21409/WZJV-P693

Péter BÁNYÁSZ
Associate Professor, PhD, Ludovika University of Public Service, HUNGARY
e-mail: banyasz.peter@uni-nke.hu

Sándor MAGYAR
Associate Professor, PhD, Ludovika University of Public Service, HUNGARY
e-mail: magyar.sandor@uni-nke.hu

Ákos ORBÓK
Assistant Professor, PhD, Ludovika University of Public Service, HUNGARY
e-mail: orbok.akos@uni-nke.hu

Rebeka TÓTH
PhD student, University of Oslo, NORWAY
e-mail: rebekat@ifi.uio.no

Article received on March 5, 2025. Article accepted on November 15, 2025.
Conflict of Interest: The author(s) declare(s) no conflict of interest.

Kincső Boróka BÁNYÁSZ-VÁCZI
Researcher, Ludovika University of Public Service, HUNGARY
e-mail: vaczi.kincso.boroka@uni-nke.hu

Ágnes VESZELSZKI
Associate Professor, PhD, Ludovika University of Public Service, HUNGARY
e-mail: veszelszki.agnes@uni-nke.hu

Gábor KOVÁCS
Associate Professor, PhD, Ludovika University of Public Service, HUNGARY
e-mail: kovacs.gabor.istvan@uni-nke.hu

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 (AI), survey, media literacy, cybersecurity education, critical thinking

***

La capacité d'identifier les images générées par l'IA

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.

1. Literature Review and Background

1.1. *The Threat of Misinformation on Society*

Misinformation, disinformation, and malinformation represent distinct yet interconnected dimensions of information disorder that collectively undermine public trust and social stability. Misinformation refers to the unintentional spread of inaccurate or misleading information, typically resulting from insufficient verification of sources—for example, exaggerated reports shared in good faith during a crisis.

Disinformation, in contrast, involves the deliberate creation and dissemination of false information for political, economic, or ideological purposes. Malinformation concerns the intentional release of authentic information in harmful contexts, such as leaking confidential data or distorting personal details to damage reputations. (Konieczniak, 2023)

While these categories differ in intent and impact, each poses serious societal risks and demands context-specific countermeasures. Misinformation often arises from cognitive biases and the failure to verify information accuracy, whereas disinformation frequently serves strategic agendas, ranging from state-sponsored propaganda to financially motivated clickbait. Malinformation, despite being grounded in truth, raises ethical and legal dilemmas when it violates privacy or national security boundaries. Experts, including those from the Council of Europe, therefore advise against the catch-all term fake news, which oversimplifies these complex dynamics and is often exploited politically to delegitimize credible journalism.

Information disorder overlaps conceptually with related constructs such as propaganda and psychological operations, both of which weaponize communication to shape perception and behavior. (Tandoc Jr. et al., 2018) Propaganda refers to manipulative, one-sided communication that distorts reality—often through half-truths or disinformation—to influence public opinion for political or ideological purposes. This selective distortion of reality links the propaganda of 20th-century totalitarian regimes with today's extremist online campaigns. (McLarin, 1982) Modern hybrid warfare strategies, exemplified by Russian information campaigns, combine disinformation, cyber operations, and propaganda to erode democratic cohesion and manipulate public opinion. In this sense, the contemporary online ecosystem represents a continuation of historical propaganda tactics, now amplified by the speed and reach of social media. The implications of these phenomena extend beyond political manipulation to the broader erosion of trust in institutions, experts, and even visual evidence itself. As communication increasingly shifts to digital platforms, the boundaries between credible and deceptive information blur further. This evolving information environment establishes the foundation for new forms of deception—particularly those enabled by artificial intelligence—which are explored in the following section.

1.2. *Generative AI's Potential in Creating Deceptive Content*

While this paper primarily examines the implications of AI-generated images in the field of disinformation, it is crucial to acknowledge that these images present various additional cybersecurity threats. They can be exploited in social-engineering and phishing attacks through deepfake technologies that impersonate real individuals or fabricate false narratives. Such images may also bypass biometric systems, enable counterfeit profiles, or conceal malicious code within steganographic files. These risks highlight the need for stronger detection tools and greater cybersecurity awareness to counter visual manipulation.

Recent advances in GenAI have increased vulnerability to online threats, especially on social-media platforms. A key concern is GenAI's ability to craft highly convincing phishing messages or fake digital identities that deceive users and gain access to sensitive data. (Hancock et al., 2020)

Although GenAI supports software development through code completion and generation, studies reveal significant security issues. Tóth et al. (2024) found high vulnerability rates in LLM-generated PHP web applications, while Tihanyi et al. (2023) reported that over half of AI-generated C programs in the FormAI dataset contained security flaws. Similarly, Chen et al. (2023) showed that deep-learning models still struggle to detect complex software vulnerabilities despite using extensive datasets. Collectively, these findings underscore the risks of deploying automatically generated code without thorough validation.

