The Role of AI Literacy and Media Education Interventions in Deconstructing Deepfakes: A Case Study of Gen Z in Croatia

ORIGINAL SCIENTIFIC PAPER
Received: 18. 3. 2026.
Accepted: 23. 4. 2026.
UDK
37.014.22+001.891.5
37:004(497.5)-053.6
https://doi.org/10.59549/n.167.1-2.1

Full. Prof. Igor Kanižaj, PhD
Catholic University of Croatia, Zagreb,
igor.kanizaj@unicath.hr;
orcid.org/0000-0002-8807-3655

Assoc. Prof. Lana Ciboci Perša, PhD
Catholic University of Croatia, Zagreb,
lana.persa@unicath.hr;
orcid.org/0000-0003-1495-9573

Abstract

One of the most concerning manifestations of artificial intelligence (AI) misuse is deepfake technology, which generates realistic but fabricated media content with the potential to deceive, manipulate, and erode public trust in digital communication. This threat is particularly acute for Generation Z, whose high exposure to synthetic media underscores the urgency of developing necessary critical skills. This study investigates the effectiveness of a short, targeted educational intervention designed to enhance deepfake recognition skills. Conducted in six secondary schools in Croatia, the intervention involved 118 students who participated in workshops focusing on the recognition, analysis, and critical evaluation of AI-generated manipulative content. A pre-test and post-test design was used to measure changes in students' competencies related to AI and media literacy. Findings indicate that even a short one-time intervention can produce measurable improvements in students' critical thinking, awareness of AI-generated content, and ability to assess digital authenticity.

Keywords: AI literacy; deepfake; Gen Z; media literacy interventions

1. Introduction

Artificial intelligence (AI) has been a topic of interest for many years (Dai & Ke, 2022), but its recent rapid growth and integration into daily life – from adults to Generation Alpha and Z – have intensified expert debate. One of the most pressing concerns is the potential misuse of AI to spread false information. Deepfake technology, which creates highly realistic but fabricated content, has emerged as a significant threat in this regard. Such AI-generated videos and images have the potential to mislead, manipulate, and harm individuals or societies, leading to the erosion of people's faith in digital content (Heidari et al., 2022).

"By definition, deepfakes are fabricated videos, images, and other media created by AI that appear to be real" (Ratner, 2021, p. 389). The term "deepfake" is derived from the words "deep learning" and "fake" (Abbas & Taeihagh, 2024; Jasserand, 2024). It "was first coined by the Reddit user u/deepfakes, who created a Reddit forum of the same name on 2 November 2017" (Ajder et al., 2019, p. 3). This forum, as emphasised by Ajder et al. (2019), focused on developing and using deep learning software to swap the faces of female celebrities into adult videos. It is believed that the field of pornography contains the most deepfake video content (Jasserand, 2024; Ajder et al., 2019). While most authors focus on the negative impacts of deepfake technology and content, Chapagain et al. (2024, p. 1) point out that such content also has positive aspects – they can entertain, educate, help society "produce accurate historical re-enactments, develop synthetic voices for speech-impaired people, and train AI systems to understand human emotions".

Deepfake content poses serious challenges in the digital age, particularly due to its ability to deceive and manipulate individuals on personal, social, and economic levels. Experts therefore call for stricter regulation and greater awareness of the potential risks associated with AI (Ratner, 2021). This issue is especially concerning for children, who are more vulnerable to such content and who spend significant amount of time on social media platforms (Dixon, 2023), where deepfakes are often encountered (Atlam et al., 2025). Recognising different types of video and audio manipulation on the Internet and in social media is becoming the ultimate challenge for our societies. Effective deepfake detection requires not only technical awareness of how AI-generated media are produced but also cognitive and ethical literacy to question digital content critically.

This article introduces the concept of AI literacy as a new approach to research of overall challenge of deepfakes deconstruction. Literacy in general is recognised as essential for successful participation in a democratic society (Turner et al., 2017) and extends far beyond reading and writing skills (McBride, 2015). From Richard Hog-

gart's perspective, it is more than just a skill or competence: "we all need literacy, imaginative and intellectual literacy, because it is an essential part of our movement towards greater critical self-awareness brought to bear on our own lives and on what society offers us as the desirable life" (Hoggart, 1980, p. 86). To empower users and citizens to transform societies and create precondition for citizens participation it is extensively promoted by many organisations such as UNESCO or OECD. From the educational perspective, the topic of literacies has evolved from the general concept of literacy to diverse multiple literacies, audio-visual, transliteracy, media literacy, visual literacy, digital literacy, algorithmic literacy and to the concept used in our article – AI literacy. It is defined as a "set of competencies that enables people to critically evaluate, communicate and collaborate effectively with AI" (Hargittai et al., 2020, as cited in Frau-Meigs, 2024, p. 5).

