

Seeing is no longer believing: a study of Deep Fake identification ability and its social impact

Yijun Xie

Department of Security and Crime Science, University College London, London, UK

leoooooxie@outlook.com

Abstract. This research explores university students' ability to identify deepfake images generated by artificial intelligence, and their criminological implications when faced with manipulated content. Through an online questionnaire with 129 participants, the study designed deepfake identification tasks and attitude scales combining demographic, cultural, and experiential factors. The results indicate an overall deepfake identification accuracy of 55.8%, consistent with previous meta-analyses. Regression analysis showed that the education level and AI tool usage experience positively predicted performance, while age and gender had no significant effect. Students who received prior deepfake training had significantly improved identification accuracy; cultural background failed to reach statistical significance. More importantly, exposure to deepfakes reduced trust in digital images and self-judgement confidence. These findings indicate that despite university students' high reliance on digital platforms, they are not naturally more resilient to deepfakes and remain susceptible to crimes such as fraud, identity theft, and forged evidence. Overall, this paper demonstrates that deepfakes present novel victimisation risks within criminology while undermining the credibility of digital evidence, highlighting the importance of advancing relevant education and prevention strategies within higher education and criminal sciences.

Keywords: deepfake detection, artificial intelligence, media credibility, user perception, AI-generated content

1. Introduction

In the digital era, Artificial Intelligence (AI) has become a leading force in technological change, reorganising how people create, communicate, and understand information. Among the many applications of AI, technologies with content generation capabilities, especially generative technologies that can produce highly realistic images, videos, and text, are becoming popular. These technologies are slowly blurring the line between real and fake in the digital medium.

Among various forms of synthetic media, deepfakes—which use artificial intelligence and machine learning technologies to generate realistic images and videos of people's appearances, speech, or behaviour—are particularly controversial. Therefore, the ability to detect deepfakes has become a key element in strategies to prevent victimisation and reduce opportunities for crime.

University students are highly exposed and at high risk in the digital environment. They rely heavily on online platforms for communication, learning, and entertainment in their daily lives [1]. However, frequent access does not necessarily mean strong identification skills [2]. Identifying real and AI-generated content requires a mix of media literacy, critical thinking, and context awareness, and these skills are not uniformly distributed among university students [3].

Although existing studies have focused on deepfake detection technologies and legal responses, there has been a lack of systematic discussion on users' actual ability to identify such content in their daily lives. This gap is critical because the ability (or inability) to identify media manipulation directly affects the risk of being misled, spreading false information, or becoming a target of fraud or other crimes facilitated by deepfake technology. In addition, existing literature often overlooks the role of social and cultural factors (such as cultural background, digital literacy levels, and prior exposure to AI-related education) in shaping individuals' susceptibility to deepfake deception.

Therefore, this study focuses on the ability of university students to identify AI-generated deepfake images, aiming to identify the factors that influence their judgements and to examine whether exposure to such content affects their trust in digital information.

2. Literature review

2.1. Background and definitions

2.1.1. The definition of AI, and the differences between generative AI and traditional AI

Artificial Intelligence (AI) is commonly defined as the study and design of agents that receive percepts from their environment and perform actions to maximise their chances of achieving goals [4]. Although AI has had numerous definitions throughout history, scholars such as Poole and Mackworth, and Nilsson generally agree that AI should be defined as agents capable of intelligent actions that adapt to diverse environments [5, 6]. According to the CSRC Content Editor, generative artificial intelligence is defined as an artificial intelligence model that creates derivative synthetic content by imitating the structure and properties of incoming data [7]. A typical instance of generative AI content creation capabilities is deepfake technology, which uses algorithms to generate synthetic but highly realistic images and videos.

2.1.2. What are deepfakes and their evolution

A deepfake is generally defined as an image, video or an audio recording, that has been edited using an algorithm to replace the person in the original with someone else. Deepfakes came into the public view in 2017, when users uploaded videos and images of famous people with their faces replaced using deepfake technology on Reddit [8]. Since then, more and more people have posted deepfake-related content online [9]. The development of deepfake technology has rapidly lowered barriers of technical skill, computing resources, and production cost, making it accessible even to non-experts through publicly available AI models and open platforms [10].

Zendran and Rusiecki summarised four commonly used methods for generating deepfake content [11]. At the same time, they reveal the ethical and social risks associated with such creations, evolving from an initial entertainment tool into a new source of risk for information manipulation and trust crises [12].

Additionally, one of the reasons for the rapid popularity of deepfake technology is the easy accessibility of its tools and the simplicity of its operation. In the past, deep learning algorithms were limited to researchers and advanced developers, but today, with the help of open-source software and visual interfaces, ordinary users can easily generate highly realistic images or videos. For example, open-source frameworks such as StarGAN-V2, AttGAN, and StyleGAN have achieved high accuracy while remaining publicly available [13]. The ease of use of these tools has significantly lowered the technical barrier, accelerating the spread of deepfake content creation among non-professional groups. Many social media platforms and mobile applications, such as TikTok and Snapchat, even have built-in features such as face replacement. This trend towards widespread accessibility and mainstream use of synthetic deepfake content has significantly lowered the threshold for creating false information and poses a challenge to traditional public opinion monitoring measures. Because this technology, which was once limited to professional developers, is now available to general users and has become popular on multiple platforms, it is becoming increasingly difficult for fact-checkers and automated detection systems to monitor and respond in real time [14].

2.2. The relationship between types of false information and deepfakes

2.2.1. Misinformation, disinformation, and fake news

Deepfakes do not exist in isolation; they often act as vehicles for different types of false information, including misinformation, disinformation, and fake news.

