What Drives the Rise in Risk Perception? —— A Study on the Mechanisms through Which Risk Perception Influences Users' Willingness to Fact-Check in the Context of AI Hallucinated Information

Yujun Liu

College of Communication Science and Art, Chengdu University of Technology, Chengdu, 610000, China
202410010215@stu.cdut.edu.cn

Abstract. AI Hallucinations are harming the reliability of information and public trust. Although previous studies hafigve explored technical countermeasures, the path by which users' risk perceptions are transformed into a desire for fact-checking is not yet known. Based on psychological reactance theory, this paper explores the mediating effect of psychological reactance on the path between risk perception and a willingness to fact-check, as well as the moderating role of initial trust. A survey was conducted on 141 adult users of AI with more than three months of use. Based on the results, risk perception positively correlates with a desire for fact-checking. Psychological reactance fully mediates this link (indirect effect = 0.24, 95% CI [0.12, 0.38]). The initial trust positively moderates the impact of risk perception on reactance (interaction beta=0.21, $p<0.05$). The research shows that a person's feeling of risk can lead to a psychological need to check the facts, thus supporting demand-side governance of AI hallucinations and trust adjustment.

Keywords: AI hallucination; Risk perception; Psychological reactance; Fact-checking intention; Initial trust

1 Introduction

Generative artificial intelligence is spreading and, at the same time, changing how information is created and spread. However, in the absence of reliable factual support, the "hallucinated" information produced by AI models is increasingly plausible in appearance yet inconsistent with reality or logically flawed; thus, the authenticity of the information ecosystem and public trust are being damaged [1-2]. The three types of spread for AI hallucinations identified in the above research are: logical inconsistencies, error masking, and temporal fluctuations [1, 2]. Technical mitigation solutions have also been proposed, such as retrieval-augmented generation and model fine-tuning [3-4].

Technical methods cannot eliminate hallucinations entirely [4]. Therefore, by addressing the psychological factors of the end-users of the information, we can help

G. E. Atalay and A. Wang (eds.), *Proceedings of the 2026 3rd International Conference on Public Relations and Media Communication (PRMC 2026)*, Advances in Social Science, Education and Humanities Research 1032,
https://doi.org/10.2991/978-2-38476-609-3_88

overcome the shortcomings in technical governance. Psychological response mechanisms for users, as the end users and evaluators of the information, have not been widely studied by scholars.

Research by previous scholars has also found a positive correlation between a user's perception of risk and their desire for fact-checking [5,6]. However, the mediating path of why risk perception prompts individuals to verify remains a "black box". Psychological reactance theory has mainly been applied in the fields of health communication and advertising persuasion, and its application to the phenomenon of AI hallucinations has not been empirically studied. Therefore, exploring how risks are perceived psychologically and how this translates into behaviour, such as verification behaviour by users, has provided some supporting theoretical and practical references.

Based on the above, the current study focuses on scenarios of AI hallucinations to construct a mediating path model of "risk perception→psychological reactance →willingness to verify facts", and also investigates the moderating effect of users' initial trust in AI. A questionnaire survey will be conducted to gather data from older users of AI in their daily lives. Mediating and moderating effects are analyzed using the PROCESS macro (Models 4 and 1), and hypotheses are tested via bias-corrected bootstrapping (5,000 resamples). This paper will provide some evidence for the demand-side governance of AI hallucinations and the calibration of human-AI trust.

2 Literature Review

2.1 The Conceptual Definition and Generation Mechanism of AI Hallucination

AI hallucination is the problem of generating false information by a generative AI model that lacks a factual basis or is self-contradictory [1, 2, 5]. Recently, some scholars abroad have begun to study this issue. Firstly, a delusion is a mental disorder of the mind. Blom referred to it as "a non-veridical perceptual experience in the absence of external stimuli" [7]. Hallucination in the early days of computer vision referred to the filling in of missing information [8]. However, with the development and popularisation of large language models, the term has gradually taken on a negative connotation and now refers specifically to outputs that are factually incorrect or ungrounded [9].

