

Developing deepfake literacy among pre-service teachers: a qualitative study in the context of immigration

Pedro Miralles-Sánchez

`pedro.miralles@unavarra.es`

Universidad Pública de Navarra (UPNA)

José María Campillo-Ferrer

Universidad de Murcia

Research Article

Keywords: artificial intelligence, digital literacy, social sciences education, migration, higher education

Posted Date: October 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11074976/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

The rise in deepfakes generated using Artificial Intelligence (AI) poses a challenge for digital literacy and the critical evaluation of information, particularly on sensitive issues such as immigration. This study analysed the perceptions of 133 students on the bachelor's degree in Primary Education at the University of Murcia regarding the identification of manipulated content. After receiving specific training in deepfake detection, the participants assessed various materials relating to immigration and justified their classification as genuine or fake. The results suggested a more systematic use of technical and source-verification criteria after the training. The most used strategies were verifying primary sources, contextualising the information and detecting anomalies in the format. However, unvalidated procedures were also observed, such as confirmation bias and the attribution of authority. In conclusion, the findings suggest that specific training may encourage more systematic approaches to evaluating potentially manipulated content, although the persistence of cognitive biases highlights the need to reinforce digital literacy and critical thinking in initial teacher training.

1 Introduction

As globalization progresses, powerful new computer programs and applications are spreading rapidly and improving the efficiency of services and products worldwide. Innovative technology-driven initiatives, such as generative AI, are capable of generating a variety of outputs, such as images, videos or code, without the need for direct human intervention. This represents a digital turning point, giving rise to a wide range of applications across various fields, including business, engineering, healthcare and, in particular, education.

Recent research has highlighted the potential benefits and drawbacks of these online tools as technology advances. Among the advantages of these online solutions, some authors highlight the creation of new career opportunities within a wide variety of transformative educational contexts worldwide (Almisad and Aleidan, 2025; Campillo-Ferrer et al., 2025). Specifically, Baidoo-Anu and Ansah (2023) listed several advantages for teaching and learning, such as the use of AI-based tutoring systems, the development of deep learning for adaptive purposes, and improvements in the automated marking of written work. Yan et al. (2024) also highlighted the contribution of these technological tools to improving the interactive exchange of ideas and information and to effective learning of programming, which can ultimately enhance students' learning experiences by fostering their critical thinking skills.

Despite the effective uses of generative AI in successfully addressing numerous critical problems, such as big data analysis, virtual tutoring and the creation of educational content, recent literature also highlights that the advancement of these online teaching materials may entail certain drawbacks in the educational sphere. Kartheeban et al. (2025) highlight algorithmic bias, data security and transparency as some of the most significant drawbacks associated with the implementation of AI-based tools. Atkinson (2023) focuses on plagiarism arising from issues of authorship. In this vein, Håkansson and Phillips-Wren (2024) raise some ethical concerns regarding the generation of incorrect information or

erroneous judgements. Deepfakes represent another key issue relating to the development of AI-based skills and the potential threats they pose in various educational settings. The term was coined by combining the words 'deep learning' and 'fake' to refer to synthetic multimedia content in a wide range of formats: code, images, video and other media.

This synthetic content is generated by AI that replicates the original appearance in order to artificially imitate it, which can facilitate manipulation and disinformation. The potential harm caused by deepfakes, as well as their accessibility and perceived authenticity, can make it difficult for internet users to distinguish between genuine and fake digital content (Overton, 2024). In this regard, the current situation may lead to two extremes: either a growing mistrust and the constant adoption of verification measures to check every piece of data, or extreme credulity, due mainly to a lack of resources to verify the information (Michałkiewicz-Kądziela, 2024). Both situations can have negative effects on relations between educators, legislators, political leaders and students due to the spread of misleading information, which may undermine democratic processes and trust in educational principles. Examples of deepfakes have been cited in recent literature; for instance, in India, where a recording was circulated the day before the elections in which a senior politician criticized his rival (Safiullah and Parveen 2022, 256). We have also seen fake videos of prominent political leaders, such as Vladimir Putin and Volodymyr Zelensky, debating the crisis surrounding the war in Ukraine (Twomey et al., 2023). Another widely circulated deepfake relates to alleged insults by former US President Obama towards current President Trump, which turned out to be false (Sorell, 2023). Other digital manipulations target government institutions and headquarters, such as the image showing a column of black smoke in 2023 in front of the Pentagon. The image appeared authentic and was shared on Twitter, causing panic over the possibility of a terrorist attack, such as the one that occurred two decades earlier, in 2001, when the attack on the Twin Towers took place – an event that became the worst in US history (Ninovic, 2024). The conclusions drawn from these studies raise several significant issues that must now be addressed when using these digital tools.