GenAI also enables large-scale disinformation. Automated systems can create and spread fake news, propaganda, and deepfakes that manipulate public opinion, particularly during elections or crises. The ability to generate highly realistic synthetic images, audio, and video facilitates abuses such as defamation, blackmail, or identity fraud. (Mirsky & Lee, 2021)

Performance evaluations of large language models (LLMs) further reveal inconsistent reliability. The Dynamic Intelligence Assessment (DIA) framework found that tool-using models, such as ChatGPT-4o, outperform others but still exhibit reasoning errors and hallucinations. Specifically, ChatGPT-4o achieved a Reliability Score of -64 and a Confidence Index of 38.7%, and lower values for other models. (Tihanyi et al., 2024) These results highlight persistent accuracy challenges that complicate responsible AI deployment.

Finally, ensuring the quality and integrity of training data remains a major challenge. If generative models rely on manipulated or synthetic inputs, they may reproduce errors or biases. Blockchain technology offers a potential safeguard by enabling decentralized, tamper-resistant data management, enhancing transparency and trust in AI systems. (Eszteri, 2020)

1.3. *Detection and Human Perception of AI-Generated Content*

Distinguishing between AI-generated and authentic photographs presents a multifaceted challenge shaped by technological, psychological, and ethical factors. In this manuscript, the term "authentic photographs" is utilized interchangeably with "human-captured photographs" to refer to non-AI images obtained from professional stock photography databases, which have been verified to have been created prior to the advent of generative AI image-generation tools. Accurate identification is essential for mitigating misinformation and maintaining digital trust. (Gupta et al., 2024) Research using diverse datasets and convolutional neural network architectures has achieved high detection accuracy up to 97.29% in identifying synthetic images. (Gupta et al., 2024; Chinta et al., 2024; Bird & Lotfi, 2024; Nayim et al., 2024) However, perceptual factors also play a role: the emotional tone of images,

particularly positive emotions, can influence users' ability to discern AI-generated visuals. (Park et al., 2024)

Recent studies have developed large-scale datasets to enhance AI-image detection. The GenImage dataset includes over one million paired real and synthetic images from diffusion and GAN models, offering a strong benchmark for evaluation. (Zhu et al., 2023) The RU-AI dataset extends this approach across text, image, and voice modalities to improve cross-media detection. (Huang et al., 2025) Yet challenges remain due to dataset biases, such as JPEG compression and image size, which reduce detector robustness. (Grommelt et al., 2024) Meanwhile, new watermarking technologies aim to ensure authenticity. Google's SynthID, developed by DeepMind, invisibly embeds metadata into AI-generated images, allowing identification without altering appearance. (Weatherbed, 2025) Similarly, the Content Authenticity Initiative (CAI) and its Content Credentials system integrate provenance data directly into files, enabling cameras and editing tools like Adobe Photoshop to verify modifications. (Wilser, 2024) These innovations represent major progress toward maintaining the integrity of digital content and reducing the impact of visual misinformation.

Psychological and cognitive factors also play a crucial role in recognizing AI-generated images. Emotional responses, creative identity, and openness influence how individuals perceive and evaluate synthetic artworks. (Park et al., 2024; Grassini & Koivisto, 2024) Participants in one study struggled to distinguish between authentic photographs and AI-generated images and often showed negative bias toward the latter when judging authenticity. (Grassini & Koivisto, 2024) From an ethical perspective, the rapid spread of AI-generated visuals heightens risks of disinformation and fraud, underscoring the need for reliable detection methods. (Li et al., 2024) These perceptual issues are particularly relevant in dynamic fields such as digital marketing, where image authenticity directly shapes consumer trust. (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 have become increasingly sophisticated and accessible, allowing even non-experts to produce highly realistic fabricated content. While generative AI offers creative and educational benefits, it also enables deception, defamation, and identity manipulation. Such misuse can damage reputations, influence politics, or facilitate fraud, eroding public trust in digital media. (Gambín et al., 2024; Abbas & Taeihagh,

2024) The convergence of deep learning, big-data resources, and powerful editing tools has lowered barriers to producing convincing fakes, expanding their malicious potential. As these technologies advance, fabricated images and videos are likely to become ever harder for the average viewer to distinguish from reality.