AI literacy is increasingly recognised as a core 21st-century skill, encompassing the ability to understand, interact with, and critically evaluate AI technologies and their societal implications (Ng et al., 2021; Long & Magerko, 2020). While AI literacy is a broad construct that includes technical understanding, ethical reasoning, and critical evaluation, this paper focuses specifically on crucial subdomains: the skills, ability and competence to recognise deepfakes. From this perspective, deepfake recognition serves as a concrete and socially relevant entry point for building AI literacy, particularly among adolescents navigating complex digital media environments. Developing these competencies is vital in fostering the ability of the average adolescent to navigate the digital environment safely and responsibly and to recognise and critically assess manipulated content online. As deepfakes become increasingly sophisticated and prevalent, a multifaceted approach that combines technological tools, human expertise, and media literacy education, may be necessary to combat the threats posed by AI-manipulated media (Somoray et al., 2025, p. 13). In this paper, we introduce the concept of educational interventions as a new path to build resilience against deepfakes.

To analyse the effectiveness of this type of interventions, we conducted a study with Croatian secondary school students from Generation Z. Anchored in the Southeast European context, this study also contributes to internationalisation of communication studies in relation to de-Westernisation of the discipline as previously emphasised by Waisbord and Mellado (2014, p. 370). The research in the Croatian cultural context involved an interactive educational session on deepfake content and methods for detecting such manipulations. Students were assessed both before and after the session by means of a test designed to evaluate their competencies in artificial intelligence literacy. The findings aim to advance understanding of how educational interventions can strengthen media and AI literacy, as well as critical thinking skills for recognising deepfake content.

2. Theoretical Background

2.1. Challenges in Deepfake Detection and Recognition

In the last five years, many authors have addressed the mechanisms and strategies for recognising deepfake content in the media (e.g., Tolosana et al., 2020; Nguyen et al., 2021; Seow et al., 2022; Dagar & Vishwakarma, 2022; Guarnera et al., 2022; Masood et al., 2023; Preeti Kumar & Kumar Sharma, 2023; Abbas & Taeihagh, 2024, Kaur et al., 2024). The strategies commonly employed for detecting deepfakes primarily rely on machine learning (ML) and deep learning (DL) techniques (Balara & Machová, 2024).

Deepfakes are difficult to recognise and deconstruct, typically appearing in audio and/or video content. Common video content techniques include face swap, face morphing, lip-synching, retouching, and face synthesis. In audio, techniques include text to speech synthesis and voice conversion (Somoray et al., 2025, p. 2). Researchers are facing additional two challenges: (a) determining the most effective methods of detecting deepfakes, and (b) identifying the methods of their production. Both basic human analysis techniques and AI-supported detection models are of great importance in this regard.

Balara and Machová (2024) identify four primary methods for creating fake human faces: (a) identity manipulation, (b) attribute manipulation, (c) expression manipulation, and (d) complete face synthesis. Identity manipulation involves swapping the face of one individual in an image with that of another, thus placing the second individual in a misleading scenario. Attribute manipulation entails altering various physical features of a person, including their face, eyes, hair, body shape, skin tone, or gender. Expression manipulation focuses on modifying the emotions or facial expressions displayed by an individual to convey different sentiments than those present in the original image. Finally, complete face synthesis refers to the generation of entirely artificial facial images using generative adversarial networks (GANs) (Balara & Machová, 2024). It is evident that the field of deepfake detection is rapidly evolving, with researchers exploring various machine learning techniques, dataset improvements, and multimodal approaches to address the challenges posed by increasingly realistic deepfakes. While significant progress has been made, the generalisation of detection models and the development of robust datasets remain critical areas for future research.

Detection and recognition of deepfakes for all exposed audiences is of great importance as deepfakes have many implications. It enables perpetrators to deceive and manipulate individuals (Abbas & Taeihagh, 2024), potentially leading to serious consequences ”on personal level (reputation loss, false accusation), social level

(political unrest, loss of trust to state or private institutions) or on economic level (financial fraud)" (Balara & Machová, 2024, p. 2). Therefore, it should come as no surprise that deepfakes are causing widespread fear (Ratner, 2021).

At the meta level, we witness the rapid development of AI that will open many new topics in this area. "The high fidelity of deepfakes, combined with their proliferation and the accessibility of the tools to generate them, will present an increasingly large challenge as investigators are expected to get confronted with growing volumes of disputed material" (de Leeuw den Bouter et al., 2024, p. 239).

2.2. Educational Interventions and AI Literacy

Analysing 32 scientific papers from relevant scientific databases, Chapagain et al. (2024) concluded that, in combating the impact of deepfake content, it is essential to raise awareness, implement regulatory measures, and promote media literacy with an emphasis on AI literacy. With the aim of raising awareness among media users, especially children and young people, "AI literacy has emerged as a new skill set that everyone should learn in response to this new era of intelligence" (Ng et al., 2021, p. 1). As well as media literacy, which many today consider an umbrella form of literacy that encompasses digital and even AI literacy (especially from the perspective of social sciences) (Frau-Meigs, 2024), AI literacy implies knowing and understanding AI (especially the technologies behind AI), knowing how to use and apply AI concepts in different contexts, and being aware of the ethical concerns related to AI technologies, as well as evaluating and creating content (Ng et al., 2021).