Based on previous studies, AI hallucinations generally refer to cases where large language models, without access to reliable factual evidence, generate content that seems reasonable but is in fact inconsistent with objective reality, logical reasoning or context. These are the forms of errors mentioned in [1, 2, 5]: factual errors, fabricated citations, etc. Maleki and others have pointed out that the word "hallucination" may lead people to believe, without intent, that the AI system is conscious. They indicated that it can be viewed as a foreseeable consequence of the "probability-driven generation" mode [9]. Typologically speaking, scholars have generally divided AI hallucinations into two categories: factual hallucinations (contradicting objective facts) and faithfulness hallucinations (not in accordance with user instructions or context) [3, 10].

2.2 Generation Mechanism: Technical Layer and User Layer

Technical Layer Factors. First, the sources of the training data are contaminated. Large Language Models are based on a vast amount of internet text, which includes various false and biased information. After indiscriminate learning, the model encodes the wrong information as language patterns [2,11]. Second, a model structure with limited attention will result in a loss of context. The model also has a weak boundary recognition capability for knowledge. With sparse training data, the model may focus more on generating fluent-sounding output rather than accurate information [2,10].

User Layer Factors. Li and Li put forward the theory of "media equivalence": people treat technical media as if they were social, attribute a thinking capacity to AI, and thus generate a cognitive projection and trust illusion [1]. Therefore, users are more likely to be fooled by AI hallucinations and are therefore at a higher risk of harm.

Generative AI is now in widespread use, and hallucinations by the AI often occur. Most of the current studies focus on technical governance and have paid little attention to the psychological reasons and ways in which users verify information. Given the above, this paper will use users as the subjects and explore the mediating effect of risk perception on the willingness to fact-check via psychological reactance. And verifies whether the starting trust is a moderator. Theoretically, this study has addressed the empirical deficiency in research on AI hallucinations by focusing on users rather than technology. In practice, it offers concrete support for the demand-side governance of AI hallucinations, calibration of human-machine trust, and literacy education.

3 Information Ecosystem Risks Triggered by AI Hallucinations

3.1 Transmission Characteristics

Comprehensive studies have shown that information produced by AI hallucinations generally has the following three characteristics in its spread [1-2].

Logical Consistency. False content is in line with the regular laws of language and therefore appears very credible.

Error Concealment. Multimodal AI-generated content is close to real life and difficult for the public to distinguish as false.

Temporal Dynamics. Due to the time lag in the model parameter update and the synchronisation of the underlying knowledge base, the same question may have different logical results at different times. Due to this dynamic inconsistency, it is more challenging for users to make an informed judgment and thus makes the problem of correcting misinformation at the level of the information ecosystem worse. It is a reason for a crisis of trust in the system.

3.2 Social Impact

Damage to society has also occurred due to AI hallucinations. Zhang and Shen employed LDA topic model analysis to identify the three types of risk indicators: data manipulation, lack of openness, and reduction in public discourse [2]. Based on foreign studies, the harms of AI hallucinations in various aspects of life are now quite serious.

The court has created AI-generated fake case citations. Legal Disputes Have Arisen in the Commercial Sector Over Chatbot Misinformation. Low-quality AI-generated false information has also spread widely in the information environment [1, 2]. These cases show that AI hallucinations are eroding the credibility of information in many parts of society and thus causing a crisis of trust. Therefore, we need to learn about the psychology of users better. Psychological reactance is used as the mediator in this paper to explain how increased risk perception leads to changes in behaviour via psychological reactance. It offers a scientific foundation for designing low-cost, highly resonant user-side intervention strategies and helps to turn the public from passive recipients into active verifiers.