Social media platforms play a central role in spreading deepfake content. A vast number of users can easily access and share manipulated images, while algorithm-driven feeds amplify emotionally charged material that captures attention. This system accelerates the reach of deceptive visuals and increases the likelihood of users encountering misinformation without verifying authenticity. As a result, deepfake content can mislead large online audiences, fostering the rapid and often uncontrolled spread of fake news and disinformation. Beyond the swift sharing capabilities, these technologies are also exploited for malicious purposes—such as harvesting data, discrediting public figures, or fueling social polarization—producing serious and lasting repercussions. In this environment, distinguishing between authentic and fabricated content has become increasingly difficult.

Although the realism of deepfakes continues to improve, detection methods remain limited. (Singh et al., 2023) Identifying and filtering manipulated media quickly and accurately is still a major challenge, undermining trust in digital information. (Chauhan et al., 2022) The field of deepfake creation and detection evolves rapidly, with both developers and adversaries deploying more advanced techniques to outpace one another. This ongoing technological race continually generates more convincing fakes while delaying the development of widely accessible and reliable detection systems.

The proliferation of disinformation in the digital field—whether manipulating public opinion, exerting political influence, provoking panic, facilitating cybercrime, or merely seeking attention—often occurs subtly. Social media enables rapid, large-scale dissemination, as users share or repost content with minimal scrutiny. These deceptive narratives can strongly shape perceptions, decisions, and behaviors. Strategically timed or carefully crafted fake news and images can distort public discourse, erode trust, and produce lasting political and economic consequences.

Images exert exceptional emotional power, often influencing how information is perceived and remembered. A single visual element can reshape the tone of a narrative or intensify emotional reactions. Therefore, identifying manipulated visual content and understanding the motives behind it are critical. Without careful verification, individuals may unintentionally spread misleading material, allowing distortions to embed in public consciousness. As information volume grows, images must be analyzed with the same rigor as text, supported by advanced detection technologies to curb the spread of visual misinformation on social media and other online platforms.

People often overestimate their ability to recognize deepfakes, assuming they are less susceptible to deception than others. This overconfidence is dangerous, as advances in artificial intelligence have made manipulations increasingly realistic,

deceiving even the most discerning observers. Deepfakes can significantly influence public opinion, particularly in political contexts, where fabricated videos may sway elections or erode trust in leaders. (Ahmed, 2023) Because misinformation evolves rapidly, detection features effective today may become obsolete tomorrow, requiring constant adaptation by practitioners. (Pennycook & Rand, 2021) A further challenge lies in the lack of transparency within digital platforms—many applications do not request or display source information when images are uploaded, complicating efforts to verify authenticity. (Grut, 2024)

Developing advanced detection and filtering tools is essential for countering deepfakes. However, technological defenses alone are insufficient without user awareness and critical evaluation skills. Educating users about deepfake risks and providing guidance on recognizing manipulation are key components of digital resilience. Public awareness initiatives and media literacy campaigns can substantially reduce the likelihood of individuals being misled by fabricated visual content.

Automated fact-checking of images requires a comprehensive, interdisciplinary approach that matches the sophistication of modern manipulation techniques. (Kavtaradze, 2025) Technologically, it depends on advances in machine learning, computer vision, and artificial intelligence to detect altered image components with precision and speed. Equally important is the contribution of psychologists, communication specialists, and ethicists, who study media behavior, deception strategies, and the societal effects of manipulated content. Integrating these disciplines can transform automated image verification from a purely technical process into a practical tool that strengthens public understanding and resilience against digital deception.

1.3. *Research Gap*

Despite substantial progress in the technological detection of AI-generated and manipulated imagery, little is known about how individuals—particularly non-specialists—perceive and judge the authenticity of such content in real social-media contexts. Most existing research concentrates on algorithmic accuracy rather than on human discernment or behavioral responses to synthetic visuals. Furthermore, the effects of demographic and visual factors, such as the gender and age of depicted individuals, on perceptual accuracy remain underexplored.

This study addresses these gaps by examining users' ability to recognize AI-generated images, the visual cues they rely on, and the factors that make some viewers more susceptible to deception. The findings aim to inform both the design of more effective detection methodologies and educational initiatives that strengthen digital literacy and resilience against visual misinformation.

1.4. *Research Objectives*

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 authentic photographs and AI-generated images, 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.