Given that deepfakes blur the boundaries of the real world, education must make progress in detecting them and equip students to navigate the internet in search of the truth and a more objective understanding of reality (Hameleers et al., 2022). To achieve this, UNESCO's guidelines on generative AI in education and research focus on integrating ethical strategies into AI-driven educational practices to ensure equity, responsible engagement and the development of professional skills during teaching and learning processes (Fengchun and Wayne, 2023). In line with this, other institutions, such as Macquarie University and the Queensland University of Technology (QUT), have in recent years established an effective framework for the appropriate implementation of AI-based resources (Smith et al., 2026). Other examples include academic integrity committees such as the Committee on Publication Ethics (COPE), which published a code of ethical principles to promote source verification and transparency (Pearson, 2024). It is therefore worth emphasizing that the veracity and accuracy of data are essential for fostering academic integrity amongst teachers, as they ensure that the content covered in class is truthful, empirically grounded and reflects real-world challenges as they occur, such as migration crises or armed conflicts.

This means that the use of these web-based resources must be supervised by teachers, recent studies have examined the use of deepfakes in higher education contexts, but it seems essential to explore the perspectives of trainee teachers in greater depth, given their key role in supervising students' learning processes when using these AI-powered resources in the classroom. In this regard, understanding the reflections, concerns and perceptions of trainee teachers is important for determining the extent to which they will be able to address the challenges posed by generative AI in their future careers. In this regard, the challenges associated with confirmation bias, examined by Ebner et al. (2024), or the relationship between higher education students' ability to verify the authenticity of web resources and their levels of digital literacy, should be the focus of current research, going beyond studies on traditional learning processes.

In this context, this research aims to examine pre-service teachers' perceptions of deepfakes and to draw meaningful comparisons between the justifications they put forward before and after the academic term regarding how to detect them. To this end, the following specific research objectives have been established, which determine the structure of the study.

- RO1: To explore the justifications put forward by participants for determining whether digital content is false, prior to the training program.
- RO2: To explore the justifications put forward by participants for determining whether digital content is genuine, prior to the training program.
- RO3: To explore the justifications given by participants for determining whether digital content is false, following the training program.
- RO4: To explore the justifications given by participants for determining whether digital content is genuine, following the training program.

2 Methods

2.1. Participants

The study involved 133 second year students on the bachelor's degree program in Primary Education at the University of Murcia, located in the Region of Murcia, Spain. As regards the sampling procedure in this research, it can be defined as non-probabilistic and convenience-based (Howarth et al., 2024), since the respondents were selected based on ease of access, once their participation in the training program focused on generative AI had been confirmed. The research fell within the field of social sciences teaching and was carried out as part of the compulsory module 'Social Sciences and their teaching' during the 2025/26 academic year. In terms of their socio-cultural characteristics, the participating students were from middle-class backgrounds, had a vocation for teaching, and shared similar academic motivations whilst studying for their undergraduate degree in Primary Education.

2.2. Instruments

Students on the bachelor's degree in Primary Education completed a questionnaire designed to assess the risks associated with AI and the proliferation of deepfakes, which had previously been implemented and validated by Neyazi et al. (2024). These tools, widely used in the social sciences, are among the most effective means of data collection, as they facilitate anonymity during the data collection process associated with academic research in this field. Furthermore, it is worth noting that previous research has highlighted certain advantages associated with their use in educational research, such as their high level of adaptability, the speed with which they can be administered, and their high degree of reproducibility. Consequently, they represent the most cost-effective way of collecting data from participants' responses, which can potentially be extrapolated to the general population (Bihu, 2021). Furthermore, to improve their adaptability to the current educational context, a focus group was formed with the aim of enhancing internal consistency. In this regard, it should be noted that the reliability of the questionnaire was calculated using Cronbach's alpha ($\alpha = 0.929$). However, this instrument also has certain disadvantages, such as the participants' possible tendency to agree with the items presented, which may lead to acquiescence bias (Danner et al., 2015).

The questionnaire used contains an introductory section which outlines the importance of the European Digital Competence Framework (DigComp) as well as the guidelines setting out the purpose of the research. This is followed by detailed instructions on how to answer the questions clearly and accurately, with additional information on the relevance of the study and its main objective, which is to obtain valid results to improve the management of digital resources based on Generative AI in higher education contexts. The instrument included both closed and open-ended questions to gather responses from university students studying education. To increase the reliability of the questionnaire items, pilot studies were conducted prior to the implementation of the final tool. During this process, several experts, such as primary school teachers and higher education lecturers, were asked to collaborate in order to adapt the instrument in line with their comments and suggestions. To measure the level of agreement amongst the collaborating teachers, the Kappa coefficient was applied ($K = 0.943$, $p < 0.001$). Having achieved a high degree of agreement amongst the collaborators, it was agreed to use the tool for this study at the start and end of the four-month term. A total of eight potential deepfakes were presented to all respondents, who were asked the same question focusing on their ability to distinguish between them and to explain the reasons for classifying the content as true or false.

