Assessing Public Ability to Distinguish AI-Generated from Real News: Accuracy, Confidence, and Influencing Factors

Khalid A.H. Alazawi
Open University Malaysia (OUM)

Nantha Kumar Subramaniam
Open University Malaysia (OUM)

ABSTRACT
The rapid advancement of generative artificial intelligence (AI) has intensified concerns about the spread of highly convincing synthetic news. This study examines the public's ability to distinguish between real and AI-generated news, investigates the misalignment between confidence and actual performance, and identifies the demographic, behavioural, and technological factors that influence detection accuracy. A total of 382 participants completed an online survey containing one real and one AI-generated news article; after rigorous preprocessing to remove bot-generated, inattentive, and uniform ("lazy") responses, 210 valid cases were analysed. Results reveal a significant detection challenge: only 8% of respondents accurately identified both articles, while 32% failed to correctly classify either one. Despite these low accuracy levels, confidence was disproportionately high, with approximately 62% reporting that they were "mostly confident" or "fully confident" in their judgments. This confidence–accuracy mismatch highlights a critical cognitive vulnerability that may amplify susceptibility to misinformation. Regression analyses further show that commonly assumed protective factors—such as education level, age, and news-checking frequency—do not reliably predict the ability to detect AI-generated content. Only technological proficiency displayed a meaningful positive correlation with performance, although the effect was modest. These findings challenge traditional assumptions about digital literacy and indicate that demographic attributes alone cannot safeguard users against sophisticated AI-driven deception. Instrument reliability was strong (Cronbach's α = .89; Composite Reliability = .80), affirming the stability of the measures used to assess credibility judgments. The implications of this study underscore the urgent need for redefined digital literacy frameworks that emphasizes critical reading, linguistic awareness, and metacognitive regulation. Technological interventions, such as AI-based detection tools and transparency mechanisms for synthetic content, are also necessary to complement user education. The study concludes that in the age of generative AI, human judgment alone is insufficient to ensure news authenticity, and coordinated efforts across education, platform design, and policy are essential to preserve information integrity.

General Terms
Artificial Intelligence; Human–Computer Interaction

Keywords
AI-generated misinformation; News credibility detection

1. INTRODUCTION
Recent advancements in artificial intelligence (AI) have fundamentally reshaped modern information ecosystems, particularly in societies increasingly dependent on digital technologies. Generative AI tools now allow users to produce text, images, audio, and video with ease, generating outputs that are frequently indistinguishable from human-created media [2]. These systems can produce large quantities of content at extraordinary speed—often within seconds—and at negligible cost, making them widely accessible to the general public [3]. While such capabilities have clear benefits for creativity and productivity, they also present significant risks related to misinformation, authenticity, and public trust.

One of the most pressing concerns associated with generative AI is its potential to amplify the production and dissemination of digital misinformation. High-quality, tailored AI-generated content can be easily misused to fabricate highly convincing false information and distribute it rapidly across social media platforms, which predominantly rely on text, images, and video. Once disseminated, such misinformation can remain unchecked, influencing public perceptions and disrupting civic discourse [6]. The effortless creation and viral spread of false content enhance the capacity for coordinated disinformation campaigns, enabling malicious actors to manipulate societal narratives and erode trust in institutions and communities [8].

Empirical evidence further underscores the severity of this issue. Research shows that individuals often cannot reliably distinguish between authentic news and AI-generated misinformation, especially when false content aligns with pre-existing cognitive biases [5]. Deep-faked media has been observed to influence attitudes even when users are explicitly warned about its credibility [4]. AI-generated deception has already affected high-stakes political processes, including events such as the U.S. 2016 election (Kertysova, 2018). Even efforts to counter misinformation through manual or automated detection face substantial limitations; AI-generated content can be produced at volumes that far exceed the capacity of existing verification systems, allowing deceptive material to spread widely and rapidly.

Although some theorised that AI-generated news might appear more neutral or objective, studies indicate that people tend to trust AI-generated news and AI reporters less than their human counterparts [7]. Nevertheless, the sophistication of AI-generated misinformation continues to escalate, raising concerns about future scenarios in which superintelligent AI systems may produce deceptive content that is even more persuasive and difficult to detect. Scholars warn that such developments could complicate verification efforts and intensify societal risks [1].

Despite the growing influence of generative AI, the academic investigation of its misinformation-related risks remains limited. Further, public availability of advanced text-generators such as ChatGPT began only in late 2022, making this phenomenon both recent and insufficiently studied. Given the escalating threat of AI-driven deception, there is a critical need to examine how individuals perceive AI-generated news, how

effectively they can distinguish authentic from synthetic content, and which factors influence their susceptibility to AI-driven misinformation.