4 Research on User Risk Perception and Information Verification Behavior

4.1 Risk Perception

Risk perception refers to an individual's subjective judgment of the potential negative consequences, such as the probability and seriousness of a risk [12]. In the context of AI hallucinations, this idea has been extended to include the whole range of expected cognitive biases, errors in judgment, and loss of social credibility that may result from the spread of false information [1,2]. Risk perception in the case of AI hallucination includes the accuracy of the information, but also other factors such as the expectation of cognitive bias, decision-making errors and social embarrassment [1,2]. Given that it is still challenging to eliminate hallucinations entirely with technology at present [3, 4], understanding the psychology of users can help to address the deficiencies in technical control.

4.2 Intention to Fact-check

The Goal of verification is a type of user behavior that arises from AI hallucinations. Research shows that people who think they are in more danger are more likely to check the facts themselves [5,13]. Most of the research so far has focused on teaching information literacy and has not studied the psychological process by which a sense of risk changes into a desire to verify. Rahman et al. have shown that most of the research on fact-checking in large language models has focused on internal verification mechanisms for the models, leaving the psychological motivations for user-initiated fact-checking largely unaddressed [13].

4.3 Theoretical Explanatory Framework

Psychological Reactance Theory (Brehm, 1966). According to this theory, when people feel that their freedom to behave has been restricted externally, they are motivated to restore this freedom and thus exhibit a sense of opposition [14]. According to the above theory of information behaviour, people are not likely to accept the persuasion in the content. Provide a theoretical basis for applying it to the problem of AI hallucination response. In cases of AI hallucination, if users believe that the information may lead to misinterpretation and thus restrict their cognitive autonomy, then psychological

reactance will occur and a demand for verification will emerge.

Rogers' Protection Motivation Theory . In addition to the above, the two-factor theory explains the emergence of protective behaviours as a result of the dual appraisal of threat severity and coping efficacy [15]. The two theories complement each other to show the psychological pathway of risk perception → psychological reactance → verification behaviour.

5 Comparison of Chinese and International Research and Research Gaps

5.1 Differences in Research Orientations: China and Abroad

Earlier in the international academic community, research has been conducted on both the technical level and the social influence of hallucination detection, law and medicine in application areas. Most of the research in China has focused on technological governance and macro-policy analysis [1, 2, 5]. Empirical studies from the user's perspective are also lacking [16]. As the primary subjects of information reception and dissemination, the psychological response mechanisms of users to AI hallucinations have not been fully studied. Most of the psychological processes of how users perceive, judge and respond to hallucinations have been treated as secondary. On the one hand, for a long time, research on AI hallucination has been led by computer science and technology to develop engineering solutions for model detection and data filtering. At the same time, the existing research on risk has focused mainly on the accuracy of the information and legal consequences; systematic studies of changes in users' psychological states in the face of risk are lacking.

5.2 Three Gaps in Existing Research

Lack of empirical research from the users' end. Most of the studies so far have focused on the technology and system level, and there is a lack of research into the psychological response mechanism of users.

The links of risk perception and verification behaviour are not yet known. The Black Box of why risk perception prompts people to verify has not been fully opened. Whether psychological reactance serves as a mediator in this way has not been determined empirically [16].

Research on the Moderating Effect of Individual Differences. The initial trust in AI and user experience may affect the activation and conversion efficiency of risk perception. Related research has been relatively sporadic [1,5].

6 The Entry Point and Research Hypothesis of This Study

Based on the above deficiencies, this paper studies the mechanism through which users' risk perceptions affect their intention to engage in fact-checking of AI-generated hallucinations. Psychological reactance theory is used as the foundation of explanation, a mediating path model of "risk perception → psychological reactance → willingness

to fact-check" is constructed, and the moderating effect of users' initial trust in AI is investigated.

The pre-survey N=37 showed that 89.19% of users had used AI for more than three months, but only 5.41% strongly agreed that AI information was trustworthy. In addition, 64.86% of users worried that AI's undoubted answers might contain inaccurate content, and 45.95% were willing to search for information verification. This paradox of high-frequency use but questionable trust signals the tangible risks that AI hallucinations introduce for ordinary users. It also reflects that the psychological mechanism between risk perception and actual action remains unclear, thus highlighting the theoretical value and practical significance of this study.