2.3. Data analysis

This research has adopted a qualitative and interpretative approach, grounded in a constructivist framework, from which a series of behavioral traits are examined and described through the analysis of ideas and the interpretation of the thoughts expressed by the participants (Mohajan, 2018). As Ernest et al. (2023) point out, the qualitative approach facilitates a deeper exploration of respondents' perceptions and mental schemas, which can help to better understand the situation under study. After collecting the information provided by the participants, data analysis was carried out in order to comprehensively interpret the respondents' perceptions. To this end, a descriptive qualitative analysis was conducted to organize, summarize, and present the main ideas and patterns in the participants' responses.

Specifically, a thematic analysis of the responses was carried out by identifying codes that grouped together common reasons or justifications when reflecting on the veracity of the resources presented.

The dataset included 265 written responses because participants completed the task at two different stages, pre-test and post-test (n = 265). Given that this number of responses was manageable, the qualitative analysis was consistent with the approach proposed by Miles et al. (2018) and Olivera (2023), whereby the questionnaire data were organized into documents according to the themes of analysis, with quotations of textual information organized into codes following an inductive strategy, which would aid in understanding and explaining the main thematic points. To this end, the qualitative research software Atlas. Ti 23, which is widely used in research within the field of Social Sciences Education. To complement this software, ChatGPT was used as a supplementary tool during the organization and refinement of the coding process, but the final coding decisions and interpretation of the data were made by the authors, who reviewed and verified all AI-assisted suggestions, organizing the main conclusions drawn from the coding of the participants’ responses (Lopezosa & Codina, 2023). When a response could be assigned to more than one category, the most relevant code was selected according to the main justification provided by the participant. Table 1 presents an example of the thematic analysis.

Table 1
Example of the thematic data analysis process

Research stage	Identified theme	Main code identified	Students’ response
Pre-test	Identification of content as false	Structural defects	‘Fake. Based on the text’s format, it does not appear to come from a newspaper or a credible news outlet.’

Source: prepared by the authors.

2.4. Training program

The research took place during the first four-month term of the 2025/26 academic year as part of the core module ‘Social Sciences and their Teaching’ within the bachelor’s degree in Primary Education at the University of Murcia, Spain. The intervention program focused on enhancing university students’ social and digital skills in the field of social sciences, including content relating to history, geography, demography and ethics. The program was based on a previous one examined by Ayuso-del Puerto and Gutiérrez-Esteban (2022), who analyzed the development of AI-based educational projects carried out in educational contexts through collaborative and project-based work.

The aim of the program was to enrich university students’ intercultural understanding, focusing on the concepts of citizenship and civic education, based on which activities were designed in which students had to tackle social problems whose resolution required the development of their reasoning skills. In this regard, they were asked to engage in critical reflection on how to address certain current social challenges, in line with the subject matter covered in the sessions on an ongoing basis. These social

issues centered on sustainability and the defense of human rights in the humanitarian crises currently facing humanity, such as poverty, widespread wars in various parts of the world, or controversies linked to the decision-making of key political leaders, as well as the migratory movements associated with all these issues.

Efforts were also made to improve students' digital literacy in AI by utilizing various AI-generated resources to explore the subject matter from different perspectives. Consequently, during the 'Social Sciences and their Teaching' sessions, students engaged with digital content by working with primary sources, analyzing previous research and completing reflective online tasks. In this way, the students produced digital projects in teams and also engaged in cognitively demanding classroom activities to discuss relevant social issues, design teaching units addressing these topics, and draw conclusions on how to tackle them as future teachers. To this end, the students used a range of web-based platforms utilizing general-purpose AI to tackle complex challenges, which required them to carry out information-seeking tasks, manage the bibliography relating to the subject under study, and find scientific answers to the questions that arose during their work. Specifically, they examined texts using Chatpdf, analyzed information provided by chatbots such as Hello History, Perplexity and ChatGPT, and created infographics, mind maps and timelines using Canva, Venngage and Slidesgo. In this way, they were able to reflect on the potential of these digital resources to review subject content by organizing information, analyzing keywords and interacting with virtual assistants to resolve queries and expand on information.