2. OBJECTIVE OF THE STUDY

The objectives of the study are listed below:

1- To evaluate users' ability to distinguish between real news and AI-generated news;
2- To examine their level of confidence in making such judgments; *and*
3- To determine how personal demographics, technological proficiency, news habits, and media trust relate to individuals' ability to identify AI-generated news

3. METHODOLOGY

The methodology for this study employed a quantitative research design to evaluate the believability and risks of AI-generated misinformation. A web-based survey was chosen as the primary research instrument due to its cost-effectiveness, wide reach, and suitability for collecting measurable data from diverse respondents. The survey consisted of two sections: demographic questions and two news-evaluation sections featuring one real and one AI-generated article.

A non-probability convenience sampling method was adopted, drawing respondents from MTurk, SurveySwap, with an intended sample size of 200–385 to achieve an acceptable margin of error. Prior to full deployment, two pilot surveys were conducted to refine the instrument and address technical issues. Data collection spanned six weeks and targeted adults from various backgrounds. After collecting responses, the data underwent cleaning, validation, and statistical analysis, including descriptive statistics and simple linear regression, to test relationships between demographic factors, confidence levels, and the ability to distinguish real from AI-generated news. The structured methodological approach ensured systematic investigation, validity of measurement, and alignment with the study's objectives.

The survey items were systematically developed to align with the study's objectives. The researcher first identified the key constructs required to assess the believability of AI-generated news, namely credibility, trustworthiness, accuracy, neutrality, presentation quality, and respondents' confidence. Guided by these constructs, a series of rating-scale items using a five-point Likert scale were designed to capture participants' perceptions in a consistent and quantifiable manner

The survey was divided into two structured sections, each serving a specific analytical purpose. Demographic items (section 1 of the survey) were created to collect background data on factors hypothesized to influence the ability to detect misinformation—such as age, gender, education level, field of study, tech proficiency, news-checking frequency, preferred news sources, and trust levels. These items were intentionally simple and categorical to allow reliable statistical analysis (e.g., regression, correlation) (Table 1).

Table 1. Survey items for demographic profile

	Item
1	What is your age? (Numerical answer)
2	What is your gender? (Male, Female, Prefer not to say)
3	What country/state do you currently live in? (Short answer)
4	What is your highest level of education attained? (High school, Diploma, Bachelor's, Master's, PhD)
5	What was your major field of study? (Short answer)
6	On a scale of 1-5, how would you describe your tech proficiency? (Low, beginner, intermediate, advanced, expert)
7	On a scale of 1-5, how frequently do you check the news? (Never, rarely, occasionally, once a day, multiple times a day)
8	What is your preferred source of news? (None, social media, traditional media, AI-generated summaries, independent media)
9	On a scale of 1-5, how much do you trust AI generated news? (1- Not at all, 2- Little trust, 3- Neutral, 4- Mostly trust, 5- Fully trust)

News evaluation (section 2 of the survey) items were designed by pairing one real news article and one AI-generated article. The researcher created identical sets of eight questions for each article to enable direct comparison of participants' perception of real vs. AI-generated content. These items captured dimensions such as believability, trust in source, perceived author expertise, factual accuracy, bias, professionalism, willingness to share, confidence in judgment, and final classification of real vs. fake as shown in Table 2.

Table 2. Survey item for news evaluation

	Item
1	Would you say that this news was fake news or real news? (fake/genuine)
2	On a scale of 1-5, how credible (believable) do you find this news article?
3	On a scale of 1-5, how much do you trust the source of this article?
4	On a scale of 1-5, how knowledgeable is the author of this news article?
5	On a scale of 1-5, how factually accurate does the information in this news article seem to be?
6	On a scale of 1-5, how neutral and unbiased do you find this news article to be?
7	On a scale of 1-5, how professional and well-organized do you find this news article to be?
8	On a scale of 1-5, how confident are you when you answer questions about this article?

Throughout the development process, the researcher ensured that all items were closed-ended, which minimizes ambiguity, encourages response completeness, and facilitates quantitative analysis. Before final deployment, the items were tested through two pilot surveys to check clarity, technical accuracy, and response validity. Feedback from these pilots informed final adjustments, ensuring that the survey items captured the intended constructs accurately and aligned with the overarching research objectives.

