International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 11 | Nov 2026 www.irjet.net p-ISSN: 2395-0072

IMPACT OF AI HALLUCINATIONS ON HUMAN TRUST AND DECISION-MAKING IN AI-BASED SYSTEMS

Aryan Keni[1], Aditya Chavan[2], Dr. Suman Saurabh[3]

[1] *Department of MCA, D. Y. Patil College of Humanities and Sciences, D. Y. Patil Deemed to be University, Navi Mumbai, Maharashtra, India*

[2] *Associate Professor, D. Y. Patil Deemed to be University, School of Humanities and Sciences, Belapur, Maharashtra, India.*

Abstract - *Artificial Intelligence (AI) systems are increasingly used to obtain information, generate content, support learning, and assist with decision-making. However, generative AI systems may produce incorrect, misleading, or fabricated information that appears convincing and authoritative. This phenomenon is commonly known as AI hallucination. Such outputs may influence how users perceive the reliability of AI systems, how much they rely on AI-generated information, and how they make decisions. This research examines the impact of AI hallucinations on human trust and decision-making in AI-based systems. The study focuses on perceived trustworthiness, reliance on AI-generated information, decision-making accuracy, and users' intention to verify information independently. A quantitative, scenario-based study is planned, using a structured questionnaire to collect primary data from college students and young adults with experience using AI tools. Data will be analysed using descriptive statistics and suitable inferential tests after collection.*

Keywords - AI Hallucinations, Human Trust, Decision-Making, Generative AI, AI Reliability, Verification Behaviour.

1. INTRODUCTION

Artificial Intelligence (AI) has become widely used in education, healthcare, business, research, and everyday activities. Generative AI tools can answer questions, generate content, summarise information, write computer programs, and support decision-making. However, these systems sometimes produce incorrect, misleading, or fabricated information that appears convincing. This phenomenon is known as AI hallucination.

AI hallucinations are a significant concern because users may find it difficult to distinguish accurate information from false or unsupported claims. When an AI system presents incorrect information confidently, users may accept it without verification. This can affect their trust in AI, their willingness to rely on AI-generated responses, and the quality of their decisions.

Trust is an important factor in human–AI interaction. Excessive trust may lead users to depend on inaccurate information, while repeated exposure to errors may reduce their confidence in AI systems. Therefore, understanding how hallucinations influence trust, reliance, and verification behaviour is essential for responsible AI use.

This research examines the impact of AI hallucinations on human trust and decision-making. It focuses on users' trust in AI-generated responses, their reliance on these responses, decision-making accuracy, and intention to verify information independently. A quantitative, scenario-based study will be used to collect and analyse responses from college students and young adults who have experience using AI tools.

1.1 Background of the Study

Generative AI systems produce responses based on patterns learned from data. Although their answers may be fluent and relevant, fluency does not guarantee factual accuracy. AI hallucinations can therefore create challenges for users who depend on these systems for information and decision support.

1.2 Importance of the Study

This study examines the human consequences of AI hallucinations. Its findings may help users improve AI literacy, encourage educational institutions to promote responsible AI use, and guide developers in making AI systems more transparent and reliable.

2. PROBLEM STATEMENT

Generative AI systems can produce incorrect or fabricated information while presenting it in a convincing manner. Users may accept such information without verifying its accuracy, potentially affecting their trust, reliance on AI, and decision-making.

Although existing research examines AI hallucinations and trust in automated systems, it is important to understand how users respond to inaccurate AI-generated information. This study investigates the relationship between AI hallucinations, human trust, reliance on AI-generated information, decision-making accuracy, and independent verification behaviour among college students and young adults.

3.1 AI Trust and Decision-Making (2024)

Klingbeil et al. (2024) [10] investigated trust and reliance on AI advice through an experiment. Their findings showed that participants sometimes followed AI advice even when it conflicted with available information, highlighting the risk of overreliance.