They also discussed the ethical challenges associated with the use of general-purpose AI, which can undermine the trust of digital consumers who encounter deepfakes. As part of the regular training for undergraduate students, they were presented with a variety of digital content and asked to determine whether it was fake or not following a thorough and detailed examination. To this end, the students adopted a series of identification techniques based on those used by Farid (2022) and Wang et al. (2024), focusing on the detection of deepfakes primarily through the identification of visual and auditory inconsistencies, such as colour differences between individual frames, discrepancies in the sound patterns used, or the use of hyper-realistic fake images. They also looked out for other online manipulation techniques, such as artificial lip-syncing, evident in inconsistencies between the flow of the spoken dialogue and the anatomical movements of the subjects' mouths.

They also monitored the main media outlets to assess the reliability of the original sources, specifically, they checked the accounts of trolls and potential fake profiles that disseminate this malicious content and are often associated with various social media platforms and online communities. Consequently, identifying the sources proved crucial not only for determining the aims of these websites but also for strengthening the students' digital literacy in terms of implementing online protection measures and raising awareness of digital security against malicious campaigns. Once the planned tasks had been completed, the university students devised teaching sequences for the social sciences to learn how to design lessons aimed at identifying deepfakes at primary school level. All of this was carried out with the aim of approaching the subject matter in an innovative way, seeking to provide an effective response to the problems and challenges arising from the use of AI in today's classrooms (Martínez et al., 2023).

3 Results

In this section, we will present the findings identified following the analysis of the data. To do this, we will categorise the students’ responses by theme using codes established during the qualitative analysis process (see Table 2).

Table 2 Themes and codes identified, and the type of verification carried out by the participating students			
Research stage	Identified theme	Main codes identified	Type of verification
Pre-test	Identification of the content as false	Structural flaws	not externally validated
		Inconsistency of data	externally validated
		Language analysis	not externally validated
Pre-test	Identification of the content as false	Confirmation bias	not externally validated
		Attribution of authority	not externally validated
		Visual effects of format	not externally validated
Post-test	Identification of the content as false	Lip-syncing (anatomical and biological anomalies)	externally validated
		Visual logical inconsistency	not externally validated
		Lack of media credibility of the sources	externally validated
Post-test	Identification of the content as false	Analysis and verification of sources	externally validated
		Contextual analysis, absence of obvious errors	not externally validated
		Narrative coherence	not externally validated
Source: prepared by the authors			

R01: To explore the justifications put forward by participants for determining whether digital content is false, prior to the training program.

Structural flaws: Some students cited formal organisational aspects of the text to justify their assessment. Among the elements highlighted were typographical consistency and the overall

appearance of the text, which in some cases was perceived as resembling a document produced using basic editing tools or even generated by artificial intelligence. One student commented: “False. Because of the text’s format, it doesn’t look like it comes from a newspaper or a credible news source.” Another respondent stated: “False, because it does not look like a journalistic text template.” These formal aspects once again play a significant role in assessing credibility.

Inconsistency of data: There is a group of responses that suggest an explicit use of external information to question the consistency of the information presented and the accuracy of the digital content: “It is false because it does not provide clear sources or specific data; it merely mentions percentages and general opinions but does not present any real evidence or verifiable statistics”. Some students cite well-known statistics or data suggesting a decline in certain crime indicators or contradicting the claims made in the text: “False because there is data showing that, although there has been an increase in the number of immigrants in our country, the crime rate has fallen compared with previous years when there were fewer immigrants.” This type of argument suggests a greater willingness to cross-check the information against additional sources.

Language analysis: There is a subgroup of students who identify the use of language perceived as alarmist or exaggerated. These responses suggest that the text adopts an excessively negative tone towards immigration or that it may be designed to generate fear or hostility towards certain groups: “It is false because it uses language that a journalist would never use”. Another participant noted: “False; an increase in violence is not usually associated with immigration.” This recognition of a sensationalist tone used to exaggerate the potential negative effects of immigration suggests a greater ability to detect ulterior motives in the media’s coverage of the issue.

RO2: To explore the justifications put forward by participants for determining whether digital content is genuine, prior to the training program.

Confirmation bias: Furthermore, some responses appeared to rely on pre-existing assumptions or stereotypes, suggesting a possible confirmation-bias pattern. In particular, they mention aspects such as historical violence in the United States, the existence of racial tensions, and the widespread presence of firearms. In this regard, the news story is perceived as credible because it fits with previous narratives about the country’s historical and social context. Specifically, one student stated: “It’s true, because if immigrants aren’t given decent housing like everyone else, they get tired of looking for it and, when they aren’t given any, they get angry and that’s why they might cause violent situations on the streets.”

Attribution of authority: Another common criterion in responses supporting the veracity of the news item is the presence of references to institutional authorities. The mention of official positions, such as a police chief or other institutional leaders, is interpreted as a guarantee of credibility, even when students do not verify whether such statements actually exist or where the information comes from. In particular, one respondent commented: “I believe this news item is genuine because it includes a statement from a police chief.” Another student stated: “Genuine, as it includes real testimonies from prominent figures such as the police chief.”