In this modern age of advanced technology, the internet has become an essential part of people's lives. Almost everybody with an internet connection can share information, leading to diverse channels and a possibility to present opposing viewpoints on a particular problem [15]. That also makes it easier for false information to spread, influencing the trust that people have in digital media [16].

Misinformation is the unintentional spreading of false information [17]. Disseminators are often unaware that the content they are spreading is false, and their motives are not harmful. Disinformation, on the other hand, is more manipulative and is defined as the purposeful creation and dissemination of false information [17]. Generally, people or organisations that disseminate such information with specific motivations, and they are aware that the information is incorrect.

Fake news can be formally defined as the deliberate presentation of false or misleading claims as news, where the claims are misleading by design [18]. Fake news normally appears in news format, using headlines, journalistic language, and media styles to package false content. Its creators know that the content is false, but they spread it to get traffic or to deceive people [19]. Therefore, fake news can also be regarded as a special type of disinformation, packaged in a more formal way, which is easier to gain trust.

2.2.2. How deepfake exacerbates information manipulation

The combination of the above three types of false information with generative artificial intelligence further accelerates the spread of false information and amplifies its effects. Due to the powerful capabilities of deepfake technology in creating videos and images, Appel and Prietzel note that it can be used to manipulate political discourse [20]. Similarly, Sandoval et al. found that deepfakes can be used to fabricate criminal evidence [21].

As misinformation, deepfakes are often spread unintentionally. Regular users may repost a synthetic video or image of unknown origin on social media without realising that the content is an AI-driven deepfake [22]. For example, a well-known case is the deepfake video of former US President Barack Obama produced by the news media company BuzzFeed. The video uses AI technology to combine actor Jordan Peele's voice with Obama's facial imagery, making it appear as though Obama is speaking. In the video, "Obama" insults President Trump, despite the fact that he never actually said those words [23].

For disinformation, deepfake content is more operational and targeted in spreading deceptive information [24]. Ranka et al. found that viewers may already form biases when watching content for the first time, even if it's later revealed to be fake [25]. This weakened trust extends beyond suspicions about social media or news sources, also affecting the credibility of digital evidence and the reliability of information in criminal investigations.

Additionally, the use of deepfakes has shifted fake news from purely text-based formats to images and dynamic videos [23]. A piece of fake news that relies solely on fictional text may not attract much attention, but when supplemented with deepfake images and the authority of the news itself, it is more likely to inspire empathy in readers [26].

2.3. Current challenges and the vulnerability of university students

2.3.1. The authenticity and social impact of generative AI content

With recent developments in generative modelling, the realism of deepfake material has steadily increased to the point where individuals frequently fail to spot influenced media content online, leading to numerous types of fraud [27]. Diel et al. conducted controlled experiments in which participants were asked to classify a series of AI-generated and real images. The study found that the average accuracy was only about 55%, suggesting that most individuals relied heavily on intuition rather than identifiable visual cues when making such judgments [28].

This highly realistic generative power has brought unexpected social challenges. First, the large-scale dissemination of deepfake content on social media platforms is weakening people's trust in the traditional belief that "seeing is believin" [29]. When people begin to doubt the authenticity of all images and videos, their trust in social media news will decline [23], and this distrust will evolve into a state of indifference to information, where they doubt not only manipulated content but also authentic sources. This breakdown of trust has wider implications, as it can breed cynicism [30] and make individuals more susceptible to alternative narratives or misinformation.

Deepfakes themselves cannot commit fraud. However, criminals are using technological advances to increase the credibility of their deception. Studies by Groh et al. and Hameleers et al. indicate that in sensitive areas such as political communication, racial conflict, and gender-based attacks, the influence of fake videos and images far exceeds that of text-based false information [31, 32]. Even after the truth is revealed, individuals' confidence in their ability to identify media will drop significantly [33].

2.3.2. The vulnerability of university students in identifying AI-generated content

University students are particularly dependent on the internet and social media [34]. Although studies have highlighted the risks associated with high exposure combined with low identification skills, scholars tend to view university students simply as a vulnerable group. This single perspective overlooks differences in digital literacy, prior media training, and cultural backgrounds, which may significantly influence their susceptibility to AI-generated deepfake content.

Foremost, although university students generally believe that they possess excellent information and media literacy, their understanding of technologies such as deepfakes and AI is actually at a lower level [35]. Research by Krupp et al. indicates that many physics students accept ChatGPT's answers without critical thinking [36]. However, this study is limited to a single subject area, so it remains to be explored whether similar overconfidence is also prevalent, especially when judging AI-generated false information.

Many university students have not received relevant deepfake education before entering university. Therefore, when content appears to be genuine, they tend to believe it is real [37], especially when it is packaged with symbols such as news formats, subtitles, and watermarks. Most students tend to judge the credibility of fake news online based on the content of the images [38]. Furthermore, many university students are willing to share false information on social media regardless of its accuracy or credibility, based simply on their interests [39].

Deepfakes enable the creation of false content depicting students in damaging or compromising situations they never participated in [40]. One of the most well-known examples is that of a law student at the University of Hong Kong who used AI

to generate more than 700 pornographic deepfake images of classmates and teachers without their consent [41]. This shows that students are easily victimised by such abuse, highlighting the need for universities to establish protective measures.

In summary, all of the above factors make university students a high-risk group for misleading information and potential direct victims of deepfakes.

2.4. Research gaps, questions, and hypotheses

2.4.1. Research gaps

Current research on deepfakes has mostly focused on technical detection methods, legal and ethical issues, or cases of political manipulation. Although a large number of studies, such as the work of Vaccari and Chadwick, have pointed out that deepfakes create challenges for social trust, there is still a lack of systematic studies on their specific impact on individual behaviour, especially among university students [23].

