RedFame

Studies in Media and Communication
Vol. 14, No. 3; September 2026
ISSN: 2325-8071 E-ISSN: 2325-808X
Published by Redfame Publishing
URL: http://smc.redfame.com

Self-Assessed Detection Ability for Identifying AI-Generated Visual Content: Evidence from Saudi Arabia

Maisoon Alsebaei[1], Marwa Atyah[2]

[1]Associate Professor of Journalism and Digital Media, King Abdulaziz University, Jeddah, Saudi Arabia

[2]Associate Professor of Journalism and Digital Media, King Abdulaziz University, Jeddah, Saudi Arabia. Email: matyah@kau.edu.sa

Correspondence: Maisoon Alsebaei, Associate Professor of Journalism and Digital Media, King Abdulaziz University, Jeddah, Saudi Arabia. Email: malsebaei@kau.edu.sa

Received: March 25, 2026 Accepted: May 11, 2026 Online Published: June 29, 2026

doi:10.11114/smc.v14i3.8611 URL: https://doi.org/10.11114/smc.v14i3.8611

Abstract

This study examines individuals' ability to distinguish between genuine and generative AI (Gen-AI) content across visual media, addressing the growing erosion of trust in digital information. Using a two-part questionnaire, participants were assessed on classification accuracy and self-assessed detection ability used to identify Gen-AI images and videos. The cross-sectional perceptual classification study design employed varied content types and randomized presentations to reduce perceptual bias. Results showed that participants demonstrated limited ability to accurately distinguish AI-generated from genuine content, with overall performance only slightly above chance level. Signal Detection Theory analysis revealed low perceptual sensitivity and a minimal authenticity bias toward classifying content as real. One-way ANOVA and MANOVA results indicated no significant differences in actual performance or self-assessed detection ability according to AI experience level. Regression analyses further showed that self-assessed detection ability generally did not predict actual performance. The findings highlight the need for stronger visual literacy, specialized training, and improved verification approaches for AI-generated media.

Keywords: AI-generated content, media authenticity, human perception, visual perception, AI experience, generative AI

1. Introduction

Artificial Intelligence (AI) has rapidly advanced in recent years, particularly in the generation of visual content. Generative Artificial Intelligence (Gen-AI), leveraging deep learning and generative models, can now produce images and videos that closely mimic human-created content (OpenAI, 2021). This innovation has expanded creative possibilities across media, education, entertainment, and marketing, while streamlining production processes and reducing costs.

Prominent applications include photorealistic image generators such as DALL-E and Stable Diffusion, which transform text prompts into detailed visuals (Bray et al., 2023), video tools such as Runway and Synthesia that generate realistic scenes from text (Rouxel, 2020). However, the increasing realism of Gen-AI content poses significant challenges, as users increasingly struggle to distinguish it from authentic media, raising concerns related to misinformation and media trust. While prior research has largely focused on Western contexts or technical detection methods (Nightingale & Farid, 2022; Diel et al., 2024), relatively few studies have examined individuals' abilities for identifying Gen-AI content within Arab populations.

This study investigates the Saudi undergraduate students' capacity to recognize Gen-AI–produced visual content and the perceived detection ability they employ. It seeks to advance understanding of Gen-AI's impact on media–audience interaction and contribute to scholarly discussions on digital literacy and responsible engagement with AI-generated content.

2. Literature Review

Reviewing the scientific literature surrounding this study reveals that generative AI has undergone rapid development, enabling it to produce highly realistic visual and textual content, which created unprecedented cognitive and communicative challenges. This new reality prompted a growing number of studies to examine individuals' ability to distinguish between genuine and Gen-AI content, assess the efficacy of computer models in detecting forgeries, and analyze the psychological, social, and cultural dimensions accompanying these transformations (Nightingale & Farid, 2022; Groh et al., 2022).

Concerning visual content, particularly human faces, Nightingale and Farid's (2022) study revealed that individuals' ability to distinguish between genuine and Gen-AI faces made with the StyleGAN2 model was close to that of random guesswork. As a matter of fact, artificial faces exhibited higher levels of trustworthiness than real faces. These findings are of particular significance as they suggest that the development of visual generative technologies has transcended the so-called "valley of the uncanny," weakening the traditional perceptual mechanisms individuals rely on to judge authenticity. The study also showed that training and feedback did not result in a significant improvement in performance, limiting the effectiveness of simple awareness-raising interventions against this type of content.

These findings are consistent with those of Bray, Johnson, and Kleinberg (2023), which demonstrated that participants' accuracy in recognizing fake facial images generated by StyleGAN2 did not exceed 62%, with no apparent effect of brief awareness-raising interventions. This recurring deficiency in human performance reflects that the problem lies not only in a lack of knowledge but also in the very nature of visual perception itself when dealing with high-quality artificial content. In a similar pursuit, Blomgren and Harald (2024) examined the impact of age differences and self-confidence levels. They found that younger groups were relatively more accurate, while self-confidence played a mixed role, suggesting that some individuals may overestimate their ability to discriminate without this actually reflecting on their performance. This finding reinforces the conclusions of Miller et al. (2023) on the phenomenon of "AI hyperrealism" as they demonstrated that some Gen-AI faces were rated as even more realistic than genuine ones. The Dunning-Kruger effect was also evident in their participant behavior as poor performance was associated with higher levels of cognitive confidence.

This cognitive deficit is not limited to still images but extends to active visual content as well. Groh et al. (2022) showed that participants' performance in detecting deepfake videos was comparable to that of a leading computer vision model, despite the differences in error patterns between humans and the model. The study showed that disabling the cognitive processing of faces led to a significant decrease in human performance, while the performance of the automated model was not affected to the same degree. This underscores humans' reliance on specialized cognitive mechanisms for facial processing when judging content authenticity. Even though combining human judgment with AI input occasionally improved accuracy, inaccurate model predictions can weaken participants' performance, which raises questions about the risks of uncritical reliance on intelligent systems.