Educational interventions, whether short or extended, have been shown to produce measurable improvements in knowledge, critical thinking, and ethical awareness, particularly in the context of AI literacy and deepfake detection (Hollands & Breazeal, 2024). Both short and long interventions can be effective (even a single 10-minute deepfake detection session yielded a 33% accuracy improvement (Tahir et al., 2021), but longer or more frequent exposure (such as daily lessons compared to weekly) is associated with greater gains (Zhang et al., 2024). Some interventions increase scepticism, which may reduce trust in real content (Tahir et al., 2021). The majority of publicly available research has focused on interventions among teachers and adults (Hollands & Breazeal, 2024; Somoray & Miller; 2023, El Mokadem, 2023; Bray et al., 2023) and students/youth (Bhalli et al., 2024; Naffi et al., 2023; Kong et al., 2022), and only a small number have focused on school students (such as Zhang et al., 2024; Theophilou et al., 2022; Van Brummelen et al., 2021). This paper attempts to fill these gaps and investigate whether secondary school students recognise deepfake content, and whether short, one-off interventions in this area are

sufficient to raise students' awareness and skills in recognising manipulative media content.

Based upon the presented literature overview and objective of the article in this paper we introduce the following research question: RQ1: How do students define and recognise deepfakes before and after educational intervention? RQ2: What are the implications of the educational intervention on the ability to identify authentic content?

Based upon the research questions we have introduced the following hypotheses.

H1: A brief educational intervention will improve students' ability to recognise deepfakes

H2: The intervention will reduce students' accuracy in identifying authentic content.

3. Methodology

Prior to the hypotheses operationalisation we considered existing research insights from the previous studies. Somoray et al. (2025) reviewed 40 studies in the area of deepfakes detection. "The studies reviewed used a variety of performance metrics to measure detection performance. These metrics include forced choices, Likertscales, open-ended questions, area under the curve (AUC), F1 score, precision, true positive rates, false positive rates, and false negative rates" (Somoray et al., 2025, p. 6). They also concluded that the most common approach is presenting a stimulus to participants one at a time and directly asking participants to indicate whether the stimulus was fake or authentic (Somoray et al., 2025).

Based upon the research questions and hypotheses we operationalised key concepts through an online pre-test and post-test questionnaire. The instrument included a mix of open- and closed-ended questions and two types of stimuli.

This research explores the extent to which secondary school students can recognize, identify and detect deepfake content in mass media and on social media. Pre-test and post-test online questionnaires were used to assess and measure the impact of an educational media literacy intervention. One-group pre-test and post-test designs do not require large samples and are an alternative analytical method when randomisation in a small sample is not feasible (Harris et al., 2006 as cited in Harerimana et al., 2023). The intervention consisted of five different segments:

1. Online pre-test on existing knowledge on deepfake technology
2. Introduction to new topics within the AI literacy concept
3. Educational segment with real-life deepfake examples

4. Tutorial on open-source applications that can be used to detect deepfake tools – within the component of OSI (open-source intelligence)
5. Online post-test on existing knowledge on deepfake technology

The intervention aimed to enhance students' media literacy skills in deepfake analysis. In a 35-minute interactive intervention, secondary school students were exposed to ten deepfake examples. The educational intervention was planned and implemented based upon DigComp 2.2. – The Digital Competence Framework for Citizens upgraded by the MIL competence framework in Algo AI literacy[1] (v. 1.0) that was produced within the ALGOWATCH project[2]. Within the intervention, we focused on key indicators of deepfake content, such as a) unnatural facial expressions, b) abnormal eye movements, and c) mismatched audio and visuals.

Two online questionnaires (pre-test and post-test) were developed to conduct the research. The post-test was conducted immediately after the lecture to measure the effectiveness of the intervention. Both tests were conducted on the spot by means of online questionnaires. Both anonymised questionnaires included sociodemographic variables, followed by indicators for assessing the understanding and definition of deepfake technology, but also questions on strategies and tools for identifying deepfake video and photo content. In addition to sociodemographic data, the pre- and post-test consisted of questions that measured theoretical knowledge of deepfake content in the media, but also practical skills in recognising deepfake content in photos and video formats.