Moreover, existing research has mainly focused on samples in a specific country or region, missing exploration of cross-cultural differences and leaving significant gaps in the research.

Additionally, although public awareness of deepfakes has increased, there is currently a lack of experimental data on whether the exposure of deepfakes affects users' overall trust in digital information and what factors may mitigate or enhance this change in trust.

2.4.2. Research questions

Based on the above research background and literature review, this study aims to address the following research questions:

1. How do demographic factors and cultural background influence the ability of university students to identify AI-generated deepfake images?
2. How does exposure to AI-generated deepfake content affect university students' social trust and confidence in digital information?

2.4.3. Research hypotheses

H1: University students will show accuracy close to the average level reported in previous studies [28] (around 50-60%) in identifying AI-generated deepfake images from real images.

H2: University students who have received professional training in deepfake-related knowledge will be significantly more accurate in identifying AI-generated deepfake images than university students who have not received such training.

H3: Cultural background will significantly affect the accuracy of university students in identifying deepfake images.

H4: Exposure to deepfake content generated by artificial intelligence will influence students' trust in digital information and media sources.

3. Methodology

3.1. Participants

This study employed convenience sampling, with current university students as the primary recruitment target, which was highly relevant to the objectives of this study. Participants were recruited through two channels: firstly, questionnaire invitations were sent to current students via university mailing lists; secondly, questionnaire links were posted on social media platforms (including WeChat student groups, Facebook student groups, and other student networks). Participation was voluntary and anonymous. Data collection was conducted using Qualtrics survey software, with duplicate detection enabled to prevent multiple submissions by the same participant.

A total of 131 questionnaires were collected. The researcher set the survey settings in Qualtrics so that incomplete responses were automatically deleted if they remained unfinished for more than four hours. In addition, two responses that were completed in less than one minute were excluded on the grounds of insufficient response quality. Consequently, 129 valid responses were retained for analysis. All participants were currently studying in higher education, including both younger students and older students who had chosen to return to school to continue their studies. Students' educational backgrounds included computer science, education, mathematics, and other fields.

3.2. Data collection

3.2.1. Data sources

The deepfake images in part 2 of the questionnaire were sourced from an open-source project called FaceForensics [42]. The author of this dissertation submitted a request and obtained permission to use the deepfake dataset. Six pairs of images were selected to provide sufficient variation to support effective analysis while controlling the length of the task to avoid participant fatigue. Figure 1 shows an example of one of the identification tasks.

Q23. Please select the image you believe is authentic (not a deepfake).



Figure 1. An example of the image pairs used in the identification task

3.2.2. Survey design

The collection of data for this study was based on a questionnaire survey. The questionnaire was structured in three parts: 1. Background Information and Initial Attitudes, 2. Deepfake Image Identification Task, and 3. Post-Task Reflection. The purpose of questionnaire questions designed is detailed in Table 1.

Table 1. Questionnaire and aims

Part	Question Type	Related RQ/Hypothesis
Part 1	Age, Gender	RQ1
	Birth country, nationality, longest residence	RQ1, H3
	Education level	RQ1
	AI tools usage; Deepfake exposure	RQ1
	Concern about AI-generated content; Trust in online images; Confidence in identifying deepfakes	RQ2, H4
	Previous deepfake training	RQ1, H2
Part 2	Deepfake Image Identification Task	Provides accuracy scores for all analysis
Part 3	Post-task: deepfake confidence	RQ2, H4
	Post-task: media trust	
	Post-task: AI concern	
Open-Ended Question	Trusted source for verification; Policy and regulation attitude	Supplementary insights

3.3. Data analysis

After collecting the data, the dataset was imported into R for analysis. Before statistical modelling, general data cleaning procedures were carried out, including converting all responses to capitals, removing extra spaces, and checking for missing values. Following data cleansing, the variables were defined and coded.

3.3.1. Variables

The model in this study used deepfake identification accuracy as the dependent variable. This was measured by participants' performance on an identification task over six groups of images. Each correct answer was scored as 2 points, while incorrect answers received 0 points. The total score ranged from 0 to 12, representing a continuous variable which was further converted into accuracy rates (0-100%).

- Independent variables were taken from demographic information, cultural background, AI-related experience, prior education, and Likert scales measuring confidence and attitudes.
- Age is an ordinal variable, numerically coded according to the following grouping scheme: 18–21 years = 0, 22–25 years = 1, 26–30 years = 2, 31–35 years = 3, 36–40 years = 4, 41–45 years = 5, 45 years and above = 6.
- Gender is a nominal variable, coded as follows: Male = 0, Female = 1, Other = 2.
- Educational background is an ordinal variable, coded as follows: High school = 0, Undergraduate degree = 1, Master's degree = 2, Doctorate = 3, and Other = 4.
- Cultural background variables included country of birth, nationality, and country of longest residence. These variables are nominal, and were effect-coded in the regression analysis.
- The usage of AI tools and whether relevant training has been received are both binary variables, coded as follows: Yes = 1, No = 0.

Finally, the two variables, trust in online images and videos and confidence in one's ability to identify deepfakes, are both ordinal variables measured using a five-point Likert scale. To quantify them, we use 1 to denote the lowest level (not at all trust, not at all confident) and 5 to denote the highest level.