Visual effects of format: A qualitative analysis of the justifications provided by those who consider the news item to be true reveals that the arguments used are based primarily on formal criteria rather than on historical verification of the information. Firstly, the importance attributed to the age of the medium stands out. Many students justify the veracity of the news item by pointing out that it comes from an old newspaper or a print publication. Expressions such as 'it's an old newspaper' or 'it's a paper newspaper' appear repeatedly, suggesting an automatic association between the age of the document and the credibility of its content.

RO3: To explore the justifications given by participants for determining whether digital content is false, following the training program.

Lip-syncing. Anatomical and biological anomalies: An identified code is based on the detection of irregularities associated with the synchronisation between the voice and lip movement. Expressions such as 'the mouth doesn't match the sound' reflect this rationale. Furthermore, they identified behaviours they consider unnatural or artificial in the people appearing in the video. Notable among the expressions used are statements relating to facial expressions or blinking, such as 'speaks like a robot', 'blinks strangely' or 'gestures strangely'. These kinds of observations correspond to indicators usually associated with deepfakes and other types of audiovisual manipulation.

Visual logical inconsistency: The participating students pointed out certain visual inconsistencies to justify identifying the digital content as fake. Specifically, they highlighted elements perceived as artificial or unrealistic, particularly in the scene's backgrounds or in the people in the background. In this regard, terms such as 'distortions' or 'AI-generated background images' were used. In particular, one participant stated the following: "Fake. The people in the image have blurred faces and there are elements that don't fit in with the rest of the picture." These responses show that the students applied visual analysis criteria to identify possible signs of digital manipulation.

Lack of media credibility of sources: A subgroup of students mentioned issues relating to the media credibility of the content. Some responses referred to the video's origin on platforms such as TikTok, the absence of verifiable sources, or the lack of coverage in traditional media. In this regard, one student remarked: "It's a lie because it's on TikTok and can spread information without being checked", whilst another participant noted: "It's a lie because they wouldn't agree on something like that, and besides, it would have been in the news rather than on TikTok". This type of argument suggests the emerging presence of verification criteria relating to the reliability of the source, although their use is still limited compared to technical and visual indicators.

RO4: To explore the justifications given by participants for determining whether digital content is genuine, following the training program.

Analysis and verification of sources: A very large subgroup of responses relates to the credibility attributed to the media outlet. Students justified the veracity of the content by referring to the institutional or verified nature of the channel, using expressions such as 'verified channel', 'official media',

‘news programme’ or ‘television programme’. There were also specific mentions of particular platforms or media outlets, as well as indicators of legitimacy such as the number of followers. This type of reasoning suggests that students externalise the criterion of truthfulness to the authority of the source, placing great weight on the media organisation’s institutional status as an indicator of reliability.

Contextual analysis, absence of obvious errors: Another line of argument is based on identifying the absence of features associated with content generated or manipulated by artificial intelligence. The responses highlight elements such as the absence of facial distortions, unusual gestures, visual distortions or mismatches between the voice and mouth movements. One student wrote: “Real, because there is nothing out of place and it could be a real situation.” Another participant noted: “It’s real, as there are no strange elements that might seem fake, nor are there any distortions.” This pattern indicates the application of technical criteria learnt during the teaching session, particularly those relating to the detection of anomalies characteristic of synthetic content.

Narrative coherence: Another group of students based their judgement on the plausibility of the information contained in the news item. In these cases, the students argued that the news item ‘makes sense’ and ‘is coherent’. There were also references to the objectivity of the digital content, describing it as ‘realistic’ and objective: ‘Real. I think it’s real because it provides objective information’. This type of reasoning suggest the application of social and contextual plausibility criteria, linked to the students’ prior knowledge of current affairs and the socio-political context.

With regard to the main objective of the research, the results reveal clear differences in the types of checks reported by the students before and after the training programme (see Table 2 and Table 3). According to the data presented, students made greater use of source-based and technical verification strategies, utilising digital tools to identify deepfakes, predominantly in the post-tests; this allowed for a more objective justification in the process of distinguishing these AI-generated resources. Equally, the interpretations based on their own intuition at the start of the term—such as those associated with confirmation bias—are noteworthy, raising doubts about the effectiveness of this technique in distinguishing deepfakes. Following the teaching intervention, a much more accurate identification of the visual and audiovisual indicators characteristic of AI-generated content was observed, such as anatomical distortions, inconsistencies in lighting, errors in lip-sync, artificial gestures or visually inconsistent backgrounds. This change is particularly noticeable in fake news based on images and videos, where participants move from uncertainty or a range of responses to a well-founded consensus.