2. 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. (Boehmer et al., 2015; Maples-Keller et al. 2019; Nurse, 2019) 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 influence the recognition of AI-generated images and the ability to distinguish them from authentic photographs. The authentic photographs were curated from a complimentary stock photo database featuring portraits of individuals aged 20, 40, and 60. In the process of selecting images from the stock photo database, we exercised careful scrutiny of the technical information and metadata provided by the platform. Only those photographs clearly indicated by the metadata as having been captured prior to 2022 were incorporated into the study. This approach aimed to mitigate the risk of including stock photos potentially generated by artificial intelligence, thereby safeguarding the authenticity and validity of the control group consisting of authentic photographs. 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. The Playground AI platform, which is publicly accessible and incorporates several sophisticated diffusion models, including Stable Diffusion XL (SDXL), within a browser-based framework. This platform was selected primarily due to its capabilities for research reproducibility, transparency in experimental parameters, and visual realism. Playground AI offers the distinct advantage of merging the high aesthetic quality characteristic of Midjourney with the open and research-friendly architecture of Stable Diffusion. In contrast to Midjourney, which operates within a closed, Discord-based environment and does not publicly disclose its generation parameters or specific model versions, Playground AI facilitates the controlled input of SDXL model settings and provides transparent documentation throughout the image generation process. Consequently, this platform reflects the prevailing trends in photorealistic portrait generation as of spring 2024, particularly with respect to the research applicability of open diffusion-based methodologies. For the image generation, we employed the following parameters: model – Stable Diffusion; filter – Realism Engine; preset – Realistic Portraits (SDXL); image dimensions – 720×1024 pixels; prompt guidance – 7; quality & details – 30; refinement – 0; sampler – Euler a. These configurations were specifically designed to produce photorealistic portraits while minimizing the presence of artificial textures and visual distortions. The prompt guidance value of 7 struck a balance between accurate adherence to textual instructions and a natural visual presentation, while a quality & details setting of 30 yielded sharp, detailed images devoid of distortions. By setting refinement to 0, we ensured that the outputs were direct representations of the model's capabilities without any additional post-processing. The Euler sampling algorithm provided a rapid and stable solution to the diffusion process, effectively balancing visual coherence and detail. The use of Playground AI facilitated controlled, well-documented, and reproducible image generation, offering a more reliable foundation for research compared to the closed-system, parameter-

obscured approach of Midjourney, while still achieving comparably high levels of visual quality.

To ensure that the AI-generated images closely aligned with the visual and technical attributes of the authentic photographs, prompts were meticulously crafted to reflect the stylistic and camera-related specifications of the original images. The following example illustrates the type of prompts used in the image generation process: "Create a photorealistic image of a Caucasian man in his 60s. The man is wearing a short grey beard, his hair is cropped short, business-like. The man is sitting in an office environment, wearing a classic navy blue suit with matching tie. The man is sitting in front of a laptop computer, giving the impression of a businessman, apparently in a senior position. The man is looking at the camera with a friendly, slightly touching look. In the background is the large window of the office, through which sunlight filters in, with the city centre in the background. The background is slightly blurred. Camera specs: 4000x6000, aspect ratio: 2:3, camera: Nikon d5600, focal: 140.0 mm, aperture: F/5.6, ISO 200, shutter speed: 1/800. The image is intended to be as photorealistic as possible"

A set of high-fidelity images was selected for inclusion in the questionnaire, and their presentation order was determined through a controlled randomization procedure to ensure balanced exposure across stimulus attributes. To mitigate habituation and serial position effects that may systematically bias perceptual judgements—particularly in cases where multiple AI-generated images are presented consecutively—the final stimulus sequence was generated using an iterative permutation algorithm. The algorithm produced random sequences until one satisfied all predefined balancing constraints: no more than three consecutive images depicting individuals of the same gender, no more than two depicting the same age group, and no more than three belonging to the same stimulus type (i.e., authentic photographs vs. AI-generated images). This procedure yielded a rigorously counterbalanced stimulus order that minimized pattern learning and enhanced the internal validity of the perceptual assessment.

The final presentation order was established in accordance with the aforementioned criteria, as depicted in Table 1.

Table 1. *Attributes of stimuli and sequence of presentation*

Position	ID	Gender	Age	Type
1.	1	man	young	authentic
2.	12	woman	elderly	AI-generated
3.	5	woman	middle-aged	authentic
4.	2	man	middle-aged	authentic
5.	10	woman	young	AI-generated
6.	9	man	elderly	AI-generated

7.	11	woman	middle-aged	AI-generated
8.	4	woman	young	authentic
9.	3	man	elderly	authentic
10.	7	man	young	AI-generated
11.	8	man	middle-aged	AI-generated
12.	6	woman	elderly	authentic

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 revealed the participants' skepticism about the genuineness of both 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.

3. 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 photographs was significantly lower at 2.27 (SD: 0.85). Results from a matched-sample t-test revealed this difference to be statistically significant: $t(214) = 28.433$; $p < 0.001$.