3.3.2. Statistical analysis

First, descriptive statistics were calculated to summarise participants' demographic characteristics, cultural backgrounds, and overall performance in the deepfake identification task (accuracy rates). Two multiple linear regression models were built with overall identification accuracy as the dependent variable: Model 1 included age, gender, educational background, and AI tool usage experience as independent variables; Model 2 used country of birth, nationality, and country of longest residence as independent variables. To evaluate the attitude changes, a paired-sample t-test was applied to compare participants' shifts in trust towards online media and confidence in identifying deepfakes. Additionally, an independent sample t-test was applied to compare differences in identification accuracy between participants who received deepfake-related training and those who did not.

3.4. Ethics

This study was approved by the UCL Departmental Ethics Committee, Project ID: 1025. Participants were notified of the information sheet and had to answer the consent question before starting the questionnaire. No personally identifiable information was collected during the study, and all responses were anonymised. Data was stored securely and confidentially, accessible only to researchers. Participants were informed that they had the right to withdraw from the study at any time without repercussions.

4. Findings

4.1. General results

This research yielded a total of 129 valid questionnaires. Among these, 67 were female (51.9%), 59 were male (45.7%), and 3 were non-binary (2.4%). Figure 2 shows the age distribution.

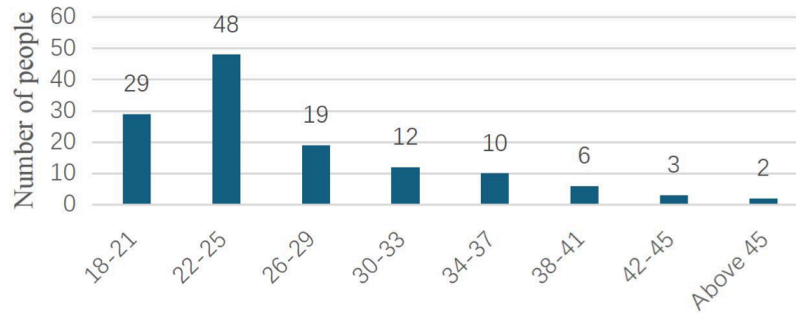


Figure 2. Age distribution of participants

53 participants were from science and engineering disciplines, 41 from social sciences, and 35 from humanities subjects. In terms of current educational status, 57 were undergraduates, 59 were master's students, 8 were doctoral students, and 5 chose "other".

In terms of cultural background, this research employed three independent indicators: country of birth, nationality, and country of longest residence. As shown in Figure 3, the majority of participants resided primarily in China and the United Kingdom, which together accounted for over three-quarters of the total sample.

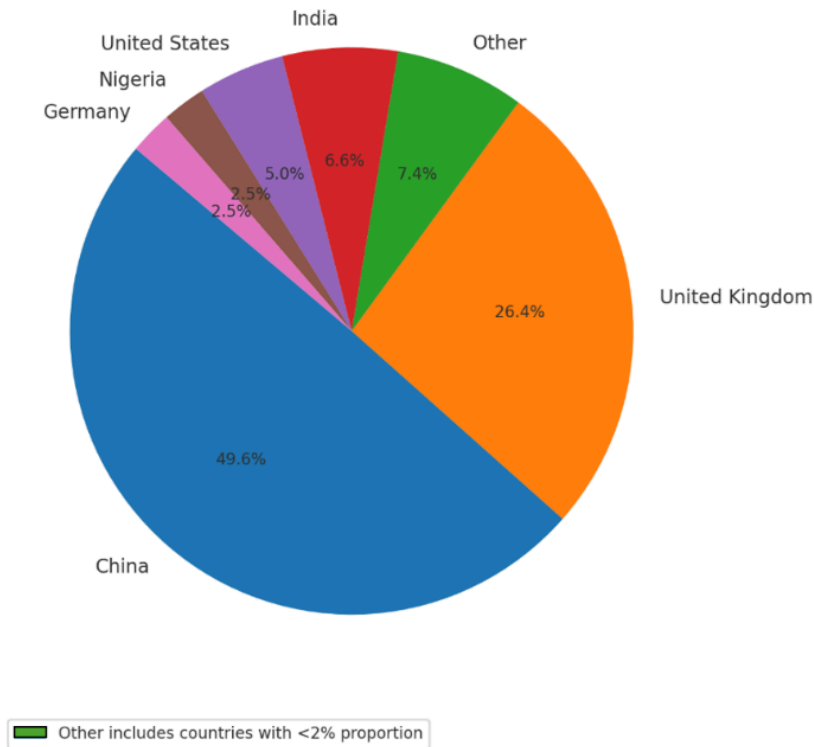


Figure 3. Country where participants spent most of their lives

4.2. Deepfake identification performance

The average score of participants was 6.70 (out of 12, standard deviation = 2.65), corresponding to an overall accuracy rate of 55.8%. There was significant variation in individual performance, with scores ranging from 0 to 12. Among these, eight participants achieved full scores (12/12), while one participant failed to answer any question correctly (0/12). This variation demonstrates the high heterogeneity of identification ability within the sample. Table 2 shows the average scores and accuracy rates for each task. The accuracy rates for the first four tasks were all above 60%, while the accuracy rate for task 5 was close to random (50.1%). The accuracy rate for task 6 was only 27.1%, making it the most difficult question.

Table 2. Average scores and accuracy rates for each deepfake identification task

	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6
Average score	1.21	1.30	1.30	1.33	1.01	0.54
Accuracy	60.5%	65.1%	65.1%	66.7%	50.1%	27.1%

Furthermore, this study analysed participants' confidence levels when identifying deepfakes. Confidence was measured using a five-point Likert scale, where 1 indicated "no confidence at all" and 5 indicated "extremely confident". Table 3 shows the average confidence scores for correct and incorrect answers.