In addition, Diel et al. (2024) provided an extensive, comprehensive guide of the limitations of human performance. Their meta-analysis of 56 studies showed that humans' average accuracy of identifying deepfake content across various media (image, text, and video) did not exceed 55.54%. The results also revealed a general cognitive bias, which is a greater ability to recognize genuine content compared to the ability to recognize fake content, regardless of the content type. While some strategies, such as AI training or support, led to some improvement in performance, this improvement remained limited and did not reach high levels of reliability.

In environments simulating everyday use of social media platforms, a study by Cooke et al. (2025) showed that participants' average accuracy in distinguishing between genuine and Gen-AI content was only about 51%, a level too close to an arbitrary coin flip. The study also indicated a clear bias toward assuming the authenticity of digital content, audiovisual content being relatively easier to detect than single-media content, and images, particularly those containing human faces, scoring the lowest on human detectability. These findings highlight the importance of environmental context in explaining cognitive performance, as conditions of rapid browsing and distraction impair individuals' ability to differentiate.

Lu et al. (2023) and Martin-Rodriguez et al. (2023) have demonstrated the superiority of computer models in differentiating between genuine and Gen-AI content. Convolutional Neural Networks and pixel analysis techniques like PRNU and ELA particularly achieved accuracy levels exceeding 95% in some experimental environments. However, this technological superiority is not without limitations; a review by Alrashoud (2025) showed that the performance of these models declines in real-world environments due to challenges such as generalization, video compression, low resolution, adversarial attacks, and the lack of standardized evaluation criteria, which constrains their practical reliability.

As for textual content, a study by Fiedler and Döpke (2025) revealed that the ability of both human experts and automated detection tools to distinguish between genuine and Gen-AI academic texts was close to chance. That was especially true when dealing with high-quality professional texts, where nearly 20% of participants failed to classify them correctly. These results highlight the fragility of traditional assessment methods in higher education and confirm that text quality is no longer a reliable indicator of its source.

In the Arab context, Alghamdi and Alowibdi's (2024) study focused on the technical aspect, demonstrating that machine learning models can distinguish between genuine and Gen-AI Arabic tweets with an accuracy of up to 93%. Despite the importance of these findings, the study was limited to automated performance and did not address how Arab users perceive this content or the self-perceived detection abilities they employ to judge its authenticity. In contrast, other Arab studies have focused on the cultural, artistic, and ethical dimensions of AI use. This includes Abu Zeid, Mohamed, and Eid's

(2023) study on visual arts; Al-Sawy's (2023) study on the risks of deepfakes; Al-Toukhy's (2024) study on the impact of the technological revolution on visual arts; and Al-Ghabari and Yousry's (2023) study on digital transformation in media institutions. Whilst these studies contribute to understanding the cultural and social context, they do not directly link these dimensions to the audience's cognitive recognition behaviors.

Within a broader explanatory framework, Pinski and Benlian (2024) presented the concept of AI culture as a conceptual framework integrating technical knowledge, skills, and ethical awareness. They demonstrated that the weak integration of these dimensions contributes to users' limited ability to critically evaluate AI outputs. This perspective helps explain the recurring findings in experimental studies regarding the poor performance of humans in recognizing Gen-AI content.

Therefore, it is evident that the existing literature, despite its richness, has dedicated its greatest attention to measuring detection accuracy or comparing human and machine performance, while overlooking broader challenges related to AI literacy, media trust, and the cultural conditions shaping how individuals interpret AI-generated content. Nonetheless, the cognitive and communicative self-perceived abilities individuals employ when attempting to recognize Gen-AI visual content remain an understudied area. Furthermore, the existing literature has often processed the different media in isolation, neglecting the integration of the visual dimensions within a single analytical framework.

3. Study questions

RQ1: How well can individuals distinguish between actual and Gen-AI content?

RQ2: To what extent do individuals use differentiation abilities to distinguish between natural and Gen-AI content (images and videos)?

4. Study Hypotheses

H1: Individuals' ability to distinguish between genuine and AI-generated content differs significantly according to their level of AI experience.

H2: Participant's self-accessed ability to distinguish between genuine and AI-generated content (images and videos) differ significantly according to individuals' level of AI experience.

H3: Participants' self-assessed ability to distinguish between genuine and AI-generated visual content (Image and videos) significantly predicts actual classification performance.

5. Methodology

The study employed a cross-sectional perceptual classification study design, where participants were presented with visual stimuli (images and videos) and asked to classify each item as genuine or AI-generated. Data were collected using a structured questionnaire administered via Google Forms. The design enabled ~~enabled~~ assessment of participants' classification performance and self-assessed detection ability.

5.1 Study Population

The study focused on undergraduate students aged 18 to 22 in Saudi Arabia using a non-random purposive sample, a demographic identified as one of the most digitally integrated and active user groups in terms of visual content consumption. Research consistently indicates that individuals within this age range are frequent users of interactive digital platforms and are highly exposed to emerging forms of Gen-AI content, including generative AI tools and applications used in educational and media contexts (Kemp, 2025; Digital Education Council, 2024). This level of exposure makes them a contextually appropriate population for examining the self-perceived detection competence employed in distinguishing between genuine and Gen-AI content.

Methodologically, selecting a relatively homogeneous age group helps reduce demographic variability associated with differences in media experience and technology adoption, thereby strengthening internal validity and ensuring that observed differences in recognizability are more directly related to the study's primary variables rather than age effects (Mertala et al., 2024). While this group is often described as part of the "digital natives" generation, the concept is used here solely to indicate intensity of digital exposure rather than assumed cognitive advantage, in line with contemporary critical perspectives (UNESCO, 2024). Given their sustained engagement with digital media, this population is particularly relevant for understanding current and future interactions with Gen-AI content (Ideis, 2024).