Wang and Ding (2024) [11] examined the effects of explanations in AI-supported decisions. Their study found that explanations could improve prediction accuracy in certain conditions, although improvements in users' reported trust were not always significant.

3.2 Trust and Reliance on AI (2025)

Research published in 2025 examined the relationship between trust, distrust, and appropriate reliance on explainable AI systems. The review highlighted that making an AI system explainable does not automatically guarantee appropriate user trust [12].

3.3 AI Hallucinations and Human Trust (2026)

Cheng et al. (2026) [13] investigated the relationship between AI hallucinations, human–AI trust, and collaboration. Their study examined how hallucinations may affect collaboration through changes in trust, making it directly relevant to the present research.

Pearson et al. (2026) [14] studied human reliance on AI in decision-making using an experiment involving judgments of real and AI-generated faces. Their research highlights the importance of understanding how people respond to AI advice when making judgments.

Overall, recent studies show that AI reliability, user trust, and reliance are closely connected. The present study builds on this research by examining how hallucinated AI responses relate to trust, decision-making accuracy, and verification behaviour among college students and young adults.

4. RESEARCH GAP

Previous studies have examined AI hallucinations, trust in automation, and the risks of large language models. However, further research is needed to understand how hallucinated responses influence human trust, reliance on AI, decision-making accuracy, and verification behaviour together. This study aims to examine these factors among college students and young adults who use AI tools.

5. RESEARCH QUESTIONS

1. RQ1: Is users' trust significantly different when they evaluate accurate AI responses compared with hallucinated responses?
2. RQ2: Does exposure to hallucinated responses change users' willingness to rely on AI-generated recommendations?
3. RQ3: Does the presence of hallucinated information reduce the accuracy of decisions made by users?
4. RQ4: Does exposure to hallucinated responses increase users' intention to independently verify AI-generated information?

5. RESEARCH OBJECTIVES

1. To examine the relationship between AI hallucinations and human trust.
2. To investigate users' reliance on AI-generated information.
3. To assess decision-making accuracy when users encounter hallucinated responses.
4. To examine users' intention to verify AI-generated information.
5. To recommend practices for responsible AI use.

7. RESEARCH HYPOTHESES

H1 (Trust): Users' mean trust score will be significantly lower for hallucinated AI responses than for accurate AI responses.

H2 (Reliance): Users' willingness to rely on AI-generated information will be significantly lower when the response contains hallucinated information.

H3 (Decision accuracy): The proportion of correct decisions will be significantly lower after users receive hallucinated AI responses than after accurate AI responses.

H4 (Verification): Users' intention to verify information independently will be significantly higher after exposure to hallucinated responses than after accurate responses.

8. RESEARCH METHODOLOGY

8.1 Research Design

A quantitative, within-subject, scenario-based design will be used. Each participant will evaluate matched decision scenarios containing either an accurate AI response or a response with a controlled factual hallucination. The order of the scenarios will be randomized or counterbalanced to reduce order effects. Participants' trust, reliance, decision accuracy, and verification intention will be compared across the two conditions.

8.2 Sample Size and Population

The target participants will be college students and young adults who have experience using AI tools. The proposed sample size is approximately 150 participants.

8.3 Sampling Technique

Convenience and purposive sampling will be used to recruit participants who are accessible and have experience with AI tools.

8.4 Data Collection Method

Primary data will be collected through a structured questionnaire using Google Forms or Microsoft Forms. Participants will review matched scenarios and answer the same set of questions after each scenario. Trust, reliance, and verification intention will be measured using multiple five-point Likert items, while decision accuracy will be scored against a predefined answer key. The questionnaire will also record basic usage experience with AI tools.

8.5 Variables of the Study

The independent variable will be response condition (accurate AI response or hallucinated AI response). The dependent variables will be (i) trust score, (ii) willingness-to-rely score, (iii) decision accuracy, and (iv) verification-intention score. Prior AI-use experience may be recorded as a background variable for exploratory subgroup analysis.