Table 3. Accuracy rates and confidence ratings for correct vs. incorrect responses

Group No.	Accuracy	Average confidence (correct)	Average confidence (incorrect)
1	60.5%	4.49	3.69
2	65.1%	4.59	4.26
3	65.1%	4.51	4.09
4	66.7%	4.63	4.28
5	50.1%	3.97	4.43
6	27.1%	4.36	4.34

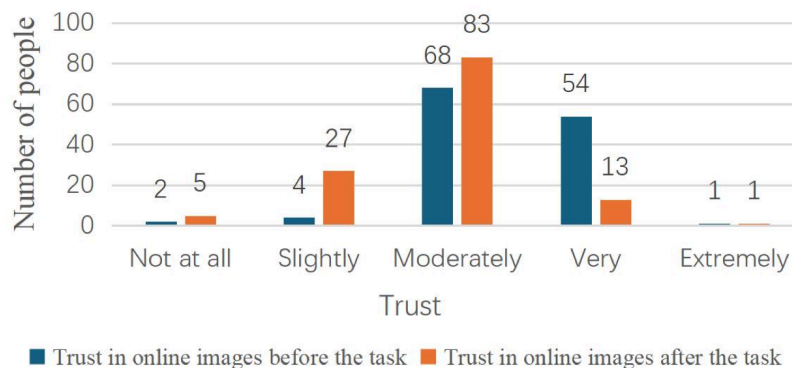
In groups 1-4, the average confidence score was generally higher when the answer was correct than when it was incorrect. However, in group 5, those who answered incorrectly had higher confidence, and in group 6, there was almost no difference in confidence levels between correct and incorrect answers. This indicates that participants' confidence levels do not always correspond to their actual performance.

When comparing participants who had received education or training related to deepfakes, task performance exhibited certain differences. The trained group achieved an average score of 7.73 (64.4% accuracy), whereas the untrained group scored an average of 6.28 (52.3% accuracy). Overall, participants who had undergone training scored higher than their untrained counterparts in the identification task.

4.3. Attitude and awareness

All attitude questions in this research used a five-point scale, where 1 represented the lowest level (e.g. "not confident at all" or "not worried at all") and 5 represented the highest level (e.g. "extremely confident" or "extremely worried"). Participants' awareness of AI-generated content, trust in online images, and confidence in their ability to identify deepfakes all changed before and after the test.

In terms of trust in the authenticity of online images, the average score before the task (Q15) was 3.37 (standard deviation = 0.64), which is considered a medium level; after the task (Q25), it dropped to 2.83 (standard deviation = 0.69), indicating that participants had more doubts about the authenticity of digital media after the task. For the detailed distribution, please refer to Figure 4.

**Figure 4.** Trust in online images before and after the task

Regarding confidence in identifying deepfakes, the average score before the task (Q16) was 4.36 (standard deviation = 0.66), with most participants selecting "very confident" or "extremely confident". However, after the task (Q24), the average score dropped to 3.78 (standard deviation = 1.00). Figure 5 illustrates the confidence in identifying deepfakes before and after the task.

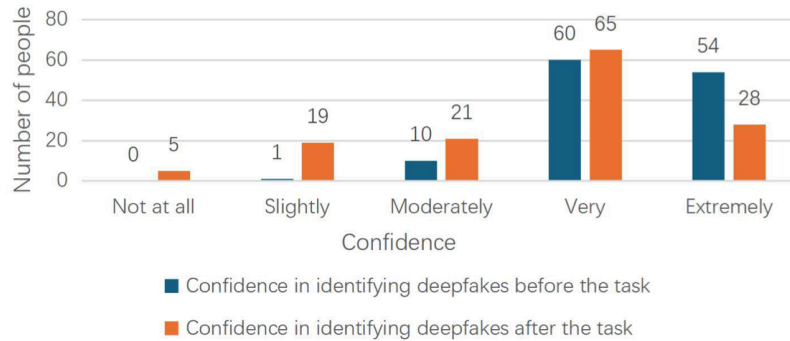


Figure 5. Confidence in identifying deepfakes before and after the tasks

Regarding concerns about AI-generated content being used for the dissemination of misleading information or fraud (such as scams or impersonation), participants scored an average of 3.29 (standard deviation = 0.72) before the task, indicating moderate concern overall. After completing the identification task, 66 participants (51.2%) explicitly stated they were "more concerned", 59 participants (45.7%) indicated they were "unsure", and only 4 participants (3.1%) stated they were "not concerned".

4.4. Regression analysis

The regression results (see Table 4.) showed that the model was statistically significant in general, but its explanatory power was relatively limited ($R^2 = 0.088$). Specifically, educational level (estimated = 0.71, $p = 0.046$) and experience using AI tools (estimated = 1.00, $p = 0.043$) had a significant positive predictive effect on identification performance. In contrast, age (estimated = -0.23, $p = 0.127$) and gender (estimated = -0.53, $p = 0.224$) did not show statistically significant effects.

Table 4. Linear regression model 1 (age, gender, education, AI tool usage)

Variable	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5.978	0.602	9.92	< 0.001
Age	-0.228	0.148	-1.54	0.127
Gender	-0.530	0.434	-1.22	0.224
Education	0.706	0.350	2.01	0.046
AI tool usage	0.997	0.487	2.05	0.043

This model explained approximately 29.6% of the variance in scores ($R^2 = 0.296$), but the adjusted R^2 was low (0.099), indicating its explanatory power was limited. The overall model approached statistical significance ($p = 0.0738$). Participants born in Norway (estimate = 1.639, $p = 0.044$) scored significantly above the mean, indicating that a Norwegian birth background correlates with stronger identification ability. For details of the regression, see Table 5.