Table 3
Qualitative comparison of pre-test and post-test responses

Item	Pre-test	Post-test
Use of technical criteria	Low	High
AI identification	Intuitive	Systematic
Trust in the source	High and uncritical	High and reasoned
Argumentation	General and contextual	Specific and analytical
Confidence in the response	Variable	High
Source: Prepared by the authors		

4 Discussion and conclusion

In this study, we have analyzed higher education students' perceptions of deepfakes and, specifically, their ability to identify them and the reasons they give for doing so, using a qualitative methodology. The relevance of this study is undeniable, as the acquisition of digital skills by future teachers is essential to enable them, in their future teaching practice, to help students analyze and evaluate a wide variety of complex information, which may sometimes prove to be false, particularly in highly complex matters, such as those relating to migratory movements (Smirnova & Permyakova, 2022). In this regard, the role of social sciences teachers as guides in students' learning processes involves not only presenting the content to be studied, but also training students on how to search for, select and scrutinize appropriate information from digital sources (Campillo-Ferrer et al., 2025).

With regard to the main objective of the research, which sought to identify differences in the techniques used by participants to detect deepfakes, the results suggested an increase in the use of techniques based on external identifiers following the completion of the training program. In this regard, the importance of equipping students with the necessary skills to identify deepfakes using appropriate digital resources is fundamental, as indicated by previous studies. For example, Tahir et al. (2021) implemented a training program focused on improving detection skills among people with lower levels of digital literacy, as they are more likely to fall victim to deepfakes. They concluded that contextualized training on how to distinguish this hyper-realistic content was highly useful and raised awareness of the issue. It should be noted that, at the start of the training program, students predominantly used unvalidated techniques such as assessing the contextual consistency of the information; this mostly led to errors, undermining the usefulness of this technique for determining the authenticity of deepfakes. For example, when some participants carried out preliminary tests and accepted the deepfake as genuine due to its plausibility within the current social context, they failed because they did not take into account other explanatory variables, such as the observation of digital inconsistencies and discrepancies in speech caused by mismatches between sound and image (Liu et al., 2024).

Furthermore, as we have seen in this study, confirmation bias poses a real risk, arising as a possible interpretation when determining whether a multimedia file is authentic or a deepfake. In fact, many respondents used this technique to confirm their preconceived ideas about immigration, based on deeply held beliefs; for example, when they claimed that digital content was a deepfake due to emerging social problems related to immigrants, or when they accepted that the content was real because it was old and, at that time, AI had not yet emerged. This finding is consistent with the conclusions drawn in recent literature, such as those of the study by Amin et al. (2025), which focused on the influence of confirmation bias, political ideology and social media when identifying these synthetic media and suggested that exposure to deepfakes can reinforce confirmation bias and affect political and ideological perspectives. This malicious intent can be exploited to promote political manipulation and carry out disinformation campaigns that harm disadvantaged social groups, such as immigrants. In this regard, it is essential to raise awareness of the risks associated with these AI-generated threats in order to prevent social manipulation that can lead to civic chaos and, ultimately, influence democratic processes such as elections and political campaigns.

Therefore, developing critical thinking skills amongst students can prevent these online risks whilst, at the same time, fostering civic engagement and responsible collective decision-making (Ibañez et al., 2020). Furthermore, more training should be included in higher education curricula to better prepare future teachers to identify emerging manipulation techniques. This training could consist of ready-to-use tutorial modules, enabling higher education students to follow step-by-step instructions to identify these synthetic media. In this regard, this conclusion is consistent with recent literature, which supports the integration and coordination of media literacy training to improve higher education students' deepfake detection skills (Wikum & Wijayanayake, 2024).

Among the limitations of this study, sample size is one of the most significant, as a larger number of participants from a wider range of higher education contexts would have increased the study's external validity. In this regard, the participation of university students from different year groups would also have improved representativeness, thereby strengthening the statistical power of the results. Furthermore, it is recommended that future research includes a control group to allow for comparison between the two groups of students and to draw more relevant conclusions. As mentioned above, the regular integration of training modules on deepfake detection into higher education would provide an opportunity to compare the results of pre-service teachers before and after the training to measure the effectiveness of the training provided. A further limitation concerns the interpretation of technical indicators of manipulation. Visual or audiovisual anomalies may raise suspicion, but they do not necessarily demonstrate that content was generated or manipulated by artificial intelligence, conversely, sophisticated deepfakes may not display easily recognizable anomalies. Future research should therefore combine the analysis of students' written justifications with objective accuracy measures, including known authentic materials, conventionally edited content and AI-generated deepfakes.