The research and class media literacy intervention were conducted from February to April 2025 in six different secondary schools across three Zagreb region communities (the City of Zagreb, Samobor and Ivanić-Grad), Croatia. A total of 118 students participated in the research (pre-test and post-test); 61% identified as female and 33.9% as male, while 5.1% chose not to disclose their gender. Among the respondents, 56.8% attended general secondary school, while 43.2% were enrolled in a four-year vocational secondary school. The average age of participants was 16.6. Both questionnaires were self-completed by students in Google Forms, before and after the intervention. We conducted a chi-square test of homogeneity to compare the distribution of the number of recognized categories between the pre-test and

---

[1] https://algowatch.eu/wp-content/uploads/2024/10/Competence-Framework-Algowatch-ENGLISH.pdf

[2] The European project Algowatch was granted as part of the CREA-MIL 2023 call for proposals. It began in October 2023 and will run for 2 years. It focuses on educating young people and the general public about the challenges of algorithms and Artificial Intelligence (Algo- and AI-literacy) in the field of information and digital citizenship. Retrieved March 31, 2025, from https://algowatch.eu,

post-test in a sample of 118 participants. A one-tailed Wilcoxon signed-rank test was used to test whether post-intervention theory scores exceed pre-intervention scores.

4. Results

4.1. Pre-testing – Deepfake Definition

Prior to the educational intervention, that is during the first part of the research, students were asked to define deepfake based upon their existing knowledge on deepfakes as phenomena. In pre-testing, one in five students admitted that they did not know how to define deepfake. Others defined it in various ways, such as "fake profiles", "false information", "lies on the Internet", "fake videos that closely mimic reality", "disinformation", "frauds related to artificial intelligence". Examples include:

- "Deepfake is artificially generated content where a person from the original recording or photo is replaced with another person to spread false information or deceive."
- "Deepfakes are videos where the person speaking is not real, but everything is created using artificial intelligence, although it looks very realistic."
- "Images or videos that appear realistic but are fake, such as using someone's voice, images, or videos."
- "Videos of real people saying something, but they are not saying it; instead, artificial intelligence is used to make it appear as though the person is speaking."
- "Creations of artificial intelligence that depict a selected person doing or saying something they never actually did or said, based on what the person instructing the AI decides."

Based upon the definitions provided by the students in the open questions, we can assert that most of the respondents were, to some extent, informed about deepfakes. Most students recognised that deepfakes are realistic enough to convincingly replicate real-life events. Noteworthy, however, was the difference in definitions provided by the students. Although some students went to the extent of explaining the definition using AI software and content inconsistencies, there were those whose explanations were overly vague and general such as "fake profiles" or "lies on the Internet". Precisely, this last definition lacks the technical scope and overall perspective, as well as significant details. In addition, almost all students had a negative perception and attitude towards deepfakes, demonstrating their previous knowledge about the phenomenon. Furthermore, most students associate deepfakes with malicious objectives, including harming or damaging a person's reputation, disinforma-

tion, or manipulating public opinion. This negative perspective is also a reflection of their previous experience. The AI-powered technological intervention helped students change their perceptions of deepfakes.

Before the intervention, 87% of students knew that deepfake technology primarily uses artificial intelligence to create fake videos or audio. This was expected due to their evident everyday exposure to these types of contents in their lives, leading to a reasonable understanding of what deepfakes can do and to what type of content are they exposed to primarily on their social media channels. Additionally, 93.9% of respondents understood deepfakes are risky because they can spread false information and influence public opinion. Our research shows that students are aware of the dangers associated with deepfakes and with the possible negative impact that deepfakes could have in their lives. Our research also shows that 66.7% of students can detect deepfakes prior to educational intervention. These skills are demonstrated by the students with the usage of sources verification, implementation and usage of detection tools, or by recognising a specific inconsistency in media. Due to a relatively huge number of respondents with the expressed ability to recognise and detect deepfakes in the pre-test, our goal, with the educational intervention, was to even further increase the number of respondents with these types of skills.

4.2. Main Changes in Deepfake Definition

After the intervention we noticed that students were able to provide more arguments to the definition as their explanations of the topic did not consist predominantly of negative characteristics of deepfakes, as registered in the pre testing. Most of them could provide a more detailed definition, with just a single student unable to explain the concept at all.

87.2% of students recognised that critical thinking and verifying sources are the best defence against deepfakes, indicating a better reasoning of prevention strategies. Furthermore, 72.4% became knowledgeable about using AI tools to detect deepfakes, showing an expanded grasp of technological solutions.

The study also explored students' views and worries about deepfakes in general. One of our goals was also to gain better insight in their existing experience and the recognition of deepfakes on their social media channels and profiles. About 66.4% reported seeing deepfake videos on TikTok, which suggests many students frequently encounter this type of content. Interestingly, 73.9% believed deepfakes are dangerous due to their ability to disseminate false information, which reflects a strong awareness of their societal impact. Despite this awareness, only 43.1% stated that they would report deepfake videos that misrepresented someone negatively, suggesting a possible lack of confidence or motivation to tackle harmful content.

Meanwhile, 42.2% expressed concern that deepfake videos could target them or their friends, pointing to personal anxieties. Lastly, 51.3% of respondents stated that schools should provide lessons on recognising deepfakes, emphasising education's importance in addressing this issue.