Figure 1. *AI-generated images and authentic photographs*

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)

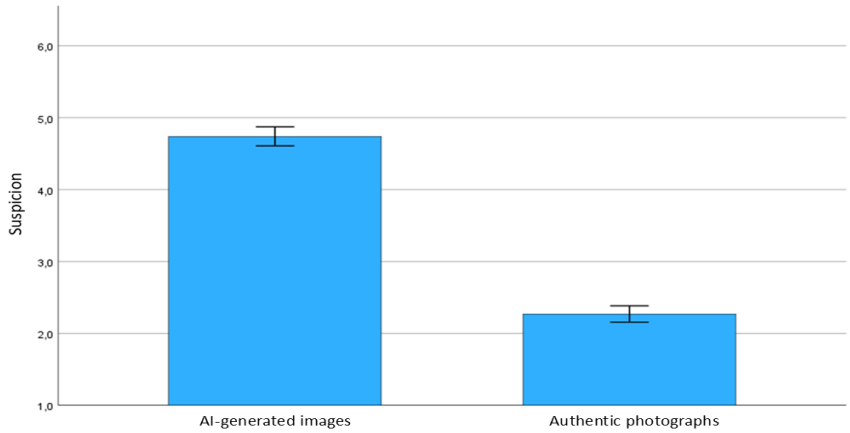


Figure 2. *The difference between the perception of AI-generated images and authentic photographs*

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

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 3)

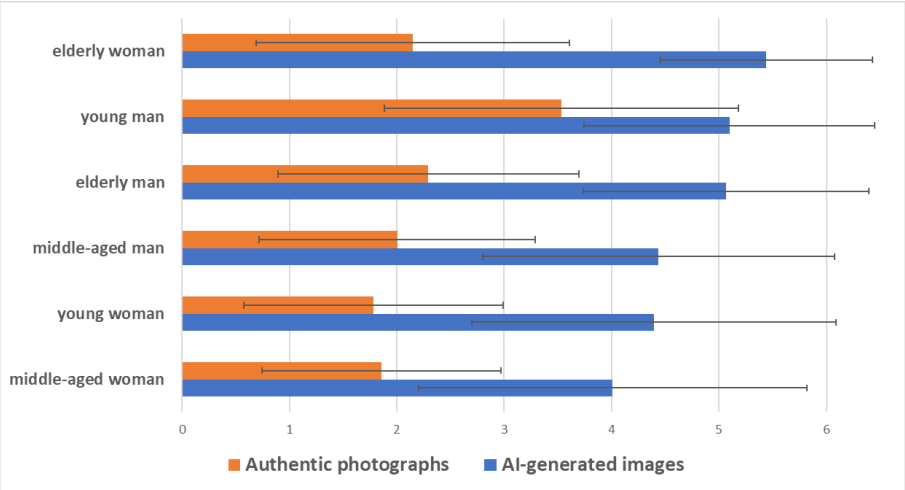


Figure 3. *The average level of suspicion for the six AI-generated images and the six authentic photographs*

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 an 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 3)

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 photograph" for all authentic photographs. 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 4)

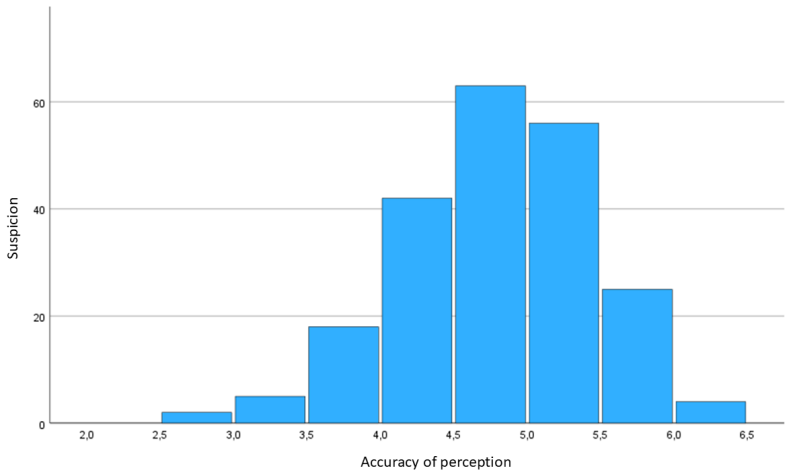


Figure 4. *Distribution of the aggregate indicator measuring the accuracy of perception*

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 pictures 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 5)