8.6 Data Collection Procedure

Participants will first receive study information and provide informed consent. They will then evaluate the scenarios in a randomized or counterbalanced order and complete the same outcome items for each condition. Decision responses will be scored using an answer key prepared before data collection. Responses will be screened for completeness, coded, and analysed using the procedures described in Section 9. Participants will be debriefed about the intentionally inaccurate scenario content.

8.7 Ethical Considerations

Participation will be voluntary, and responses will be kept confidential. The collected data will be used only for academic research. Any intentionally inaccurate scenarios will be low-risk and explained appropriately to participants.

9. DATA ANALYSIS AND HYPOTHESIS TESTING

The analysis will compare responses from the accurate and hallucinated conditions for the same participants. Microsoft Excel will be used for data cleaning, coding, descriptive summaries, and charts; SPSS, R, or Python may be used for inferential testing. The analysis plan is specified below. Since primary data have not yet been collected, no empirical results or statistical conclusions are reported in this paper.

9.1 Descriptive Statistics

For each condition, frequency and percentage will summarize categorical responses, while mean and standard deviation (or median and interquartile range where appropriate) will summarize trust, reliance, and verification scores. Decision accuracy will be reported as the number and percentage of correct responses. Tables and comparative bar charts will display the differences between accurate and hallucinated conditions.

9.2 Reliability Analysis

Where trust, reliance, or verification are measured using multiple questionnaire items, internal consistency will be assessed separately for each scale using Cronbach's alpha. An alpha value of approximately 0.70 or higher will generally be treated as acceptable for exploratory research. Items will be combined into a scale score only when their coding and reliability support doing so.

9.3 Inferential Statistics

Because each participant evaluates both response conditions, paired tests will be used. A paired-samples t-test will compare condition means for scale scores when the paired-difference scores are reasonably consistent with the test assumptions. If assumptions are not met, the Wilcoxon signed-rank test will be used. McNemar's test will compare paired correct/incorrect decision outcomes. Effect sizes and 95% confidence intervals will be reported alongside p-values. Pearson or Spearman correlation may be used as an exploratory analysis of the relationship between trust and reliance.

9.4 Hypothesis Testing

Each hypothesis will be tested using a two-sided significance level of 0.05. For H1, H2, and H4, the paired condition scores will be compared using the appropriate paired test. For H3, paired decision-accuracy outcomes will be compared using McNemar's test. The null hypothesis will be rejected when the adjusted p-value is below 0.05; otherwise, it will not be rejected. Because four primary hypotheses are tested, the Holm procedure will be applied to control the family-wise error rate. Statistical significance will be interpreted together with effect sizes and confidence intervals.

9.5 Illustrative Statistical Calculations

The following calculations illustrate the planned analysis only. The values are hypothetical examples and are not findings from this study. Actual statistics will be calculated

only after the questionnaire data have been collected and checked.

A. Mean

The mean is calculated by dividing the sum of all observations by the number of observations.

Mean = Sum of observations / Number of observations
For example, if five participants give trust scores of 2, 3, 3, 4, and 3:

$$\text{Mean} = (2 + 3 + 3 + 4 + 3) / 5 = 3.00$$

B. Percentage

Percentage = (Number in a category / Total number of responses) × 100

For example, if 30 out of 50 participants select an option:

$$\text{Percentage} = (30 / 50) \times 100 = 60\%$$

C. Paired-Samples t-Test

For a scale score measured under both conditions, the paired-samples t-test evaluates whether the average within-participant difference is different from zero. The difference score is calculated for each participant by subtracting the accurate-condition score from the hallucinated-condition score (or vice versa, consistently).

The test statistic is:

t = Mean of paired differences / (Standard deviation of paired differences / $\sqrt{n}$)

Here, n is the number of participants with complete paired observations and the degrees of freedom are n − 1. If the paired-difference assumptions are not suitable, the Wilcoxon signed-rank test will be considered instead.