Interestingly, as many as 47.4% of students disagreed with the idea that recognising a fake image is more straightforward than spotting a fake video. This indicates that images and videos are complex for students to identify as fake, likely because deepfake technology is improving and becoming more complex. As a result, there is a clear need for more advanced training and better tools to help students effectively distinguish between manipulated images and videos.

4.3. Effectiveness of the Educational Intervention – Ability to Recognise Deepfakes

Students rated their personal skills in recognising deepfake content in media and social media on a scale from 1-5, with an average of 3.31 before and 3.56 after the intervention. This is a positive change in students' confidence regarding their ability to identify deepfake content. The increase of 0.25 points on the scale reflects the effectiveness of the 35-minute interactive lecture, which provided the students with practical examples and strategies for recognising deepfake technology. These results suggest that the educational intervention enhanced the students' media literacy and awareness of deepfake content.

Pre-test and post-test results indicate a significant improvement in student's ability to identify specific signs of deepfake content after the educational intervention. The percentage of students recognising key indicators of deepfake manipulation increased across all categories (see Table 1). These improvements suggest that the educational intervention effectively enhanced students' ability to detect inconsistencies in deepfake content. The increase in recognition of more nuanced indicators, such as spots between the background and the face (from 44% to 74.1%) and asymmetrical display of glasses (from 27.6% to 63.8%), highlights the effectiveness of the media literacy intervention.

In relation to earlier findings that students' self-assessed skills in recognising deepfake content improved (from 3.31 to 3.56), it is evident that the intervention increased their confidence and provided them with new insight into practical tools to identify deepfake content more effectively. This aligns with the broader goal of the study, which was to enhance media literacy and critical thinking skills among secondary school students in the context of emerging technologies like deepfakes.

We conducted a chi-square test of homogeneity to compare the distribution of the number of recognized categories between the pre-test and post-test in a sam-

ple of 118 participants. The resulting test statistic was $\chi^2 = 10.95$ with 8 degrees of freedom and a p-value of approximately 0.205, indicating no statistically significant difference in the overall distribution between the two measurements. Although some individual categories appear different in terms of percentages, overall the change in distribution is not large enough to be statistically significant according to this test.

Table 1: Signs that Help Recognise that an Image or Video Might Be a Deepfake? (%) (N=118)

	Pre-test	Post-test
Eyes look unnatural, often a different colour	60.3%	81.9%
Facial movements are not coordinated with the voice	82.8%	96.6%
Background looks strange or blurry	62.9%	80.2%
Skin looks too smooth or plastic-like	64.7%	80.2%
Disproportionate lip movements, especially lip and voice synchronisation	72.4%	91.4%
Unusual blinking and various other facial irregularities	53.4%	77.6%
Spots between the background and the face	44.0%	74.1%
Asymmetrical display of glasses	27.6%	63.8%
Blurred or unclear edges of a person	62.1%	70.7%

Source: own processing, 2025

4.4. Distinguishing Real from Deepfake Content – Reduced Accuracy in Identifying Authentic Content

Within the media literacy intervention, one of the assignments required students to distinguish actual from deepfake content. This activity was introduced after the knowledge intervention with guidelines on how to reveal mixed outcomes regarding the effectiveness of the educational intervention. While there was a notable improvement in recognising deepfake content, the ability to identify actual content decreased significantly (Table 2). After being exposed to selected techniques and approaches to deconstruction of deepfakes, the students became more sceptical about the content to which they were exposed. The percentage of students correctly identifying deepfake photographs increased from 81.4% to 88.7%, and for deepfake videos from 85.2% to 90.3%. This suggests that the intervention successfully enhanced students' ability to detect manipulated content, likely due to the emphasis on specific indicators such

as unnatural facial movements, mismatched audio-visual cues, and other inconsistencies.

Conversely, the percentage of students correctly identifying real photographs dropped from 74.6% to 35.3%, and for real videos from 72.1% to 58.4%. This decline may indicate that the intervention inadvertently caused students to become overly sceptical, leading them to misclassify authentic content as deepfake. As they learned to scrutinise various manipulation indicators, they may have developed a heightened suspicion regarding the authenticity of real images and videos. This scepticism could confuse them, causing them to misidentify genuine content as manipulated. Similar findings were reported by other authors, e.g. Tahir et al. (2021).

Table 2: Correct Identification of Real and Deepfake Content (%) (N=118)

	Pre-test	Post-test
Real photo recognition	74.6%	35.3%
Deepfake photo recognition	81.4%	88.7%
Real video recognition	72.1%	58.4%
Deepfake video recognition	85.2%	90.3%

Source: own processing, 2025

At the end, students' performance for each of the four questions was scored as correct (1) or incorrect (0), and these scores were totalled to yield a total "correct" score (range 0–4) for each student, both before and after the intervention. The normality of the paired difference scores was assessed using the Shapiro–Wilk test, which indicated a significant departure from normality ($p < .05$). Consequently, non-parametric Wilcoxon signed-rank tests were employed for both hypotheses. A one-tailed Wilcoxon signed-rank test was used to test whether post-intervention theory scores exceed pre-intervention scores. Results showed a significant positive shift ($V = 1617$, $p = .0219$), indicating that students' ability to recognise deepfake content was greater after the educational intervention than before. A parallel one-tailed Wilcoxon signed-rank test examined whether post-intervention photo-based scores exceeded their pre-intervention counterparts. This test was not significant ($V = 1216.5$, $p = .998$), providing no evidence that students' understanding of deepfake indicators (as measured by the photo task) improved in the hypothesised direction.