Table 5. Linear regression model 2

Variable	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.000	1.753	3.42	0.001
Birthplace: Norway	+1.639	0.8037	2.040	0.044
Birthplace: India	-1.04	0.580	-1.786	0.077
Nationality: Japan	+0.963	0.581	1.660	0.100
Nationality: UK	-1.3091	0.8171	-1.6021	0.1121
Longest-lived country: Norway	+1.548	0.8267	1.989	0.049
Longest-lived country: Nigeria	-0.8964	0.5231	-1.7137	0.0895

4.5. T-tests

To further examine changes in participants' trust in online media and confidence in identifying deepfakes before and after the task, this study conducted a paired samples t-test on Q15/Q25 (trust) and Q16/Q24 (confidence).

The results (see Table 6.) of the sample t-test indicate that participants' level of trust prior to the task ($M = 3.37$, $n = 129$) was significantly higher than that after the task ($M = 2.82$, $n = 129$), with the difference reaching statistical significance ($t(128) = 7.84$, $p < 0.001$). Moreover, participants' confidence in their ability to identify deepfakes averaged 4.35 ($n = 129$) prior to the task and 3.80 ($n = 129$) afterwards, with the difference again being significant: $t(128) = 7.36$, $p < 0.001$.

Table 6. Sample t-test: trust and self-confidence (pre-task and post-task)

Variable	Average score (before task)	Average score (after task)	<i>t</i> -value	<i>p</i> -value
Trust	3.37	2.82	7.84	< 0.001 ***
Confidence	4.35	3.80	7.36	< 0.001 ***

Finally, in comparing those with and without training, the sample t-test revealed that participants who had received education or training on deepfakes ($n = 37$) achieved an average score of 7.73 (64.4% accuracy), while those without training ($n = 92$) scored an average of 6.28 (52.3% accuracy). The difference between the two groups reached statistical significance, $p = 0.0099$.

4.6. Overview of open questions

Table 7 summarises the main themes emerging from participants' open-ended responses and provides short illustrative quotes for each theme.

Table 7. Key themes in open-ended questions

Theme	Example
Stronger regulation	"Absolutely should make policies/laws to protect against deepfakes."
Freedom	"No, freedom matters too."
Deepfake education	"Schools should run workshops."
Technical measures	"AI tools help verify."

Participants' views on whether relevant policies would influence their choice of residence or study destination were divided. Approximately one-third of participants said they would be more likely to choose countries or regions with stricter regulations, believing this would provide a safer online environment. Over half of the participants, however, believed that policy would not influence their decision-making, with career prospects, educational opportunities, or freedom being more significant considerations.

5. Discussion

5.1. Interpretation of findings

Firstly, the overall accuracy of deepfake identification tasks in this research stands at 55.8%, which is highly consistent with Hypothesis 1 and in line with the latest meta-analysis findings from Diel et al. [28]: across 56 empirical studies, humans achieved an overall accuracy of 55.54% (95% confidence interval [48.87, 62.10]) in deepfake detection, with approximately 53.16% accuracy for image-based media. This indicates that although university students are highly dependent on digital platforms, they do not demonstrate better identification capabilities than the general public.

The model was statistically significant overall but demonstrated limited explanatory power ($R^2 = 0.088$), explaining only 8.8% of the variance. Within this model, educational levels ($p = 0.046$) and AI tool usage experience ($p = 0.043$) tended to be significant positive predictors, indicating that both higher educational levels and practical familiarity with technology contribute to improved identification capabilities. By contrast, age and gender proved insignificant, which is consistent with existing research by Tambe and Hussein indicating that basic demographic characteristics do not determine deepfake identification ability in a systematic manner [2].

This conclusion is further supported by additional findings: participants who received relevant training demonstrated an identification accuracy rate of 64.4%, while the untrained group achieved only 52.3% ($p = 0.0099$). This significant difference implies that structured educational intervention can effectively enhance identification capabilities, thereby verifying Hypothesis

2. Regarding cultural background, the regression model failed to achieve statistical significance ($p = 0.0738$), thus failing to support Hypothesis 3.

Following the identification task, participants' trust in the authenticity of digital images decreased significantly from 3.37 to 2.82 ($p < 0.001$), while their self-judgement confidence also declined from 4.35 to 3.80 ($p < 0.001$). This finding is consistent with existing research [33], which found that exposure to deepfakes induces generalised doubt, causing individuals to question not only fabricated content but also authentic material.

5.2. Insights and unexpected findings

On a practical level, prior education can effectively improve deepfake literacy and should be included within the overall framework of digital literacy courses in higher education. Research shows that identification capabilities are not fixed but can be cultivated through structured training, which holds direct value for course design and educational policy.

Firstly, the influence of cultural background was limited, which contrasts with the assumptions of existing research. Secondly, while training significantly improved identification capabilities, the exposure of deepfakes simultaneously reduced trust and confidence.

5.3. Limitations

Firstly, the explanatory power of regression models is limited (e.g., $R^2 = 0.088$ in the demographic model), indicating that unmeasured factors of greater significance exist within deepfake identification.

Furthermore, the analysis of cultural backgrounds was constrained by data limitations. The sample size was relatively small, and the excessive number of cultural categories resulted in sparse data for many countries, with either no participants or only one or two. Consequently, the findings of this research cannot be simplistically interpreted as indicating that 'cultural factors are unimportant'.

The trust and confidence measures in this study rely on self-reporting, which may be impacted by biases such as social expectations or inadequate reflection through single Likert-scale questions.