5.2 Visual Stimuli

The visual materials used in researching the self-perceived detection competence for identifying visual Gen-AI content included images and videos created in different ways. The visual material counts varied on each side and during each phase of the study. The variation avoided priming participants into expected patterns and prevented reliance on numerical distributions as a reference for guessing. This reduced perceptual bias, enhancing the validity of the results. The distribution of the materials was:

1-Images:

- **Genuine Images**: To represent human photographs, ten genuine images captured by cameras were sourced from **Pixabay**. Characterized by their accuracy in capturing visual elements, such as natural lighting, as well as the subtle details of faces, living creatures, and places, these images precisely reflect the real world.
- **Images created by Gen-AI:** Nine images were generated by tools, such as Adobe Firefly. Totally fake images of people, animals, and places were generated with realistic features. These images represent the output of AI models that learn from massive databases to generate realistic-looking digital content.

These two types of images were used in this study to measure intelligent models' and humans' ability to distinguish between real and Gen-AI content. The stimuli were selected to allow in-depth analysis of visual details that may reveal artificial generation, such as unnatural patterns in shadows, faces, and texture, compared to genuine images captured by cameras. The visual materials represent a key focus of this research, which aims to measure the ability of individuals to distinguish between Gen-AI and genuine content.

2-videos:

The videos measured the ability of individuals to distinguish between Gen-AI and genuine content. The videos were selected as follows:

- **Genuine videos:** Three video scenes were recorded in the real-world using cameras. The videos featured a high level of realistic details regarding motion, lighting, and interaction between objects and natural elements. This level of detail enabled participants to distinguish between genuine and Gen-AI videos.
- **Gen-AI videos:** Five videos were created using AI models, such as generative adversarial networks (GANs) or diffusion models. The visual content, featuring scenes that mimic nature or living creatures, was generated to be realistic. However, they also exhibit artificial patterns of movements or interactions with light and shadows. OpenAI's Sora in particular was used to generate those videos.

Participants compared the Gen-AI videos with the genuine videos to assess how well they could distinguish between the two types of visual content.

5.3 Gen-AI Tool Selection and Justification

This study employed Sora (OpenAI) and Adobe Firefly to generate the visual materials used in the study. These tools were selected because they enable the production of high-quality, realistic content that closely resembles real-world visual media, which is essential for accurately assessing audience response**s (**OpenAI, 2024; Adobe, 2023).

In addition to realism, the selected tools provide a high degree of control over generated content. Sora supports the creation of temporally and narratively consistent videos, and Adobe Firefly enables the generation of diverse and high-quality still images. This combination allowed for consistent manipulation of experimental variables across stimuli, strengthening methodological rigor (Ho et al., 2020; Dwivedi et al., 2023).

Finally, these tools operate within established ethical and regulatory frameworks, relying on licensed training data and transparency policies, and produce outputs that resemble contemporary digital content encountered on social media platforms. This alignment enhances ecological validity while maintaining ethical research standards (Floridi et al., 2018; OpenAI, 2024). Overall, the selected tools balance realism, experimental control, and ethical considerations, supporting the reliability of the study's findings.

5.4 Data Collection

The study used a questionnaire designed to measure individuals' ability to distinguish between genuine and Gen-AI content as well as their detection competence for identifying such content across different media. The tool included two main parts. The first tested the accuracy of participants in classifying images and video recordings. The second focused on the metrics for analyzing the self-assessed detection of image quality, audiovisual synchronization, and motion consistency, along with specialized tools to verify the authenticity of digital content. Each of these parts is explained below.

The stimuli consisted of 19 images (10 genuine and 9 AI-generated) and 8 videos (3 genuine and 5 AI-generated). The stimuli were presented in randomized order to reduce response bias. Each item was presented once, after which participants made a classification decision without the option for replay. Differentiation accuracy was calculated as the percentage of correct classifications across all presented items. Thereafter, participants responded to a screening question after the classification task, where they responded regarding their ability to distinguish AI-generated from real media. Those who responded "Yes" went on to fill out the self-assessment items, while those who responded "No" were asked only completed the demographic section. This branching logic of classification resulted from separating the participants on the basis of their perception of their capacity in the task of classification.

5.5 Measuring Individuals' Ability to Distinguish Between Gen-AI and Genuine Content

Differentiation accuracy of individuals between genuine and Gen-AI content was measured by the percentage of correct

classifications they made when evaluating the presented visual stimuli. A media content set, including images and video recordings, was presented to the participants. Some items in this set were naturally captured, whereas others were generated using AI. The ability of the participants to differentiate between the two types of content was evaluated based on correctly categorizing each piece of content. The differentiation ability was calculated based on the percentage of correct answers.

5.6 Measures of Participants Self-Accessed Detection Abilities for Gen-AI content Identification

Participants' self-assessed detection ability was measured using Likert-scale items adapted from previous studies on human recognition of AI-generated media (Lu et al., 2023; Frank et al., 2024; Epstein et al., 2023). The items tested the participants' perceived ability in detecting inconsistencies, irregular motions, and audiovisual synchronization errors, as well as other clues often found in AI-generated videos (Epstein et al., 2023).

Gen-AI Images of people - created by Adobe Firefly Genuine images taken by cameras - obtained from PixaBay

Figure 1. Samples of the images used as experimental materials

5.7 Study Instrument Validity and Reliability

To measure the internal consistency of the study questionnaire, Cronbach's alpha coefficient was calculated to ensure the reliability of the study using a sample of 26 individuals who were later excluded from the total sample. Table 1 lists the consistency coefficients of the study tool along various axes.

Table 1 shows that the overall consistency coefficient of all the study axes in the questionnaire was high (0.894 for all 21 items). The consistency coefficients of the three axes ranged from 0.733 to 0.828. This indicates that the questionnaire had a high degree of consistency and could be relied upon and applied to larger study samples. This is according to Nunnally's threshold, which sets 0.7 as the minimum consistency coefficient for exploratory research.