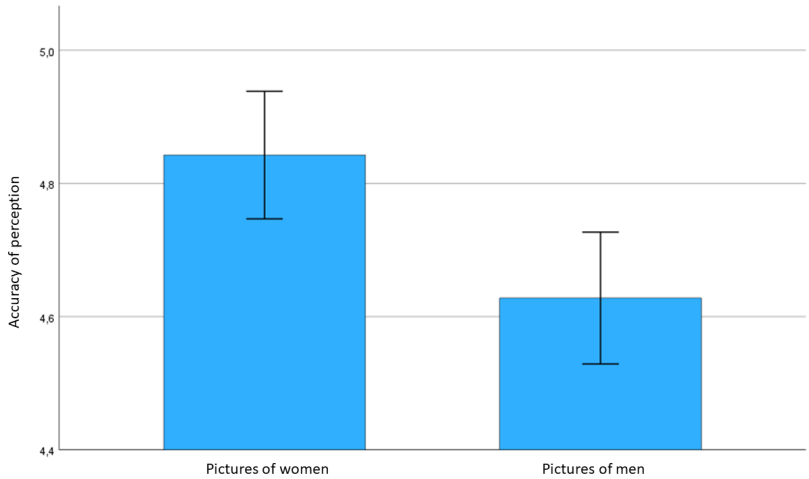


Figure 5. *Accuracy of perception according to the gender of the person in the picture*

It is worth noting that digital and visual media content typically features younger female faces in greater proportions, so users encounter these images more frequently and thus become more familiar with them. (Kramer et al., 2018) This frequent visual exposure may contribute to the fact that subtle differences such as anatomical flaws are more easily detected in young female faces than in less common face types. (Zhou et al., 2021) Thus, the higher recognition accuracy for female faces can be partly explained by the fact that there are younger female faces in the media, which increases perceptual sensitivity to these faces. (Fardouly et al., 2015)

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 pictures of individuals across different age groups: $F(2, 428) = 39.498$, $p < 0.001$. (Figure 6) 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.

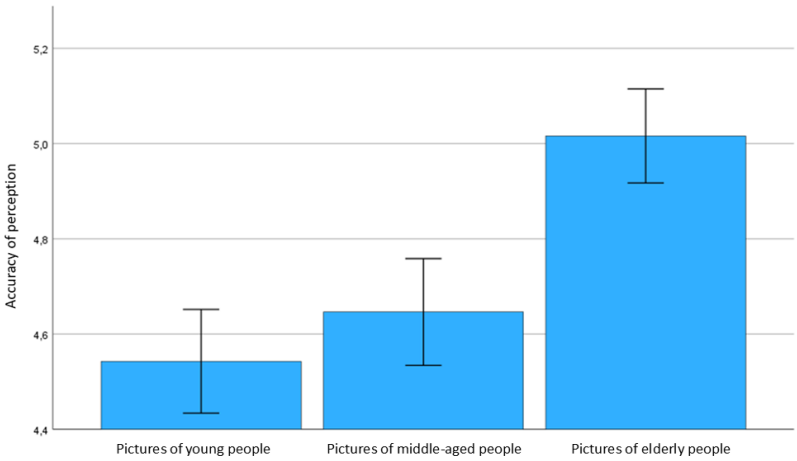


Figure 6. *Accuracy of perception according to the age of the person in the picture*

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 photographs 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.

Conclusion

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. Regarding RO1, 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. However, the accuracy of perception was influenced by a number of factors. Notably, with reference to RO2, 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. In relation to RO3, 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 AI-generated images 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 AI-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. Alternatively, the potential effect of pre-existing levels of AI literacy and the long-term effect of AI literacy development on the accuracy of perception could be further explored to obtain a complete understanding of the interplay between short and long-term factors.

The research findings emphasize the necessity of integrating specific components of AI literacy into current media and digital education curricula, thereby enhancing traditional media literacy frameworks that have primarily focused on source evaluation, textual analysis, and the detection of misinformation. As generative AI increasingly generates hyper-realistic visual content that challenges perceptual certainty, educational initiatives must actively foster AI visual literacy—a skill set that includes the critical interpretation, identification, and contextual evaluation of synthetic imagery. These curricular improvements should progress beyond superficial recognition activities to facilitate systematic skill development. This involves the incorporation of comparative visual analysis exercises that equip students to identify consistent artifacts and coherence discrepancies commonly associated with AI-generated visuals (e.g., disproportionate anatomical features, textural inconsistencies, variable light–shadow dynamics, or misalignments between context and background). Additionally, it is crucial that these activities are paired with instruction addressing the cognitive mechanisms that influence visual judgment, such as susceptibility to realism heuristics and social-categorical biases that may contribute to misclassification. Given the growing prevalence of generative AI in cyber-enabled deception, fraudulent activities, and influence campaigns, AI visual literacy has emerged as an essential cybersecurity competency. Enhancing individuals' capacity to critically assess the authenticity of visual content is vital for mitigating risks linked to AI-assisted social engineering attacks (for instance, visually manipulated executive impersonation in business email compromises) and bolstering societal resilience against visual disinformation aimed at swaying political decision-making, electoral choices, or public confidence in institutions. By incorporating these competencies within educational frameworks, we can promote a more security-conscious civic culture capable of recognizing and countering malicious synthetic media. Consistent with contemporary scholarship in metacognition and critical digital literacy, reflective practice should serve as a foundational aspect of AI literacy training. Learners ought to be encouraged to articulate and evaluate the reasoning processes underlying their assessments of authenticity, adjust their confidence levels, and identify the perceptual thresholds at which visual evaluations may be deemed unreliable. By nurturing analytical, metacognitive, and epistemic vigilance skills, such educational approaches

can enhance both individual and collective resilience against misleading synthetic visuals, thereby preparing participants for an information ecosystem in which the differentiation between authentic photographs and AI-generated images is progressively ambiguous.