5.4. Future research directions

Future research on university students should expand sample sizes to cover student populations across more countries, in order to enhance the robustness and cross-cultural applicability of conclusions. Also, employing a longitudinal research design would help examine the persistence and evolution of training and exposure effects over time. Furthermore, the survey design requires improvement to allow for more nuanced measures of trust, confidence, and behavioural changes.

6. Conclusion

This research showed that university students' overall accuracy in identifying deepfake content stood at 55.8%, consistent with existing studies indicating that human performance in identifying manipulated content remains close to random. Demographic variables such as age and gender showed no significant effect, while education level and experience with AI tools presented positive predictive effects. Prior deepfake training significantly enhanced identification accuracy. It is important to note that the exposure of deepfakes has significantly decreased trust in digital images and confidence in one's own judgement, highlighting the psychological risks posed by synthetic media. The findings indicated that university students have limited capabilities in identifying deepfakes.

However, this study also has some limitations, including a small sample size, unbalanced distribution of cultural groups, reliance on self-reported measures of trust and confidence, and a survey design restricted to image comparisons.

References

- [1] Dong, R., Yuan, D., Wei, X., Cai, J., Ai, Z., & Zhou, S. (2025). Exploring the relationship between social media dependence and internet addiction among college students from a bibliometric perspective. *Frontiers in Psychology*, *16*, 1463671. <https://doi.org/10.3389/fpsyg.2025.1463671>
- [2] Tambe, S. N., & Hussein, N. A.-H. K. (2023). Exploring the impact of digital literacy on media consumer empowerment in the age of misinformation. *MEDAAD*, *2023*, 1–9. <https://doi.org/10.70470/medaad/2023/001>
- [3] Hasan, A. B., Alsabri, M. A., Alharbi, A., & Okela, A. H. (2024). Artificial intelligence literacy among university students—a comparative transnational survey. *Frontiers in Communication*, *9*. <https://doi.org/10.3389/fcomm.2024.1478476>
- [4] Russell, S. J., & Norvig, P. (2022). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