Table 1. Cronbach's alpha coefficient: Results used to measure the reliability of the study tool

Axes	Number of items	Axis consistency
Axis 1 - Image Identification	9	0.828
Axis 2 - Video Identification	6	0.733
Overall questionnaire consistency	**15**	**0.863**

5.8 Internal Consistency Validity

To verify the internal consistency of the questionnaire items, Pearson's correlation coefficient was calculated between the scores of each item on the three axes and the total score on the axis to which the item belonged, as shown in Tables 2 and 3.

Table 2 shows a statistically significant correlation between the items on the first axis and the total score of the axis. The Pearson correlation coefficient for the first item was statistically significant at a significance level of 0.05, whereas the Pearson correlation coefficients of all the other items were statistically significant at a significance level of 0.01. All correlation coefficients ranged between 0.485 and 0.819, indicating that all items on the first axis were internally consistent with the axis itself. This validates the internal consistency of all first-axis items.

Table 2. Correlation coefficients between first-axis items and the overall axis

Item	Correlation coefficient	P-value
1	0.485*	0.012
2	0.698**	0.000
3	0.690**	0.000
4	0.819**	0.000
5	0.607**	0.001
6	0.621**	0.001
7	0.615**	0.001
8	0.723**	0.000
9	0.683**	0.000

** The correlation coefficient is significant at a significance level of 0.01.

* The correlation coefficient is significant at a significance level of 0.05.

Table 3 shows a statistically significant correlation between the items on the second axis and the total score on the second axis. The Pearson correlation coefficient for the sixth item was statistically significant at a significance level of 0.05, whereas the Pearson correlation coefficients of all the other items were statistically significant at a significance level of 0.01. All correlation coefficients ranged between 0.477 and 0.773, indicating that all the items on the second axis were internally consistent with the axis itself. This validates the internal consistency of all second-axis items.

According to the results of the reliability and internal consistency tests in the previous tables, it is evident that the questionnaire prepared for this study has a high level of reliability and that the items within the questionnaire axes are internally consistent. This allowed us to apply the questionnaire to the entire sample.

Table 3. Correlation coefficients between second-axis items and the overall axis

Item	Correlation coefficient	P-value
1	0.557**	0.003
2	0.641**	0.000
3	0.765**	0.000
4	0.697**	0.000
5	0.773**	0.000
6	0.477*	0.014

** The correlation coefficient is significant at a significance level of 0.01.

* The correlation coefficient is significant at a significance level of 0.05.

5.9 Signal Detection Theory Analysis

In addition to descriptive accuracy measures, Signal Detection Theory (SDT) was employed to provide a more rigorous assessment of perceptual sensitivity and response bias in distinguishing between genuine and AI-generated content (Macmillan & Creelman, 2004). SDT distinguishes between participants' actual sensitivity to AI-generated stimuli and their response tendencies when making classification decisions. For each participant, hit rates (correct identification of AI-generated content) and false alarm rates (incorrect classification of genuine content as AI-generated) were calculated. These measures were further used to compute sensitivity scores (d′) and response bias values (criterion c). Higher d′ values indicate stronger perceptual sensitivity, while positive criterion values indicate a bias toward classifying content as genuine.

5.10 Statistical Analysis

Descriptive statistics were used to summarize participants' demographic characteristics and classification accuracy rates. One-way ANOVA was conducted to examine differences in overall classification accuracy according to AI experience level (H1). A multivariate analysis of variance (MANOVA) was employed to assess differences in participants' self-assessed detection ability across AI experience groups (H2). Simple linear regression analyses were conducted to evaluate whether self-assessed detection ability predicted actual classification performance for images, videos, and overall visual content (H3). Statistical significance was evaluated at the .05 level using IBM SPSS version 27.

6. Results

This section reports the statistical results of the study, including participants' ability to distinguish between genuine and

AI-generated content and the outcomes of the hypothesis testing. All results are presented in the following tables.

A total of 193 participants participated in the study. Table 4 showed that there were more females than males in the study; the participants are predominantly BA degree holders, while majority are either limited or well experienced; as regards specializations, participants were mainly Public relations and Audio-visual experts.

Table 4. Demographic Variables

	Frequency	% Frequency
Gender		
Male	27	13.8%
Female	166	85.1%
Education Level		
Diploma	9	4.6%
BA	173	88.7%
Graduate Studies	11	5.6%
Experience in AI		
No Experience	29	14.9%
Limited Experience	77	39.5%
Experienced	86	44.1%
Specialization		
Audio Visual Production	49	25.1%
Public Relation	62	31.8%
Marketing Communication	18	9.2%
Journalism and Digital Media	35	17.9%
Other	29	14.9%

n = 193

Table 5 presents the percentage of correct responses for questions assessing participants' ability to distinguish between genuine images and AI-generated images. Correct identification rates varied by image content, with the highest accuracy observed for images of animals and birds (70.5%), followed by images of people (52.3%), while images of natural places yielded the lowest accuracy (47.3%). Differences were also observed according to image type, as participants identified genuine images with a correct answer rate of 60.1%, compared to 56.0% for AI-generated images. Overall, participants correctly answered an average of 58.1% of the image-based discrimination questions.

Table 5. Distinguishing between genuine and Gen-AI images

		Count	Average correct answers	Ratio (%)
Image content	People	4	2.09	52.3
	Animals and birds	8	5.64	70.5
	Natural places	7	3.31	47.3
Image type	Genuine	10	6.01	60.1
	AI Generated	9	5.04	56.0
Overall Average			**11.04**	**58.1**

Table 6 shows the percentage of correct answers to questions that distinguished between genuine videos and videos generated by AI. The correct answer rates varied according to video content, as participants were able to distinguish between genuine and Gen-AI videos of animals and birds with a correct answer rate of 64.8%, while participants' correct answers for videos of natural places were less common, at a rate of 42.3%. The correct answer rates also varied according to video type, with participants identifying genuine videos with a correct answer rate of 63%, compared to a correct answer rate of 52.4% for Gen-AI videos. Overall, participants correctly answered an average of 56.4% of the questions on distinguishing genuine videos from Gen-AI videos.