In addition to its significance in the realms of education and cybersecurity, the findings possess considerable implications for the fact-checking and open-source intelligence (OSINT) communities, where the verification of visual content has increasingly become a vital operational capability. As AI-generated images become more prevalent within information ecosystems, it is essential for both professional fact-checkers and OSINT analysts to implement advanced verification workflows that surpass mere human perceptual judgment. The perceptual blind spots identified in this investigation could guide targeted training for verification specialists by highlighting specific image categories that demand heightened scrutiny—especially hyper-realistic portraits situated in business or politically sensitive contexts, which are frequently leveraged in influence operations and cyber-enabled deception. Crucially, the results underscore the necessity of normalizing multimodal OSINT-based verification practices that amalgamate human review with technical and contextual intelligence-gathering methodologies. In practical terms, verification protocols ought to encompass methods such as metadata extraction and analysis (e.g., utilizing ExifTool or FotoForensics), reverse image search and image-tracing tools (e.g., Google Lens, TinEye, InVID), geolocation and chronolocation verification, cross-platform source triangulation, and forensic assessment capable of identifying diffusion-model artifacts. While these OSINT techniques are prevalent in journalistic investigations and intelligence work, they are often inaccessible to the general public due to their perceived technical complexity and the higher digital literacy levels they necessitate. To bridge this gap, user-centered pedagogical strategies and tool development are crucial for making OSINT-informed verification more intuitive for non-experts. This includes translating expert workflows into simplified, step-by-step guidance, integrating built-in prompts or "credibility nudges" within platforms, and creating user-friendly interfaces that automate essential OSINT functions—such as one-click metadata checks, provenance tracing, and visual authenticity assessments with confidence scores. Furthermore, incorporating micro-learning modules and scenario-based exercises into digital literacy programs can further elucidate OSINT practices, enabling lay users to routinely and confidently apply basic verification techniques when encountering dubious visual content. By reinforcing synergies among fact-checking, OSINT methodologies, and public-facing digital literacy initiatives, societies can transition from reactive debunking to proactive visual resilience-building. This integrated approach not only enhances professional verification capacities but also empowers citizens to critically assess visual information, thereby diminishing vulnerability to AI-assisted deception, social engineering campaigns, and politically motivated disinformation efforts.

Collectively, these findings establish a significant foundation for future investigations into the perceptual, cognitive, and socio-technical factors that influence the assessment of visual authenticity. Subsequent research should build upon the current study by examining how individual variations in sensory processing, cognitive styles, and user backgrounds impact the detection of AI-generated images. A particularly important area of inquiry involves populations with sensory impairments—such as individuals who are deaf or hard of hearing—whose perceptual environments and attentional strategies diverge from those of hearing individuals. Existing literature indicates that deaf individuals frequently exhibit enhanced visual attention, more effective peripheral scanning, and distinctive gaze allocation patterns, which may affect both detection accuracy and confidence in evaluating visual authenticity. Gaining insights into these perceptual processes would aid in the formulation of inclusive AI literacy programs and specialized resilience-building interventions. From a methodological standpoint, future studies should integrate eye-tracking technologies to investigate how participants visually scan and cognitively process both authentic photographs and AI-generated images. Eye-movement data—such as fixation duration, saccade patterns, heatmaps, and scan paths—could elucidate the visual features that users depend on when making authenticity judgments, how attention allocation varies among demographic or sensory groups, and whether perceptual cues evolve as AI-generated images advances in sophistication. Merging eye-tracking with self-reported confidence measures and behavioral performance assessments would yield a more comprehensive cognitive-perceptual framework for evaluating synthetic images. Moreover, expanding the research scope to encompass diverse cultures, age demographics, and media formats would substantially enhance the generalizability of findings. Cross-cultural studies may reveal how cultural schemas, media conventions, and exposure to manipulated visuals influence visual trust and susceptibility to synthetic media. Longitudinal approaches could trace the evolution of detection capabilities in tandem with advancements in generative models and increasing public awareness. Finally, interdisciplinary collaboration with computer scientists, open-source intelligence professionals, and experts in human–computer interaction could facilitate the creation of adaptive training tools and personalized AI detection support systems tailored to individual perceptual patterns. Pursuing these avenues would enrich the understanding of visual resilience and foster the development of evidence-based strategies to mitigate the societal risks associated with progressively indistinguishable synthetic media.