5. Discussion

„It is evident that the deepfake detection methods are continuously struggling to catch up to deepfake generation methods." (Saif & Tehseen, 2022, p. 3004). The challenge of the rapid AI and deepfakes development remains. Two main approaches focus on (a) the competence to detect and recognise deepfakes and (b) the techniques and tools distinguishing AI-generated content from human-generated content.

Previous research concludes that media literacy interventions in deepfake recognition and identification are welcomed (Somoray et al., 2025). We believe, that with the effective combination of the two previously mentioned approaches, the relevance of the interventions will not be questioned by the rapid development of the deepfake technology. Our research confirms that short, well-structured educational interventions can positively influence students' ability to recognise and critically evaluate deepfake content, thereby enhancing key aspects of AI and media literacy. The findings highlight the value of incorporating deepfake detection as an entry point to broader AI literacy, particularly in educational settings where exposure to such content is high, and formal curricula lack targeted content in this area.

This study contributes to the growing body of literature emphasising the need for interdisciplinary and age-appropriate AI education. It also provides empirical support for integrating such interventions into formal schooling, using adaptable models that can be replicated in various European contexts. While the intervention demonstrated immediate learning gains, future research should examine the long-term effects of such programmes and explore how repeated or integrated instruction influences skill retention and critical awareness. Further investigations should also address the development of comprehensive AI literacy curricula that span multiple educational levels and subjects, in line with recommendations from existing educational frameworks. Longitudinal studies, mixed-method approaches, and cross-national comparisons will be crucial to refine instructional strategies and inform policy aimed at ensuring that younger generations are not only consumers of AI-powered content but also critical, ethical, and informed users of such technologies.

6. Conclusion

Based on our findings, we identify seven key challenges:

1. Gen Z students in Croatia have insufficient knowledge about deepfakes and synthetic media, likely due to curricula needing updates to include new AI competencies and skills.

2. The school environment is a great opportunity for suitable media literacy interventions. In comparison with other communities and settings, it provides a good structure and environment for other similar types of interventions.
3. Effective media literacy interventions can lead to an increase in critical awareness about deepfake content, even during a limited period, which is considered as a significant argument for the inclusion and engagement of the teachers in the workshops.
4. Educational interventions, as presented in this paper, can also cause a higher level of scepticism towards original/genuine content in comparison with synthetic. This needs to be taken into account especially during implementation of similar programmes at the European level.
5. A comprehensive, systematic and long-term approach is required to provide in-depth analysis, training and education on algorithms.
6. We must focus on a new paradigm that will enable us to critically evaluate the work of algorithms based upon media algorithm and media literacy interventions.
7. Students showed high motivation and interest for the topic, and this topic should be promoted on a larger scale as a part of wide-ranging media literacy interventions.

We believe that the presented model of media literacy interventions could be replicated in other countries as well. The structure of the intervention, as well as its applicability, including the type and structure of the examples used in the intervention could be contextualised to most of education systems, particularly in Southeast Europe. Although our focus in this research was not the regulatory aspect of deepfakes, there is an urgent need for a detailed regulatory framework that would provide guidance and guidelines for the average user.

Finally, another important segment of the intervention needs to be further developed and investigated – the teachers' perspective and experience with the AI challenges. A more qualitative approach that could provide valuable feedback and evaluation for future research and analysis is needed.

Acknowledgement: A part of this research was conducted within the Creative Europe project ALGOWATCH and funded by the European Union.

Bibliography

Abbas, F., & Taeihagh, A. (2024). A systematic review of deepfake detection and generation techniques using artificial intelligence. *Expert Systems with Applications*, *252*, Article 124260. https://doi.org/10.1016/j.eswa.2024.124260

Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019). *The state of deepfakes: landscape, threats, and impact*. Deeptrace.