Table 6. Distinguishing between genuine and Gen-AI videos

		Count	Average correct answers	Ratio (%)
Video content	Animals and birds	5	3.24	64.8
	Natural places	3	1.27	42.3
Video type	Genuine	3	1.89	63.0
	AI Generated	5	2.62	52.4
Overall Average			**4.51**	**56.4**

Table 7. Distinguishing between genuine and Gen-AI content

		Count	Average correct answers	Ratio (%)
Content type	Genuine	13	7.89	60.7
	AI Generated	14	7.66	54.7
Overall Average			**15.55**	**57.6**

Table 7 shows the percentage of correct answers to the questions distinguishing between genuine and Gen-AI content. Participants were able to identify genuine content with a correct answer rate of 60.7%, whereas Gen-AI content had a correct answer rate of 54.7%. Overall, participants correctly answered an average of 57.6% of the questions on distinguishing genuine content from Gen-AI content.

Table 8. Measuring strategies used to distinguish between genuine and Gen-AI images

Statement		Sample responses			Mean	Standard Deviation	Sample Direction	Order
		Disagree	Neutral	Agree				
I can detect inconsistencies in image details such as shadows and skin texture.	Count	1	34	96	2.73	0.465	Agree	3
	%	0.8	26.0	73.3				
I am able to identify common errors in Gen-AI images such as facial distortions or repeating patterns.	Count	6	36	89	2.63	0.571	Agree	8
	%	4.6	27.5	67.9				
I can identify a genuine image by the quality of its colors and its fine details.	Count	4	28	99	2.73	0.512	Agree	4
	%	3.1	21.4	75.6				
I observe unrealistic details in Gen-AI images, such as uneven shapes or floating elements.	Count	1	22	108	2.82	0.408	Agree	1
	%	0.8	16.8	82.4				
I can identify genuine images by focusing closely on small details like light reflections.	Count	8	32	91	2.63	0.598	Agree	9
	%	6.1	24.4	69.5				
I use my image analysis expertise to identify Gen-AI images through recurring patterns and unnatural details.	Count	1	34	96	2.73	0.465	Agree	5
	%	0.8	26.0	73.3				
I notice minor distortions in Gen-AI images such as misaligned hands or feet.	Count	5	29	97	2.70	0.536	Agree	6
	%	3.8	22.1	74.0				
I can identify a genuine image by the background details and quality of integrated elements.	Count	3	22	106	2.79	0.464	Agree	2
	%	2.3	16.8	80.9				
I can identify Gen-AI images by comparing natural textures and elements with those in the image.	Count	5	29	97	2.70	0.536	Agree	7
	%	3.8	22.1	74.0				
Overall Average		**2.72**			**Standard Deviation**		**0.317**	

Table 8 presents participants' self-assessed detection abilities for distinguishing Gen-AI from genuine images, with responses shown as counts and percentages. The overall mean was 2.72 (SD = 0.317), indicating general agreement with these strategies. The most endorsed approach was noticing unrealistic details, such as uneven shapes or floating elements (M = 2.82, SD = 0.408; 82.4%), followed by evaluating background details and element integration (M = 2.79, SD = 0.464; 80.9%). Recognizing common AI errors like facial distortions or repeating patterns (M = 2.63, SD = 0.571; 67.9%) and attending to subtle cues such as light reflections (M = 2.63, SD = 0.598; 69.5%) were less frequently endorsed.

Table 9. Measuring strategies used to distinguish between genuine and Gen-AI videos

Statement		Sample responses			Mean	Standard Deviation	Sample Direction	Order
		Disagree	Neutral	Agree				
I can spot subtle differences in image quality that might indicate a video is Gen-AI.	Count	2	34	95	2.71	0.488	Agree	3
	%	1.5	26.0	72.5				
I often notice inconsistencies between voices and lip movements in a video edited or generated by AI.	Count	3	23	105	2.78	0.469	Agree	2
	%	2.3	17.6	80.2				
I rely on specialized tools to analyze and detect fake or Gen-AI videos.	Count	25	39	67	2.32	0.777	Neutral	6
	%	19.1	29.8	51.1				
I always check video sources to ensure their credibility before believing their content.	Count	8	29	94	2.66	0.592	Agree	4
	%	6.1	22.1	71.8				
I feel confident in my ability to distinguish between genuine and Gen-AI videos based on my personal experience.	Count	6	49	76	2.53	0.586	Agree	5
	%	4.6	37.4	58.0				
I believe that receiving training or education on how to identify Gen-AI videos is important for everyone.	Count	2	23	106	2.79	0.442	Agree	1
	%	1.5	17.6	80.9				
Overall Average		**2.63**			**Standard Deviation**		**0.356**	

Table 9 presents participants' self-assessed detection abilities for distinguishing Gen-AI from genuine videos, with responses reported as counts and percentages. The overall mean was 2.63 (SD = 0.356), indicating general agreement with the strategies. The most endorsed approach was valuing training or education on identifying Gen-AI videos (M = 2.79, SD = 0.442; 80.9%), followed by noticing mismatches between voices and lip movements (M = 2.78, SD = 0.469; 80.2%). Confidence in distinguishing videos based on personal experience (M = 2.53, SD = 0.586; 58.0%) and reliance on specialized detection tools (M = 2.32, SD = 0.777; 51.1%) were less frequently endorsed.

Signal Detection Theory (SDT) analysis was conducted to obtain a more stringent assessment of perceptual sensitivity and response bias. Table 10 showed that there is a low perceptual sensitivity in the distinguishing between AI-generated and real visual contents, $d' = 0.51$. Additionally, the mean criterion value was $c = 0.12$, indicating a modest response bias to mark content as genuine rather than AI-generated.