4. RESULTS

A total of 382 individuals participated in the survey. Following data preprocessing—which involved removing bot-generated entries, inattentive respondents, and uniform "lazy" responses—210 valid cases were retained for analysis (55% of the original dataset). Bots accounted for the largest portion of invalid responses (33%), followed by respondents who failed the attention check (7%) and those who submitted non-informative repeated answers (5%). The final sample produced an acceptable sampling error of 6.77%. Respondents represented a broad age range. The largest group was aged 20–29 (37%), followed by 30–39 (27%), 50+ (16%), and 40–49 (15%), with only 5% under the age of 20, indicating underrepresentation of teenagers and early undergraduates. A total of 61% of respondents were male and 39% were female, and the sample was generally well educated, with 66% holding at least a bachelor's degree.

4.1 Accuracy in Identifying AI-Generated and Real News

Participants demonstrated substantial difficulty distinguishing authentic news from AI-generated content. Only 8% accurately identified both news articles, while 32% misidentified both. The remaining 60% correctly identified only one article, indicating inconsistent performance and widespread confusion between real and AI-generated material.

4.2 Confidence–Accuracy Relationship

In this study, respondents' confidence was assessed immediately after they evaluated each news article using six credibility-related indicators—namely, how believable the article was, how trustworthy its source appeared, how knowledgeable the author seemed, how factually accurate the content appeared, how neutral and unbiased the writing was, and how professional and well-organized the article seemed (Table 2). Using a five-point Likert scale (where 1 = not supportive and 5 = very supportive), the mean scores for all evaluated items (items 2 to 7 of Table 2) across both news articles ranged from 3.22 to 3.86, indicating generally supportive and favourable assessments from the respondents. For all six questions, the average respondent rated the AI-generated news higher than the human-written news.

After completing these six judgments, participants were asked to rate how confident they felt in answering these questions (item 8 of Table 2), using a five-point Likert scale ranging from "zero confidence" to "fully confident." When the confidence ratings for both articles were combined, 61.9% of respondents identified themselves as either "mostly confident" or "fully confident." However, actual performance data showed a stark contrast: only 8% of respondents correctly identified both the real and AI-generated articles, while 32% failed to identify either one correctly. This substantial disparity demonstrates that respondents' perceived confidence in evaluating credibility-related attributes was inflated and poorly aligned with their actual detection ability, thus empirically supporting the hypothesis that confidence is an unreliable predictor of performance in distinguishing real news from AI-generated misinformation.

4.3 Correlation Analysis Using Simple Regression

This section evaluates the strength of the linear relationships between respondents' characteristics and their ability to distinguish human-written news from AI-generated news. Simple linear regression was performed for each independent variable against the normalized score. Because each regression model contained only one predictor, the coefficient of determination (R^2) is used as an indicator of the magnitude of the linear association. In a simple regression model, R^2 represents the squared correlation coefficient (r^2), and therefore provides a valid basis for comparing relationship strength across variables. While R^2 does not capture directionality, it is appropriate for assessing the relative influence of each factor on respondents' detection ability.

Age

Age groups were categorised into five levels to align with other Likert-type variables. The resulting regression yielded a weak linear association between age and normalized score (R^2 = 0.48). Although minor differences in performance were observed across age categories, the results suggest that age does not meaningfully account for variation in respondents' ability to identify AI-generated content.

Education Level

The association between education level and detection accuracy was also weak (R^2 = 0.461). Despite varying sample sizes across education groups, the linear relationship between education and normalized score remained limited. This result indicates that higher educational attainment did not correspond to increased accuracy in distinguishing authentic news from AI-generated news.

Technological Proficiency

Technological proficiency showed the strongest linear association with detection accuracy among all variables tested. The model produced a high R^2 value (0.834), indicating that proficiency with technology explains a substantial portion of variance in respondents' performance. This highlights technological competency as a key factor influencing detection capability.

Frequency of Checking the News

A moderate association was observed between news-checking frequency and normalized score (R^2 = 0.669). Respondents who checked the news more frequently tended to perform slightly better, although the magnitude of the relationship suggests only partial explanatory power.

Frequency of Encountering Fake or False News

The regression model for perceived frequency of encountering fake or false news demonstrated a very weak linear association (R^2 = 0.241). Participants who believed they frequently encountered false information did not perform significantly better in identifying AI-generated content.

Trust in Traditional News

Trust in traditional news sources exhibited a moderate degree of association with detection accuracy (R^2 = 0.635). Respondents with higher levels of trust in traditional media tended to achieve slightly higher normalized scores. However, this association must be interpreted alongside the associations observed for other news trust variables, as considered in the concluding summary.