Notwithstanding the contributions of this study, several limitations warrant consideration. First, the sample was confined to participants with internet access and predominantly higher educational backgrounds, which may constrain the generalizability of the findings to broader or less digitally literate populations. Second, while efforts were made to mitigate order and habituation effects during stimulus presentation, the reliance on a fixed set of static images introduces an inherent limitation, as real-world encounters with synthetic media typically occur within dynamic and contextually rich environments. Third, the images utilized in this research were generated in early 2024 using Stable Diffusion-based technology, and

the rapid evolution of generative models may outstrip the perceptual cues identified in this study. Lastly, the dependence on self-reported confidence measures may introduce subjective biases into the findings.

This study demonstrates that non-expert users can differentiate AI-generated images from authentic photographs with commendable accuracy; however, this ability is influenced by the characteristics of the images and the perceptual cues present. As generative models progress toward near-photographic realism, the assumption of visual authenticity must be reconsidered, highlighting the necessity for improved AI visual literacy, verification skills, and resilience in the face of synthetic media. Enhancing educational initiatives, fact-checking methodologies, and interdisciplinary research will be crucial to safeguarding public trust and fostering informed decision-making in an era where authentic and AI-generated visuals increasingly intersect.

Funding and Acknowledgements. Supported by the EKÖP-24-4-II-23 and EKÖP-2025-NKE-1-006 University Research Scholarship Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund. 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. Volume 252* (Part B), 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. 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. In *2022 IEEE International Conference on Data Science and Information System (ICDSIS)* (1–5). IEEE. 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. In *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses (RAID '23)* (654–668). Association for Computing Machinery. 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. In *2024 Third International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, (1–6). https://doi.org/10.1109/ICEEICT61591.2024.10718485

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, 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

Council of Europe. (n.d.). *Dealing with propaganda, misinformation and fake news—Democratic Schools for All*. Retrieved on 17 March 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

Fardouly, J., Diedrichs, P. C., Vartanian, L. R., & Halliwell, E. (2015). The mediating role of appearance comparisons in the relationship between media usage and self-objectification in young women. *Psychology of Women Quarterly.* 39(4), 447–457. https://doi.org/10.1177/0361684315581841

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* [Preprint]. *arXiv.* https://arxiv.org/abs/2403.17608

Grut, S. (2024). Source-Critical Affordances in Social Media Apps. *International Journal of Communication.* 18, Article 0. https://ijoc.org/index.php/ijoc/article/view/21892

Gupta, A. S., Shreneter, K. P., & Sehgal, S. (2024). Visual Veracity: Advancing AI-Generated Image Detection with Convolutional Neural Networks. In *2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)* (1–6). IEEE. 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. In *Companion Proceedings of the ACM on Web Conference 2025 (WWW '25)* (733-736). Association for Computing Machinery. 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*. 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/

Kramer, R. S. S., Young, A. W., & Burton, A. M. (2018). Understanding face familiarity. *Cognition.* 172, 46–58. https://doi.org/10.1016/j.cognition.2017.12.005

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 Volume 14982* (330–331). 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.

Mirsky, Y., & Lee, W. (2021). The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surveys*, 54(1), Article 7. 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. In *2024 International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS*) (1–7). IEEE. 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. HCII 2024 Volume 14702* (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. In *2023 5th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N*), 1104–1107. IEEE. https://doi.org/10.1109/ICAC3N60023.2023.10541758

Tandoc Jr., E. C., Lim ,Zheng Wei, & and 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. In *2024 IEEE International Conference on Big Data (BigData),* 3313–3321. IEEE. 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. In *Proceedings of the 19th International Conference on Predictive Models and Data Analytics in Software Engineering (PROMISE 2023)* 33–43. Association for Computing Machinery. https://doi.org/10.1145/3617555.3617874

Tóth, R., Bisztray, T., & Erdődi, 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 SAFECOMP 2024 Volume 14989* (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), 1146-1151. 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-photossynthid-ai-watermarks-magic-editor

Wilser, J. (2024, October 30). Content Credentials: The 200 Best Inventions of 2024. *TIME.* https://time.com/7094554/content-credentials/

Zhou, X., Burton, A. M., & Jenkins, R. (2021). Two factors in face recognition: Whether you know the person's face and whether you share the person's race. *Perception.* 50(6), 524–539. https://doi.org/10.1177/03010066211014016

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. [Preprint]. *arXiv.*https://arxiv.org/abs/2306.08571