Table 10. SDT Results

Measure	Mean	Interpretation
d′ (Sensitivity)	0.51	Low perceptual sensitivity
Criterion c (Bias)	0.12	Slight bias toward classifying content as genuine

Hypotheses Results

H1: Individuals' ability to distinguish between genuine and AI-generated content differs significantly according to their level of AI experience.

Table 11. One-Way ANOVA of difference in Overall accuracy between AI experience groups

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	292.076	2	146.04	1.556	.214
Within Groups	17734.403	189	93.83		
Total	18026.479	191			

A one-way ANOVA was conducted to examine whether overall classification accuracy differed according to participants' level of AI experience. The analysis revealed no statistically significant difference among the groups, $F(2, 189) = 1.56$, $p = .214$, $\eta^2 = .016$, indicating a small effect size.

H2: Participant's self-accessed ability to distinguish between genuine and AI-generated content (images and videos) differ significantly according to individuals' level of AI experience.

A one-way multivariate analysis of variance (MANOVA) was conducted to examine whether participants' self-assessed detection ability to distinguish between genuine and AI-generated visual content differed according to AI experience level. The multivariate effect of AI experience was not statistically significant, Wilks' $\Lambda = .975$, $F(4, 250) = 0.79$, $p = .531$.

Table 12 showed that there was no statistically significant differences between AI experience groups for self-assessed image detection ability, $F(2, 126) = 0.05$, $p = .949$, $\eta p^2 = .001$, or self-assessed video detection ability, $F(2, 126) = 1.34$, $p = .265$, $\eta p^2 = .021$.

Table 12. MANOVA Results

	F	**df**	**p**	**Partial η²**
Image self-assessed ability	0.05	(2, 126)	.949	.001
Video self-assessed ability	1.34	(2, 126)	.265	.021

Wilk's = .975, F(4, 250) = 0.79

H3: Participants' self-assessed ability to distinguish between genuine and AI-generated visual content (Image and videos) significantly predicts actual classification performance.A simple linear regression was conducted to examine whether participants' self-assessed ability to distinguish between genuine and AI-generated images predicted actual image classification accuracy. The regression model was not statistically significant, $F(1, 129) = 2.18$, $p = .142$, explaining approximately 1.7% of the variance in image classification accuracy ($R^2 = .017$). Self-assessed image detection ability did not significantly predict actual performance, $\beta = -.129$, $t(129) = -1.48$, $p = .142$.

Table13. Linear Regression Predicting Image Classification Accuracy from Self-Assessed Image Detection Ability

Predictor	**B**	**SE B**	**β**	**t**	**p**	**95% CI for B**
Constant	71.63	7.26	—	9.87	< .001	[57.27, 85.99]
Image self-assessed ability	-4.65	3.15	-.129	-1.48	.142	[-10.88, 1.58]

A simple linear regression was conducted to examine whether participants' self-assessed ability to distinguish between genuine and AI-generated videos predicted actual video classification accuracy. The regression model was statistically significant, $F(1, 128) = 4.49$, $p = .036$, explaining approximately 3.4% of the variance in video classification accuracy ($R^2 = .034$). Participants' self-assessed video detection ability significantly and positively predicted actual video classification performance, $\beta = .184$, $t(128) = 2.12$, $p = .036$.

Table 14. Linear Regression Predicting Video Classification Accuracy from Self-Assessed Video Detection Ability

Predictor	**B**	**SE B**	**β**	**t**	**p**	**95% CI for B**
Constant	40.10	9.95	—	4.03	< .001	[20.41, 59.80]
Video self-assessed ability	8.80	4.15	.184	2.12	.036	[0.58, 17.02]

A simple linear regression was conducted to examine whether participants' overall self-assessed ability to distinguish between genuine and AI-generated visual content predicted overall classification accuracy. The regression model was not statistically significant, $F(1, 129) = 0.13$, $p = .721$, explaining approximately 0.1% of the variance in overall classification accuracy ($R^2 = .001$). Overall self-assessed detection ability did not significantly predict actual classification performance, $\beta = .031$, $t(129) = 0.36$, $p = .721$.

Table 15. Linear Regression Predicting Overall Classification Accuracy from Overall Self-Assessed Detection Ability

Predictor	B	SE B	β	t	p	95% CI for B
Constant	58.49	7.21	—	8.11	< .001	[44.21, 72.76]
Overall self-assessed ability	1.10	3.09	.031	0.36	.721	[-5.01, 7.22]

7. Discussion

This study investigated the accuracy of Saudi university students in detecting AI-generated content from real content, with emphasis on the link between students' self-perceptions of their skill at detecting AI-generated images and the actual accuracy of their detection. Results offer fresh insights into audience perceptions of authenticity in digital media and expand the body of knowledge on the nature of human interactions with increasingly realistic AI-generated reports.

The study showed participants had a relatively low ability to detect real vs AI-generated imagery and videos. The accuracy of classification of the images and videos was modestly greater than chance. This is in line with the previous studies that found that the reliability of previous perceptual cues for evaluating media authenticity is significantly impaired by the developments of generative AI technologies (Nightingale & Farid, 2022; Diel et al., 2024; Cooke et al., 2025).

The results of the SDT analysis also confirmed these findings with the participants' perceptual sensitivity being low; with participants having experienced significant difficulties distinguishing between AI-generated and real content. Furthermore, there was a hint of authenticity bias, meaning that participants tended to believe the images were real rather than created by an AI tool. This is consistent with previous studies indicating that there may be a wrong impression of credibility without strong cues from the host source (Cooke et al., 2025). Consequently, these findings argue that the issue of authenticity in AI generated media is beyond incorrect information; rather, it might also reveal perceptual and cognitive issues in the perception of highly realistic synthetic media.