The test produces a p-value and confidence interval for the mean paired difference. The direction and size of the difference will be interpreted in relation to each hypothesis.

D. Decision Accuracy and Hypothesis Decision

- McNemar's test will be used for paired binary outcomes (correct or incorrect) to determine whether the proportion of correct decisions differs between accurate and hallucinated conditions.
- For all primary tests, $p < 0.05$ after Holm adjustment will indicate statistical evidence against the corresponding null hypothesis. If $p \geq 0.05$, the null hypothesis will not be rejected; this does not prove that the conditions are equivalent.

The actual test statistics, p-values, effect sizes, confidence intervals, and hypothesis decisions will be reported only after the study data have been collected and analysed. No values will be estimated or presented as observed results in advance.

E. Association Analysis

An exploratory correlation analysis will examine whether participants with higher trust scores also report greater willingness to rely on AI-generated information. Pearson's correlation will be used when its assumptions are reasonable; otherwise, Spearman's rank correlation will be preferred. This association will be interpreted as a relationship, not evidence of causation.

9.6 Hypothesis-to-Test Mapping

Table I links each research hypothesis to its outcome measure, comparison, and planned statistical test. Trust, reliance, and verification intention will be scored consistently across both conditions. Decision accuracy will be scored as correct or incorrect using an answer key prepared before data collection.

Hypothesis	Outcome measure	Comparison	Planned test
H1	Trust scale score	Accurate vs hallucinated	Paired t-test / Wilcoxon
H2	Reliance scale score	Accurate vs hallucinated	Paired t-test / Wilcoxon
H3	Decision correct/incorrect	Paired binary outcomes	McNemar test
H4	Verification-intention score	Accurate vs hallucinated	Paired t-test / Wilcoxon

9.7 Illustrative Graphs (Not Actual Findings)

The following figures demonstrate how the results may be displayed after data collection. The plotted values are hypothetical examples created only to explain the proposed analysis. They are not responses from participants, and they must be replaced with the actual survey results before submission as a completed empirical study.

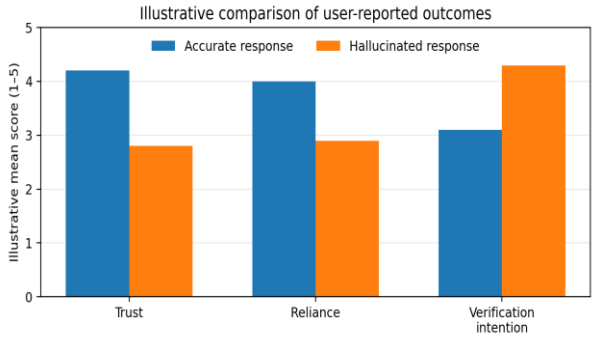

Fig. 1. Illustrative mean scores for trust, reliance, and verification intention (hypothetical values only).

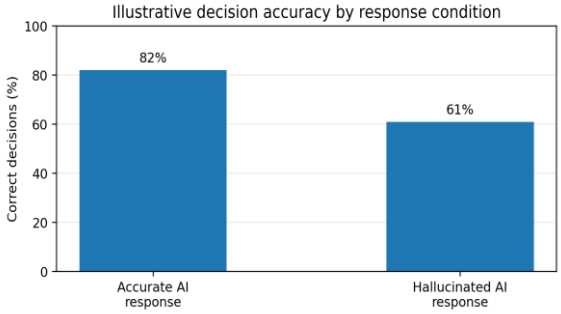

Fig. 2. Illustrative decision accuracy under accurate and hallucinated response conditions (hypothetical values only).

9.8 Illustrative Numerical Summary

For demonstration only, a hypothetical dataset might show a lower average trust score after hallucinated responses and a higher intention to verify information. A similar illustrative pattern is shown for decision accuracy in Fig. 2. These values are not estimates of the expected effect and must not be described as observed outcomes.