Declarations

Funding:

This research is result of the project 'Pensamiento histórico y cultura democrática: Herramientas educativas para luchar contra la desinformación' (PID2024-155318NB-C31), funded by MICIU/AEI/10.13039/501100011033/ FEDER, UE.

Conflict of interest: The authors declare no conflict of interest.

Ethics approval and consent to participate: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of Ethics Research Committee of the University of Murcia (protocol code 3113/2020 and approved on 3 August 2021). Informed consent was obtained from all subjects involved in the study.

Consent for publication. All authors have read and agreed for publication the version of the manuscript.

Data, materials and code availability: The research data will be provided by the authors upon request via email.

Author Contribution

Conceptualization, P.M.S and J.M.C.F.; methodology, P.M.S. and J.M.C.F.; software, P.M.S.; validation, P.M.S. and J.M.C.F.; formal analysis, P.M.S.; investigation, J.M.C.F.; resources, J.M.C.F.; data curation, P.M.S.; writing—original draft preparation, P.M.S. and J.M.C.F.; writing—review and editing, P.M.S.; visualization, P.M.S.; supervision, J.M.C.F.; project administration, P.M.S. and J.M.C.F.; funding acquisition, P.M.S. and J.M.C.F. All authors reviewed the manuscript.

Data Availability

The research data will be provided by the authors upon request via email.

References

1. Almisad B, Aleidan A. Faculty perspectives on generative artificial intelligence (GenAI): insights into awareness, benefits, concerns, and uses. *Front Educ.* 2025;10:1632742. <https://doi.org/10.3389/feduc.2025.1632742>.
2. Atkinson CF. ChatGPT and computational-based research: benefits, drawbacks, and machine learning applications. *Discover Artif Intell.* 2023;3(1):42. <https://doi.org/10.1007/s44163-023-00091-3>.
3. Ayuso-del Puerto Desirée, Gutiérrez-Esteban P. La Inteligencia Artificial como recurso educativo durante la formación inicial del profesorado. *RIED-Revista Iberoamericana de Educación Distancia.* 2022;25(2):347–58. <https://doi.org/10.5944/ried.25.2.32332>.

4. Baidoo-Anu D, Ansah LO. Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *J AI*. 2023;7(1):52–62. <https://doi.org/10.61969/jai.1337500>.
5. Bihu R. (2021) Questionnaire Survey Methodology in Educational and Social Science Studies. *International Journal of Quantitative and Qualitative Research Methods*, 9, No.3, pp 40–60. Available at SSRN: <https://ssrn.com/abstract=4435809>
6. Campillo-Ferrer José, María, López-García A, Miralles-Sánchez P. Student Perceptions of the Use of Gen-AI in a Higher Education Program in Spain. *Digital*. 2025;5(3):29. <https://doi.org/10.3390/digital5030029>.
7. Danner D, Aichholzer J, Rammstedt B. Acquiescence in personality questionnaires: Relevance, domain specificity, and stability. *J Res Pers*. 2015;57:119–30. <https://doi.org/10.1016/j.jrp.2015.05.004>.
8. Ebner Jörg, Lutz B, Neumann D. Deep Fakes and Political Attitudes: An Analysis of Confirmation Bias and Cognitive Dissonance Using Neurophysiological Measurements. *NeuroIS Retreat*. Cham: Springer; 2024. pp. 299–311. https://doi.org/10.1007/978-3-031-71385-9_26. Nature Switzerland.
9. Farid H. Creating, using, misusing, and detecting deep fakes. *J Online Trust Saf*. 2022;1(4). <https://doi.org/10.54501/jots.v1i4.56>.
10. Håkansson A, Phillips-Wren G. Generative AI and Large Language Models-Benefits, Drawbacks, Future and Recommendations. *Procedia Comput Sci*. 2024;246:5458–68. <https://doi.org/10.1016/j.procs.2024.09.689>.
11. Hameleers M, van der Meer TG, Dobber T. You won't believe what they just said! The effects of political deepfakes embedded as vox populi on social media. *Social Media+ Soc*. 2022;8(3):20563051221116346.
12. Howarth A, Cooke SJ, Nguyen VM, Hunt LM. Non- probabilistic surveys and sampling in the human dimensions of fisheries. *Rev Fish Biol Fish*. 2024;34(2):597–622. <https://doi.org/10.1007/s11160-023-09831-1>.
13. Kartheeban K, Aanandaram V, Lakshminarasimman D, Kabalishwaran D. (2025). The Potential and Drawbacks of Machine Learning for Student Assessment. En Murugan, Thangavel, editor *Adopting Artificial Intelligence Tools in Higher Education* (pp 62–98). CRC Press.
14. Liu W, She, Tianyi, Liu, Jiawei, Li B, Yao, Dongyu, Liang Z, Wang R. Lips are lying: Spotting the temporal inconsistency between audio and visual in lip-syncing deepfakes. *Adv Neural Inf Process Syst*. 2024;37:91131–55. <https://doi.org/10.52202/079017-2892>.
15. Lochmiller CR. Conducting thematic analysis with qualitative data. *qualitative Rep*. 2021;26(6):2029–44. <https://doi.org/10.46743/2160-3715/2021.5008>.
16. Lopezosa C, Codina Lluís. (2023). ChatGPT y software CAQDAS para el análisis cualitativo de entrevistas. Pasos para combinar la inteligencia artificial de OpenAI con ATLAS.ti, Nvivo y MAXQDA.
17. Miao F, Holmes W. Guidance for generative AI in education and research. UNESCO Publishing. 2023. <https://doi.org/10.54675/EWZM9535>.