From the self-assessed detection empirical results, it can be surmised that the participants largely depended on general perception judgment and discrepancies that could be seen when judging authenticity of content. The AI-generated media covered issues such as visual abnormalities, inconsistencies in movement and synchronization issues between the audio and visual components are issues the participants were confident with identifying. The association between perceived ability and performance was generally, weak. All subjects rated their ability at moderate to high levels of confidence but this confidence was not always reflected in higher classification accuracy. The results mirror other previous research indicating that the ability to identify AI generated media might be higher than actual performance detection capabilities (Miller et al. 2023, Lu et al. 2023).

Furthermore, results for the first hypothesis did not show statistically significant differences in accuracy of actual classification based on participants' level of experience with AI. This implies that being exposed generally to AI technologies doesn't necessarily lead to better abilities among individuals to distinguish genuine media from AI-generated ones. This discovery is evidence of earlier studies showing that knowing about AI tools doesn't necessarily enhance perceptual discrimination skills, especially in the realm of highly realistic synthetic media (Nightingale & Farid, 2022; Diel et al., 2024).

Likewise, there were no significant differences in the self-assessed detection ability for the second hypothesis as related to AI experience level. The participants from each experience group had staying power perceptions of their own ability to detect AI-generated content. The perceptions of participants within each experience group were relatively similar in terms of their capacity to detect AI-generated content. This could indicate the general uptake of information about AI-generated media by digitally active consumers regardless of their familiarity with AI.

Furthermore, findings showed, for the third hypothesis that self-assessed image detection ability was not significantly predictive of actual image classification ability and an overall self-assessed detection ability was not significantly predictive of overall classification accuracy. However, self-assessed video detection ability was found to have a positive and statistically significant association with actual video classification performance. This indicates that participants' metacognitive awareness might be slightly higher in judging dynamic AV stimuli than static ones, perhaps due to the presence of motion and synchronization cues provided by the videos.

In sum, the results indicate that the barrier to perceiving AI generated media is not necessarily unaware of or unfamiliar with it, but actually the ineffectiveness of perceptual judgment, when faced with increasingly sophisticated media productions. The findings reveal that individual confidence in one's self is insufficient to discriminate with accuracy and even those experienced with digital modes are vulnerable to highly realistic AI-produced media. The study represents an innovation into the literature by combining the components of classification accuracy, self-assessed detection ability and Signal Detection Theory analysis. Specifically, SDT enabled the study to surpass descriptive accuracy measures, as well as evaluating perceptual sensitivity and authenticity bias in audience reactions to AI-produced media. Additionally, the research enriches the currently scarce studies in the Arab context on these issues, and it reveals the need to develop a critical visual interpretation and specialized training in media verification to assist in engagement with AI-generated content.

8. Conclusion

This study examined the classification accuracy, self-reported detection capabilities, and levels of AI experience among individuals in images and videos, comparing real content to AI-generated content. The results revealed the participants' inability to accurately selectively differentiate between AI generated between actual media, especially in video content when compared to static/image content. Moreover, the participants had a small authenticity bias, showing a preference for authentic over AI-generated content.

Furthermore, the skill acquired as a result of self-reporting in relation to AI experience did not consistently improve participants' actual classification ability, nor their measures of detection ability. Most participants felt confident in their capacity to identify AI-created media, but there was a significant difference between participants' skills in actually recognizing the media. This indicates that there's a limit to how much one is exposed to AI and how much they understand about common indicators of AI content, and that might not be enough.

The principal finding of this research is that recognizing AI-generated media is not merely a technical challenge, but also a perceptual and cognitive process. The integration of classification accuracy, self-reports of detection ability, AI experience and Signal Detection Theory analysis allow the study to prove that the lack of human detection is linked with a lack of perceptual sensitivity and poor subjective judgment.

Based on these findings, a critical approach to dealing with risks posed by AI-generated contents should transcend beyond general awareness campaigns; instead, there should be an increasing demand for systematic visual literacy training that can help individuals to critically evaluate and be aware of their perceptual limits. Further studies are needed to examine the efficacy of specialized cognitive training interventions, cultural variations in recognizing AI-generated media content, and the combined methods of taking a hybrid approach that include human intervention and the use of automated verification tools. Finally, the research points out that with highly realistic generative AI, accurately distinguishing between real and fake versions has become a cognitive/perceptual issue, beyond just a technical issue.

9. Limitations

This study is subject to several methodological and contextual limitations. First, although considerable effort was made to balance the realism and quality of genuine and AI-generated stimuli, differences in visual fidelity between media types and generative models may have influenced participants' classification performance. In particular, the quality and realism of AI-generated videos varied depending on the capabilities and limitations of the generation tools used.

Second, since the study was conducted with Google Forms participants had to classify the tasks under uncontrolled viewing conditions. Screen sizes, screen type, screen resolution, internet connection and distractions could have affected perceptual judgements and detection performance. Furthermore, the online design did not provide the opportunity to confirm if participants used help or if content was accessed and replayed in other ways than intended in the procedure.

Third, the measures in the study involved some self-assessment questions for detecting types, which are subjective and could not truly assess the processes and strategies people use for actual classification. Therefore, participants' reported confidence/abilities could differ from their actual detection performance.

Fourthly, all the participants were Saudi undergraduate students, mainly from media related fields. Culturally, this population was a reasonable target for investigating interactions with digital media, but caring must be taken to ensure that the results are not broadly applicable to other demographic groups or cross-cultural groups.

Lastly, the cross-sectional design of the study constrains the generalization of the present results on possible changes in performance over time due to increased exposure, training or technological advances of generative AI systems. Data and research for better understanding of how perceptual sensitivity and detection skill is developed through more elaborate media produced by AI systems should be further explored with longitudinal and experimental methods.

Acknowledgments

Not applicable.