Trust in AI-Generated News Summaries

Trust in AI-generated news summaries had the lowest observed association with detection accuracy (R^2 = 0.123). This indicates that participants who expressed greater trust in AI-generated summaries did not demonstrate enhanced ability to differentiate between AI-generated and human-written content.

Trust in Social Media

Trust in social media as a news source showed a moderate linear association with detection accuracy (R^2 = 0.69).

However, this result should be interpreted cautiously due to the small number of participants who reported very high trust levels. Nonetheless, the overall pattern suggests a moderate positive association.

Trust in Independent News
Trust in independent news sources produced a moderate association with detection accuracy ($R^2 = 0.655$). Participants who trusted independent outlets tended to perform somewhat better at distinguishing AI-generated content.

4.4 Instrument Reliability and Validity

A factor analysis conducted on all Likert-scale items of the survey indicated strong psychometric robustness. The measurement instrument demonstrated high internal consistency and construct validity, as evidenced by Cronbach's $\alpha = .89$, Composite Reliability = .80, and an Average Variance Extracted (AVE) of .579.

5. DISCUSSION

The findings of this study highlight significant challenges in the public's ability to distinguish real news from AI-generated misinformation. Despite the rapid advancement of generative models and their increasing presence in digital information ecosystems, this study reveals a substantial gap between individuals' perceived and actual ability to detect AI-generated content. Only a small minority of respondents (8%) accurately distinguished both articles, while nearly one-third misclassified both. These results suggest that AI-generated misinformation can convincingly mimic the style and characteristics of authentic journalism, reinforcing concerns raised in emerging misinformation literature.

A striking outcome of this research is the confidence–accuracy mismatch. Although most respondents expressed high or very high confidence in their judgments, their performance showed considerable inaccuracy. This overconfidence phenomenon echoes previous research showing that individuals who are least equipped to identify false news content are also the least aware of their own limitations and, therefore, more susceptible to believing it and spreading it further [9]. In the context of AI-generated misinformation, this overconfidence may increase susceptibility to false information, as individuals may not seek verification when they believe they already possess sufficient judgment capabilities. Thus, the findings reinforce the need for digital literacy interventions that address not only skills but also metacognitive awareness.

The regression analysis further provides nuanced insights into which user characteristics predict better discrimination ability. Technological proficiency emerged as the only strong and consistent predictor, suggesting that familiarity with digital systems may offer a partial protective factor. However, this relationship was still insufficient to materially improve performance for most respondents. Education level—often assumed to be a strong driver of critical thinking—did not correlate significantly with the ability to identify AI-generated news. Similarly, frequently checking the news or exhibiting higher skepticism offered limited benefits. Collectively, these results indicate that traditional assumptions about who is "more resilient" to misinformation may no longer hold in the age of sophisticated AI content generation.

The findings also reveal that trust in traditional media does not substantially enhance detection ability. Although partially supported, this correlation was weak and inconsistent. This may reflect broader global trends in which trust in legacy news organisations is insufficient to safeguard individuals from deceptive content encountered in digital spaces. Because AI-generated articles can emulate journalistic cues such as tone, structure, and citation style, users may rely on superficial features rather than content verification when forming judgments.

Overall, this study contributes to a growing body of research emphasizing the urgent need for targeted interventions to mitigate AI-driven misinformation risks. The findings suggest that simple awareness or higher education alone will not equip individuals to adequately identify deceptive AI content. Future efforts should focus on developing structured digital literacy programmes, enhancing platform-level detection tools, and exploring how users can be trained to critically evaluate content beyond surface-level cues. As AI-generated misinformation becomes increasingly indistinguishable from genuine journalism, protecting public information ecosystems will require coordinated strategies involving educational institutions, technology developers, and policymakers.

6. CONCLUSION

This study set out to evaluate the public's ability to distinguish real news from AI-generated misinformation and to identify the factors that influence this judgment. The findings demonstrate that AI-generated content poses a significant threat to information integrity, as the majority of respondents were unable to reliably differentiate between authentic and fabricated articles. Despite relatively low accuracy levels, participants expressed high confidence in their decisions, indicating a substantial misalignment between perceived and actual competence. This overconfidence represents a critical vulnerability, as individuals may unknowingly accept or share misinformation while believing they are exercising sound judgment.

Regression analyses showed that demographic and behavioural factors traditionally associated with critical information evaluation—such as education level, age, and news consumption habits—did not reliably predict the ability to detect AI-generated news. Only technological proficiency exhibited a consistent positive relationship with performance, suggesting that familiarity with digital systems provides a modest but insufficient advantage. Collectively, the results highlight the limitations of relying solely on individual characteristics to understand misinformation resilience in the era of generative AI.