Atlam, E.-S., Almaliki, M., Elmarhomy, G., Almars, A. M., Elsiddieg, A. M. A., & ElAgamy, R. (2025). SLM-DFS: A systematic literature map of deepfake spread on social media. *Alexandria Engineering Journal, 111*, 446-455. https://doi.org/10.1016/j.aej.2024.10.076

Balara, V., & Machová, K. (2024). Detection of artificially created Faces with convolutional Networks. *2024 International Conference on Emerging eLearning Technologies and Applications (ICETA),* 10.1109/ICETA63795.2024.10850853

Bhalli, N.; Naqvi, N.; Evered, C.; Mallinson, C.; & Janeja, V. P. (2024). Listening for expert identified linguistic features: Assessment of audio deepfake discernment among undergraduate students. *arXiv:2411.14586v1.* https://arxiv.org/pdf/2411.14586

Bray, S. D., Johnson, S. D., & Kleinberg, B. (2023). Testing human ability to detect 'deepfake' images of human faces. *Journal of Cybersecurity, 9*(1), 1-18. https://doi.org/10.1093/cybsec/tyad011

Chapagain, D., Kshetri, N., & Aryal, B. (2024). Deepfake disasters: A comprehensive review of technology, ethical concerns, countermeasures, and societal implications. *2024 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC).* 10.1109/ETNCC63262.2024.10767452

Dagar, D., & Vishwakarma, D. K. (2022). A literature review and perspectives in deepfakes: Generation, detection, and applications. *International Journal of Multimedia Information Retrieval, 11*(3), 219–289. https://doi.org/10.1007/s13735-022-00241-w

Dai, C.-P., & Ke, F. (2022). Educational applications of artificial intelligence in simulation-based learning: A systematic mapping review. *Computers and Education: Artificial Intelligence,* 3, Article 100087. https://doi.org/10.1016/j.caeai.2022.100087

Dixon, S. J. (2023) Most used social media platforms among Gen Z and internet users worldwide as of September 2023. *Statista.* https://www.statista.com/statistics/1446950/gen-z-internet-users-social-media-use/

El Mokadem, S. S. (2023). The effect of media literacy on misinformation and deepfake video detection. *Arab Media & Society*, *35*. https://doi.org/10.70090/SM23EMLM.

Frau-Meigs, D. (2024). *User empowerment through media and information literacy responses to the evolution of generative artificial intelligence (GAI).* Paris: UNESCO.

Guarnera, L., Giudice, O., Guarnera, F., Ortis, A., Puglisi, G., Paratore, A., Bui, L. M. Q., Fontani, M., Coccomini, D. A., & Caldelli, R. (2022). The face deepfake detection. *Challenge. J. Imaging, 8*, 263. https://doi.org/10.3390/jimaging8100263

Harerimana, A., Duma, S. E., & Mtshali, N. G. (2023) Measuring perceived learning gains of undergraduate nursing students in ICT skills: One group pre-test and post-test design. *Contemporary Nurse, 59*(2), 114-131, https://doi.org/10.1080/10376178.2023.2230309

Heidari, A., Navimipour, N. J., Dag, H., & Unal, M. (2022). Deepfake detection using deep learning methods: A systematic and comprehensive review. *WIREs Data Mining Knowl Discov., 14*, e1520. https://doi.org/10.1002/widm.1520

Hoggart, R. (1980) The importance of literacy, *Journal of Basic Writing*, 3 (1), 74-87.

Hollands, F., & Breazeal, C. (2024) Establishing AI literacy before adopting AI, *The Science Teacher, 91*(2), 35-42, https://doi.org/10.1080/00368555.2024.2308316

Jasserand, C. (2024). Deceptive deepfakes: Is the law coping with AI-altered representations of ourselves? *2024 International Conference of the Biometrics Special Interest Group (BIOSIG).* https://ieeexplore.ieee.org/document/10786729

Kaur, A., Hoshyar, A. N., Saikrishna, V., Firmin, S., & Xia, F. (2024). Deepfake video detection: challenges and opportunities. *Artificial Intelligence Review*, *57*(159), https://doi.org/10.1007/s10462-024-10810-6

Kong, S.-C., Cheung, W. M.-Y., & Zhang, G. (2022). Evaluating artificial intelligence literacy courses for fostering conceptual learning, literacy and empowerment in university students: Refocusing to conceptual building. *Computers in Human Behavior Reports*, 7, 100223. https://doi.org/10.1016/j.chbr.2022.100223

Lee, J., & Kwon, K. (2024). A systematic review of AI education in K-12 classrooms from 2018 to 2023: Topics, strategies, and learning outcomes. *Computers and Education: Artificial Intelligence 6,* 100211. https://doi.org/10.1016/j.caeai.2024.100211

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1-16, https://doi.org/10.1145/3313831.3376727

Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A., & Malik, H. (2023). Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied Intelligence, 53*(4), 3974–4026. https://doi.org/10.1007/s10489-022-03766-z

Naffi, N., Charest, M., Danis, S., Pique, L., Davidson, A. L., Brault, N., & Barma, S. (2023). Empowering youth to combat malicious deepfakes and disinformation: An experiential and reflective learning experience informed by personal construct theory. *Journal of Constructivist Psychology*, *38*(1), 119–140. https://doi.org/10.1080/10720537.2023.2294314

Ng, D. T. K., Leung. J. K. L., Chu, S. K. W., & Qiao, M. S. (2021) Conceptualizing AI literacy: an exploratory review. *Computers and Education: Artificial Intelligence, 2*: 100041. https://doi.org/10.1016/j.caeai.2021.100041