Author contributions

Not applicable.

Funding

This research received no external funding.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Informed consent

Obtained.

Ethics approval

The Publication Ethics Committee of the Redfame Publishing.

The journal's policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Data sharing statement

Not applicable.

References

Abu Zaid, M., Mohamed, Z., & Eid, A. (2023). Artificial intelligence and contemporary trends in fine arts: A descriptive and analytical study. *Journal of Fine Arts and Art Education*, 53–84. (in Arabic)

Adobe. (2023). *Firefly and responsible generative AI*. Adobe.

Al-Ghabbari, M., & Youssry, B. (2023). The role of artificial intelligence technologies in developing digital media: A future vision. *Arab Journal of Media and Communication Research*, 619–653. (in Arabic)

Alghamdi, N. S., & Alowibdi, J. S. (2024). Distinguishing Arabic GenAI-generated tweets and human tweets utilizing machine learning. *Engineering, Technology & Applied Science Research*, *14*(5), 16720-16726. https://doi.org/10.48084/etasr.8249

Al-Sawy, M. M. R. (2023). Greening the workplace: the power of a positive climate and leader influence on employee behavior*, Journal of Commercial Research,45*(3), 3-45.

Al-Tawkhi, I. (2024). The artificial intelligence revolution and its impact on creativity and design. *Heritage and Design Magazine*. (in Arabic)

Blomgren, E., & Harald, R. (2024). Age disparities in discerning AI-generated imagery. *KTH Royal Institute of Technology.*

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

Cooke, D., Edwards, A., Barkoff, S., & Kelly, K. (2025). As Good as a Coin Toss: Human Detection of AI-Generated Content. *Communications of the ACM*, *68*(10), 100-109. https://doi.org/10.1145/3729417

Diel, A., Lalgi, T., Schröter, I. C., MacDorman, K. F., Teufel, M., & Bäuerle, 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

Digital Education Council. (2024). *AI or not AI: What students want: Global AI student survey 2024*.

Dwivedi, Y. K., et al. (2023). Generative AI and the future of content creation. *International Journal of Information Management*, *68*, 102642.

Epstein, D., Jain, I., Wang, O., & Zhang, R. (2023). Online detection of AI-generated images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (pp. 382–392). Paris, France. https://doi.org/10.1109/ICCVW60793.2023.00045

Epstein, Z., Fang, C., Arechar, A., & Rand, D. (2023). What label should be applied to content produced by generative AI? *arXiv preprint*. https://doi.org/10.31234/osf.io/v4mfz

Fiedler, A., & Döpke, J. (2025). Do humans identify AI-generated text better than machines? Evidence based on excerpts from German theses. *International Review of Economics & Education*, 100321. https://doi.org/10.1016/j.iree.2025.100321

Floridi, L., et al. (2018). AI4People—An ethical framework for a good AI society. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Frank, J., Herbert, F., Ricker, J., Schönherr, L., Eisenhofer, T., Fischer, A., Durmuth, M. & Holz, T. (2024, May). A representative study on human detection of artificially generated media across countries. In *2024 IEEE Symposium on Security and Privacy (SP)* (pp. 55-73). IEEE. https://doi.org/10.1109/SP54263.2024.00159

Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2022). Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, *119*(1), e2110013119. https://doi.org/10.1073/pnas.2110013119

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ideis, T. (2024). The reality of artificial intelligence (AI) and its applications in Arab countries (Challenges and recommendations). *Journal of Arts, Literature, Humanities and Social Sciences*, *112*, 364–381. https://doi.org/10.33193/JALHSS.112.2024.1241

Kemp, S. (2025, March 3). *Digital 2025: Saudi Arabia*. DataReportal.

Lu, Z., Huang, D., Bai, L., Qu, J., Wu, C., Liu, X., & Ouyang, W. (2023). Seeing is not always believing: Benchmarking human and model perception of AI-generated images. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*. https://doi.org/10.52202/075280-1105

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide* (2nd ed.). Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410611147

Martin-Rodriguez, F., Garcia-Mojon, R., & Fernandez-Barciela, M. (2023). Detection of AI-created images using pixel-wise feature extraction and convolutional neural networks. *Sensors*, *23*(9037). https://doi.org/10.3390/s23229037

Mertala, P., Lopez-Pernas, S., Vartiainen, H., Saqr, M., & Tedre, M. (2024). Digital natives in the scientific literature: A topic modeling approach. *Computers in Human Behavior*, *152*, 108076. https://doi.org/10.1016/j.chb.2023.108076

Miller, E., Steward, B., Witkower, Z., Clare, S., Krumhuber, E., & Dawel, A. (2023). AI hyperrealism: Why AI faces are perceived as more real than human ones. *Psychological Science*, *34*(12), 1390–1403. https://doi.org/10.1177/09567976231207095

Mohammed ALrashoud, L., & Ali Althubaiti, H. (2025). Exploring assessment for learning and assessment as learning in Saudi EFL classrooms. *Journal of the Faculty of Education, 91*(1), 1-25. https://doi.org/10.21608/mkmgt.2025.357883.1893

Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, *119*(8), e2120481119. https://doi.org/10.1073/pnas.2120481119

OpenAI. (2021, January 5). *DALL·E: Creating images from text*. https://openai.com/index/dall-e/

OpenAI. (2024). *Video generation models as world simulators*.

Pinski, M., & Benlian, A. (2024). AI literacy for users: A comprehensive review and future research directions. *Computers in Human Behavior: Artificial Humans*, *2*(1), 100062. https://doi.org/10.1016/j.chbah.2024.100062

Rouxel, A. (2020). AI in the media spotlight. In *Proceedings of the 2nd International Workshop on AI for Smart TV Content Production* (pp. 1–2). ACM. https://doi.org/10.1145/3422839.3423059

UNESCO. (2024, May 13). *UNESCO supports Arab States in developing AI competencies for students and teachers*.