- [5] Poole, D. L., & Mackworth, A. K. (2023). *Artificial intelligence: Foundations of computational agents* (3rd ed.). Cambridge University Press. <https://artint.info>
- [6] Nilsson, N. J. (2009). *The quest for artificial intelligence*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511819346>
- [7] CSRC Content Editor. (2025). *Generative artificial intelligence - Glossary*. Computer Security Resource Center, NIST. https://csrc.nist.gov/glossary/term/generative_artificial_intelligence
- [8] Babaei, R., Cheng, S., Duan, R., & Zhao, S. (2025). Generative artificial intelligence and the evolving challenge of deepfake detection: A systematic analysis. *Journal of Sensor and Actuator Networks*, 14(1), 17. <https://doi.org/10.3390/jsan14010017>
- [9] Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, 9(11), 39–52. <https://timreview.ca/article/1282>
- [10] Hawkins, W., Russell, C., & Mittelstadt, B. (2025). *Deepfakes on demand: The rise of accessible non-consensual deepfake image generators* [Preprint]. arXiv. <https://arxiv.org/abs/2505.03859>
- [11] Zendran, M., & Rusiecki, A. (2021). Swapping face images with generative neural networks for deepfake technology – experimental study. *Procedia Computer Science*, 192, 834–843. <https://doi.org/10.1016/j.procs.2021.08.086>
- [12] Citron, D. K., & Chesney, R. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107(6), 1753–1820. https://scholarship.law.bu.edu/faculty_scholarship/640/
- [13] Mukta, M. S. H., Ahmad, J., Raiaan, M. A. K., Islam, S., Azam, S., Ali, M. E., & Jonkman, M. (2023). An investigation of the effectiveness of deepfake models and tools. *Journal of Sensor and Actuator Networks*, 12(4), 61. <https://doi.org/10.3390/jsan12040061>
- [14] Kaur, A., Hoshyar, A. N., Saikrishna, V., Firmin, S., & Xia, F. (2024). Deepfake video detection: Challenges and opportunities. *Artificial Intelligence Review*, 57(6), 161. <https://doi.org/10.1007/s10462-024-10810-6>
- [15] Ilchenko, A. (2023). *Fake news as a distortion of media reality: Tell-truth strategy in the post-truth era* (Publication No. 30792942) [Doctoral dissertation, University of Missouri-Columbia]. ProQuest Dissertations & Theses Global.
- [16] Pfänder, J., & Altay, S. (2025). Spotting false news and doubting true news: A systematic review and meta-analysis of news judgements. *Nature Human Behaviour*. Advance online publication. <https://doi.org/10.1038/s41562-024-02086-1>
- [17] Soe, S. O. (2019). A unified account of information, misinformation, and disinformation. *Synthese*, 198(6), 5929–5949. <https://doi.org/10.1007/s11229-019-02444-x>
- [18] Gelfert, A. (2018). Fake news: A definition. *Informal Logic*, 38(1), 84–117. <https://doi.org/10.22329/il.v38i1.5068>
- [19] Gelfert, A. (2021). What is fake news? In M. Hannon & J. de Ridder (Eds.), *The Routledge handbook of political epistemology* (pp. 171–180). Routledge. <https://doi.org/10.4324/9780429326769-22>
- [20] Appel, M., & Prietzel, F. (2022). The detection of political deepfakes. *Journal of Computer-Mediated Communication*, 27(4), zmac008. <https://doi.org/10.1093/jcmc/zmac008>
- [21] Sandoval, M.-P., De Almeida Vau, M., Solaas, J., & Rodrigues, L. (2024). Threat of deepfakes to the criminal justice system: A systematic review. *Crime Science*, 13(1), 20. <https://doi.org/10.1186/s40163-024-00239-1>
- [22] Ahmed, S. (2021). Who inadvertently shares deepfakes? Analyzing the role of political interest, cognitive ability, and social network size. *Telematics and Informatics*, 57, 101508. <https://doi.org/10.1016/j.tele.2020.101508>
- [23] Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, 6(1). <https://doi.org/10.1177/2056305120903408>
- [24] Hameleers, M. (2024). Cheap versus deep manipulation: The effects of cheapfakes versus deepfakes in a political setting. *International Journal of Public Opinion Research*, 36(1), edae004. <https://doi.org/10.1093/ijpor/edae004>
- [25] Ranka, H., Surana, M., Kothari, N., Pariawala, V., Banerjee, P., Surve, A., Reddy, S. S., Jain, R., Lalwani, J., & Mehta, S. (2024). *Examining the implications of deepfakes for election integrity* [Preprint]. arXiv. <https://arxiv.org/abs/2406.14290>
- [26] Farmer, L. (2022). Visual literacy and fake news: Gaining a visual voice. *Studies in Technology Enhanced Learning*, 2(2). <https://doi.org/10.21428/8c225f6e.b34036b2>
- [27] Croitoru, F.-A., Hiji, A.-I., Hondru, V., Ristea, N. C., Irofti, P., Popescu, M., Rusu, C., Ionescu, R. T., Khan, F. S., & Shah, M. (2024). *Deepfake media generation and detection in the generative AI era: A survey and outlook* [Preprint]. arXiv. <https://arxiv.org/abs/2411.19537>
- [28] Diel, A., Lalgı, T., Schröter, I. C., MacDorman, K. F., Teufel, M., & Bächerle, A. (2024). Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. *Computers in Human Behavior Reports*, 16, 100538. <https://doi.org/10.1016/j.chbr.2024.100538>
- [29] Feng, K. J., Ritchie, N., Blumenthal, P., Parsons, A., & Zhang, A. X. (2023). *Examining the impact of provenance-enabled media on trust and accuracy perceptions* [Preprint]. arXiv. <https://arxiv.org/abs/2303.12118>
- [30] da Gama Batista, J., Bouchaud, J.-P., & Challet, D. (2015). Sudden trust collapse in networked societies. *The European Physical Journal B*, 88(3), 71. <https://doi.org/10.1140/epjb/e2015-50645-1>
- [31] Groh, M., Sankaranarayanan, A., Singh, N., Kim, D. Y., Lippman, A., & Picard, R. (2022). *Human detection of political speech deepfakes across transcripts, audio, and video* [Preprint]. arXiv. <https://arxiv.org/abs/2202.12883>
- [32] Hameleers, M., van der Meer, T. G. L. A., & Dobber, T. (2024). Distorting the truth versus blatant lies: The effects of different degrees of deception in domestic and foreign political deepfakes. *Computers in Human Behavior*, 152, 108096. <https://doi.org/10.1016/j.chb.2023.108096>
- [33] Weikmann, T., Greber, H., & Nikolaou, A. (2024). After deception: How falling for a deepfake affects the way we see, hear, and experience media. *The International Journal of Press/Politics*, 30(1), 96–118. <https://doi.org/10.1177/19401612241233539>
- [34] Kim, J.-H. (2022). The excessive use of social media among college students: The role of mindfulness. *Open Access Journal of Addiction and Psychology*, 5(5), 1–8. <https://irispublishers.com/oajap/fulltext/the-excessive-use-of-social-media-among-college-students-the-role-of-mindfulness.ID.000624.php>

- [35] Attewell, S. (2025, May 21). *Student perceptions of AI 2025*. National Centre for AI. <https://nationalcentreforai.jiscinvolve.org/wp/2025/05/21/student-perceptions-of-ai-2025/>
- [36] Krupp, L., Steinert, S., Kiefer-Emmanouilidis, M., Avila, K. E., Lukowicz, P., Kuhn, J., Küchemann, S., & Karolus, J. (2023). *Unreflected acceptance—Investigating the negative consequences of ChatGPT-assisted problem solving in physics education* [Preprint]. arXiv. <https://arxiv.org/abs/2309.03087>
- [37] Roe, J., Perkins, M., & Furze, L. (2024). Deepfakes and higher education: A research agenda and scoping review of synthetic media. *Journal of University Teaching and Learning Practice*, 21(10). <https://doi.org/10.53761/2y2np178>
- [38] Nygren, T., Wiksten Folkeryd, J., Liberg, C., & Guath, M. (2020). Students assessing digital news and misinformation. In S. Tauber, M. A. Livingston, M. J. Smith, & A. J. Cowen (Eds.), *Disinformation in open online media* (Vol. 12259, pp. 63–79). Springer. https://doi.org/10.1007/978-3-030-61841-4_5
- [39] Leeder, C. (2019). How college students evaluate and share “fake news” stories. *Library & Information Science Research*, 41(3), 100967. <https://doi.org/10.1016/j.lisr.2019.100967>
- [40] Alexander, S. (2025). Deepfake cyberbullying: The psychological toll on students and institutional challenges of AI-driven harassment. *The Clearing House*. Advance online publication. <https://doi.org/10.1080/00098655.2025.2488777>
- [41] Cosme Torres, L. (2025, April 4). *Law student allegedly used AI to create porn of fellow students — then tried to apologize*. People. <https://people.com/law-student-allegedly-used-ai-create-porn-fellow-students-11773557>
- [42] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 1–11). IEEE. <https://github.com/ondyari/FaceForensics>