Outcome	Accurate condition	Hallucinated condition	Illustrative direction
Trust (1–5)	4.2	2.8	Lower after hallucination
Reliance (1–5)	4.0	2.9	Lower after hallucination
Verification intention (1–5)	3.1	4.3	Higher after hallucination
Decision accuracy	82%	61%	Lower after hallucination

9.9 Interpretation and Reporting of Results

If H1 is supported, the findings would indicate that participants report lower trust in hallucinated responses than in accurate responses. If H2 is supported, participants would be less willing to rely on hallucinated information. Support for H3 would indicate a difference in the proportion of correct decisions across conditions. Support for H4 would indicate a change in participants' intention to verify information independently. Each conclusion must be based on the actual test statistic, adjusted p-value, effect size, and confidence interval.

A non-significant result should be reported as insufficient evidence of a difference, not as proof that the two response conditions are identical. Results should also be interpreted in light of the scenario wording, participant familiarity with AI, sample composition, and the limitations of self-reported measures.

10. TOOLS AND TECHNOLOGIES

Google Forms or Microsoft Forms will be used to collect questionnaire responses. Microsoft Excel will be used to organise data and calculate descriptive statistics. Suitable statistical software, such as SPSS, R, or Python, may be used for hypothesis testing and further analysis. Reliable sources will be used to verify the accuracy of the AI-generated scenarios.

11. EXPECTED FINDINGS

The study is expected to provide insights into how hallucinated AI responses relate to users' trust, reliance, decision-making accuracy, and verification behaviour. It may show whether users respond differently to accurate and inaccurate AI-generated information. However, the actual findings will depend on the collected data and statistical analysis.

12. RECOMMENDATIONS

1. **Verify AI-generated information:** Users should cross-check important claims using reliable and independent sources.
2. **Promote AI literacy:** Educational institutions should teach students about AI hallucinations, factual accuracy, and responsible AI use.
3. **Improve AI transparency:** Developers should clearly communicate system limitations and uncertainty.
4. **Provide reliable sources:** AI systems should provide verifiable references where appropriate.
5. **Encourage human judgement:** Users should evaluate AI-generated advice critically instead of relying on it blindly.

6. **Strengthen evaluation:** AI systems should be tested for factual accuracy and their ability to communicate uncertainty.

13. CONCLUSION

AI hallucinations present a challenge to the reliability of generative AI systems. Incorrect information may influence users' trust, reliance, and decisions. This study aims to examine these effects through a quantitative, scenario-based approach. Its results may help encourage independent verification, appropriate trust, and responsible use of AI-based systems.

14. LIMITATIONS OF THE STUDY

The study may be limited by its sample size, convenience-based sampling, and reliance on questionnaire responses. Participants' reported behaviour may differ from their actual behaviour. The results may not represent all AI users, and conclusions will depend on the final research design and collected data.

15. FUTURE SCOPE

Future research could include larger and more diverse samples, compare different AI systems, and examine whether AI literacy training improves users' ability to identify hallucinations. Long-term studies could also investigate how repeated exposure to incorrect AI responses changes trust and decision-making over time.

16. REFERENCES

[1] Z. Ji et al., "Survey of hallucination in natural language generation," ACM Computing Surveys, vol. 55, no. 12, 2023.

[2] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 1906–1919.

[3] P. Manakul, A. Liusie, and M. J. F. Gales, "SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models," in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 9004–9017.

[4] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," Human Factors, vol. 46, no. 1, pp. 50–80, 2004.

[5] K. A. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," Human Factors, vol. 57, no. 3, pp. 407–434, 2015.

[6] A. Glikson and A. W. Woolley, "Human trust in artificial intelligence: Review of empirical research," Academy of Management Annals, vol. 14, no. 2, pp. 627–660, 2020.

[7] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," Human Factors, vol. 39, no. 2, pp. 230–253, 1997.

[8] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 610–623.

[9] R. Bommasani et al., "On the opportunities and risks of foundation models," arXiv preprint arXiv:2108.07258, 2021.