18. Miralles-Martínez. Pedro, Campillo-Ferrer, José María & Prats-Cuevas, Joaquim (2023). La enseñanza y el aprendizaje de las ciencias sociales en tiempos de incertidumbre. *Areas*, 45, 5–9. <https://doi.org/10.6018/areas.598491>
19. Michałkiewicz-Kądziela E. The Impact of Deepfakes on Elections and Methods of Combating Disinformation in the Virtual World. *Teka Komisji Prawniczej PAN Oddział w Lublinie*. 2024;17(1):151–61. <https://doi.org/10.32084/tkp.8615>.
20. Miles MB, Huberman MA, Saldana J. *Qualitative Data Analysis: A Methods Sourcebook*. 4th ed. SAGE; 2018.
21. Mohajan HK. Qualitative research methodology in social sciences and related subjects. *J economic Dev Environ people*. 2018;7(1):23–48.
22. Negou E, Fonkem M, Nkenganyi, Abenwi JS, Ibrahima. Qualitative research methodology in social sciences. *Int J Sci Res Manage (IJSRM)*. 2023;11(09):1431–45. <https://doi.org/10.18535/ijssrm/v11i09.sh01>.
23. Ninovic V. Deepfake crime: Trends, threats and implications. *Int J Contemp Intell Issues*. 2024;1(2):41–55. 10.3316/informit.T2024061400011102001728699. <https://search.informit.org/doi/>.
24. Oliveira G. *Int J Res Method Educ*. 2023;46(3):300–12. <https://doi.org/10.1080/1743727X.2022.2128745>. Developing a codebook for qualitative data analysis: Insights from a study on learning transfer between university and the workplace.
25. Overton S. Overcoming racial harms to democracy from artificial intelligence. *Iowa L Rev*. 2024;110:805.
26. Pearson GS. Artificial intelligence and publication ethics. *J Am Psychiatr Nurses Assoc*. 2024;30(3):453–5. <https://doi.org/10.1177/10783903241245423>.
27. Sánchez-Ibáñez R. Campillo-Ferrer, José María & Guerrero-Romera, Catalina (2020). Percepciones del profesorado de primaria y secundaria sobre la enseñanza de la historia. *Revista interuniversitaria de formación del profesorado*, 34(3), 57–76. <https://doi.org/10.47553/rifop.v34i3.83247>
28. Smirnova E, Permyakova T. Terminology of Migration Studies: A Corpus Analysis of Research Papers in Social Sciences. *Migration Lett*. 2022;19(4):401–12. <https://migrationletters.com/index.php/ml/article/view/1799>.
29. Smith S, Michelle B, Lane M. (2026). A university framework for the responsible use of generative AI in research. *Journal of Higher Education Policy and Management*, 48(1), 17–36. <https://doi.org/10.1080/1360080X.2025.2509187>
30. Twomey J, Ching, Didier A, Peter M, Quayle M, Linehan C, Murphy G. Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine. *PLoS ONE*. 2023;18(10):e0291668. <https://doi.org/10.1371/journal.pone.0291668>.
31. Wang T, Liao, Xin C, Pui K, Lin X, Wang Y. Deepfake detection: A comprehensive survey from the reliability perspective. *ACM-CSUR*. 2024;57(3):1–35. <https://doi.org/10.1145/3699710>.

32. Wikum T, Wijayanayake J. (2024). A systematic literature review of deepfake literacy, societal impacts, and AI-driven prevention strategies for platform-specific vulnerabilities. 2024 International Conference on Advances in Technology and Computing (ICATC) (pp 1–6). <https://doi.org/10.1109/ICATC64549.2024.11025231>
33. Yan W, Nakajima T, Sawada R. Benefits and challenges of collaboration between students and conversational generative artificial intelligence in programming learning: An empirical case study. *Educ Sci.* 2024;14(4):433. <https://doi.org/10.3390/educsci14040433>.