This research underscores an urgent need to rethink digital literacy frameworks, which have traditionally focused on evaluating source credibility and identifying overt signs of manipulation. As AI-generated misinformation becomes more linguistically sophisticated and stylistically indistinguishable from real journalism, detection will require new strategies—ranging from enhanced critical reading skills to stronger platform-level safeguards. By demonstrating the scale of misclassification and the psychological overconfidence that accompanies it, this study contributes important evidence to the growing scholarly concern surrounding AI-driven misinformation and the vulnerabilities it introduces into public information ecosystems.

7. IMPLICATIONS, LIMITATIONS AND FUTURE RESEARCH

The findings of this study carry important implications for education, technology development, and policy formulation. From an educational perspective, traditional digital literacy

programmes are no longer sufficient in an era where AI-generated misinformation closely mimics authentic journalism. Curricula must evolve to incorporate deeper linguistic, contextual, and narrative analysis skills, alongside metacognitive training to help learners recognise and regulate overconfidence in their judgments. On the technological front, platforms and news aggregators should adopt stronger AI-based detection mechanisms and implement clearer labelling systems for synthetic content. Relying solely on user judgment—particularly in light of the confidence–accuracy mismatch observed—poses significant risks. At the policy level, governments and regulators may need to establish frameworks ensuring transparency in AI-generated content, including watermarking, provenance tracking, and accountability structures to curb misuse.

Despite its contributions, the study is subject to several limitations. The sample under-represented teenagers and early undergraduates, which restricts generalisability to younger populations who are frequent consumers of digital content. Additionally, the survey-based design introduces potential self-report biases, such as social desirability and self-overestimation, although the large confidence–accuracy gap suggests that these biases did not obscure the central findings. The use of only two news articles limits the diversity of content styles assessed, meaning that results may differ with other topics, genres, or narrative formats. Finally, the prevalence of bot-generated responses—even with strict filtering—highlights a methodological challenge inherent to conducting online surveys in environments where AI-driven automation is widespread.

These limitations provide clear directions for future research. Expanding the variety of content tested—including multiple AI models, diverse topics, and different media formats—would offer a broader understanding of how misinformation is perceived. Longitudinal studies could examine whether training, repeated exposure, or platform safeguards improve individuals' detection abilities over time. Experimental designs that manipulate cues, warnings, or structural features of articles would help isolate the elements that assist or hinder recognition of AI-generated content. Cross-cultural comparisons may reveal how cultural norms, linguistic differences, and media ecosystems influence susceptibility to misinformation. Finally, future studies should explore the cognitive and emotional mechanisms underlying judgment, including how biases, heuristics, and affective responses shape the interpretation of AI-generated news.

8. REFERENCES

[1] Baum, S. D. (2018). Countering superintelligence misinformation. *Information, 9*(10), 244.

[2] Dunn, C., Hunter, J., Steffes, W., Whitney, Z., Foss, M., Mammino, J., … Nathoo, R. (2023). Artificial intelligence–derived dermatology case reports are indistinguishable from those written by humans: A single-blinded observer study. *Journal of the American Academy of Dermatology*.

[3] Galanter, P. (2019). Artificial intelligence and problems in generative art theory. In *Proceedings of EVA London 2019* (pp. 112–118).

[4] Hughes, S., Fried, O., Ferguson, M., Hughes, C., Hughes, R., Yao, X. D., & Hussey, I. (2021). *Deepfaked online content is highly effective in manipulating people's attitudes and intentions* (No. FERMILAB-PUB-21-182-T). Fermi National Accelerator Laboratory. https://doi.org/10.31234/osf.io/4ms5a

[5] Kreps, S., McCain, R. M., & Brundage, M. (2022). All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science, 9*(1), 104–117.

[6] Liu, Y. (2023). *Analysis on the effects of misinformation: Taking Facebook as an example* (Unpublished undergraduate thesis). Boston University, College of Communication.

[7] Longoni, C., Fradkin, A., Cian, L., & Pennycook, G. (2022, June). News from generative artificial intelligence is believed less. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 97–106). https://doi.org/10.1145/3531146.3533077

[8] Stanton, B., & Jensen, T. (2021). *Trust and artificial intelligence*. National Institute of Standards and Technology, U.S. Department of Commerce.

[9] Lyons, B. A., Montgomery, J. M., Guess, A. M., Nyhan, B., & Reifler, J. (2021). *Overconfidence in news judgments is associated with false news susceptibility*. *Proceedings of the National Academy of Sciences, 118*(23), e2019527118. https://doi.org/10.1073/pnas.2019527118