Nguyen, X. H., Tran, T. S., Le, V. T., Nguyen, K. D., & Truong, D.-T. (2021). Learning Spatio-temporal features to detect manipulated facial videos created by the Deepfake techniques. *Forensic Science International: Digital Investigation*, *36*, 301108. https://doi.org/10.1016/j.fsidi.2021.301108

Preeti, Kumar, M., & Kumar Sharma, H. (2023). A GAN-based model of deepfake detection in social media. *Procedia Computer Science*, *218*, 2153-2162. https://doi.org/10.1016/j.procs.2023.01.191

Ratner, C. (2021). WHEN "Sweetie" is not so sweet: Artificial intelligence and its implications for child pornography. *Family Court Review, 59*(2), 386–401. doi: 10.1111/fcre.12576

Saif, S., & Tehseen, S. (2022). Deepfake videos: synthesis and detection techniques – a survey. *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*, 42 (4), 2989–3009. https://doi.org/10.3233/JIFS-210625

Seow, J. W., Lim, M. K., Phan, R. C. W., & Liu, J. K. (2022). A comprehensive overview of Deepfake: Generation, detection, datasets, and opportunities. *Neurocomputing, 513*, 351–371. https://doi.org/10.1016/j.neucom.2022.09.135

Somoray, K., & Miller, D. J. (2023). Providing detection strategies to improve human detection of deepfakes: An experimental study. *Computers in Human Behavior, 149*, 107917. https://doi.org/10.1016/j.chb.2023.107917

Somoray, K., Miller, D. J., & Holmes, M. (2025). Human performance in deepfake detection: A systematic review, *Human Behavior and Emerging Technologies,* 1-15 https://doi.org/10.1155/hbe2/183322

Tahir, R., Batool, B., Jamshed, H., Jameel, M., Anwar, M., Ahmed, F., Zaffar, M. A., & Zaffar, M. F. (2021). Seeing is believing: Exploring perceptual differences in deepFake videos. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. https://manwar.web.illinois.edu/files/DeepFakes___After_Rebuttal.pdf

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A Survey of face manipulation and fake detection. *Information Fusion*, *64*, 131–148. https://doi.org/10.1016/j.inffus.2020.06.014

UNESCO (2022). *K-12 AI curricula. A mapping of government-endorsed AI curricula*. UNESCO.

Van Brummelen, J., Heng, T., & Tabunshchyk, V. (2021). Teaching tech to talk: K-12 conversational artificial intelligence literacy curriculum and development tools. *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*. https://ojs.aaai.org/index.php/AAAI/article/view/17844

Waisbord, S., & Mellado, C. (2014). De-westernizing communication studies: A reassessment. *Communication Theory, 24*(4), 361-372. https://doi.org/10.1111/comt.12044

Zhang, H., Lee, I., & Moore, K. (2024). An effectiveness study of teacher-led AI literacy curriculum in K-12 classrooms. *The Thirty-Eighth AAAI Conference on Artificial Intelligence* (AAAI-24). https://doi.org/10.1609/aaai.v38i21.30380

Uloga pismenosti o umjetnoj inteligenciji i intervencija medijskoga obrazovanja u dekonstrukciji lažnih generiranih sadržaja: Studija slučaja generacije Z u Hrvatskoj

Sažetak

Jedna od najzabrinjavajućih zlouporaba umjetne inteligencije (UI) jest tzv. *deepfake* tehnologija koja generira realističan, ali izmišljen medijski sadržaj s potencijalom obmanjivanja, manipulacije i narušavanja javnog povjerenja u digitalnu komunikaciju. Ta je prijetnja posebno izražena za generaciju Z, čija visoka izloženost sintetičkim medijima naglašava hitnost razvoja potrebnih kritičkih vještina. Ovaj rad ispituje učinkovitost kratke, ciljane obrazovne intervencije osmišljene za unaprjeđenje vještina prepoznavanja *deepfake* sadržaja. Provedena u šest srednjih škola u Hrvatskoj, intervencija je obuhvatila 118 učenika koji su sudjelovali u radionicama usmjerenima na prepoznavanje, analizu i kritičko vrednovanje manipulativnog sadržaja generiranog umjetnom inteligencijom. Za mjerenje promjena u kompetencijama učenika povezanima s umjetnom inteligencijom i medijskom pismenošću primijenjen je nacrt s predtestom i posttestom. Nalazi pokazuju da čak i kratka jednokratna intervencija može dovesti do mjerljivih poboljšanja u kritičkom mišljenju učenika, njihovoj svijesti o sadržaju generiranom umjetnom inteligencijom te sposobnosti procjene digitalne autentičnosti.

Ključne riječi: *deepfake* sadržaji; generacija Z; intervencije medijske pismenosti; pismenost o umjetnoj inteligenciji