The following are the hypotheses of this study:

H1 (Main effect): A higher perceived risk of AI-generated hallucinations by a user is associated with a stronger desire for verification.

H2 (Mediator effect): Psychological reactance is a mediator in the path of risk perception to intention to verify.

H3 (Moderating effect): A person's initial trust in AI is expected to change the extent to which risk perception leads to psychological reactance; that is, highly trusting individuals may be more susceptible to psychological reactance under high-risk circumstances.

This paper intends to explore why people are affected psychologically by artificial intelligence's false information, and in doing so, offer new references for the demand-side regulation of AI-generated hallucinations.

7 Data Analysis Methods

7.1 Data Preprocessing and Reliability Testing

SPSS 26.0 was employed to organise and analyze the data gathered in this study. Given that the experience of AI use may affect people's perception and evaluation of hallucination, this study set a condition of at least three months of AI use for inclusion and excluded 15 samples of participants who had used AI for less than three months or never used it. A total of 156 questionnaires were distributed, and after excluding invalid ones, 141 were collected; the actual collection rate was 90.4%. An internal-consistency test was performed on the measurement items of the core variables (risk perception, psychological reactance, willingness to fact-check, initial trust). Cronbach's α coefficient for the risk perception scale (three items) was 0.82; for the psychological reactance scale (three items) it was 0.79; for the willingness to fact-check scale (two items) it was 0.84; and for the initial trust scale (two items) it was 0.76. All exceeded the recommended cutoff of $\alpha = 0.70$, and thus were considered reliable.

7.2 Descriptive Statistics and Correlation Analysis

Mean and standard deviation of all the above were calculated. Pearson product-moment correlation analysis was used to investigate the correlations among risk perception, psychological reactance, willingness to fact-check, and initial trust, and some initial data for subsequent hypothesis testing were obtained.

7.3 Hypothesis Testing

Main Effect Test (H1). A single linear regression was used to study the effect of risk perception on the level of willingness to fact-check. This investigated the direct effect of risk perception on the will to fact-check.

Mediation Effect Test (H2). Hayes' PROCESS macro (Model 4) was employed to examine whether risk perception led to a decrease in willingness to fact-check through psychological reactance, thereby conducting a mediation analysis. Resampling 5,000 times with bias correction in the percentile bootstrap method was used to obtain the 95% confidence interval of the indirect effect. If the interval did not include 0, then mediation was considered to have taken place.

Test of Moderating Effect (H3). PROCESS macro (Model 1) was used to examine the moderation effect of initial trust on the "risk perception → psychological reactance" path; thus, initial trust was added as a moderator. Simple slope analysis was used to determine whether there were significant differences in the impact of risk perception on psychological reactance at various levels of trust (mean ± 1 standard deviation).

8 Results

8.1 Descriptive Statistics and Correlation Analysis

The means, standard deviations, and correlation coefficients of each variable are shown in Table 1. Risk perception was significantly positively correlated with willingness to fact-check $r=0.46, p<0.01$, providing preliminary support for H1. Risk perception was also significantly positively correlated with psychological reactance $r=0.53, p<0.01$. Psychological reactance was significantly positively correlated with willingness to fact-check $r=0.49, p<0.01$. Initial trust was weakly negatively correlated with risk perception ($r = -0.11$, $p > 0.05$), but significantly negatively correlated with psychological reactance ($r = -0.28$, $p < 0.01$). This indicates that users with higher trust levels exhibited lower psychological reactance when confronted with hallucinatory information.

Table 1. Means, Standard Deviations and Correlation Matrix of Variables (N = 141)

Variable	M	SD	1	2	3	4
1. Risk Perception	3.78	0.67	1			
2. Psychological Reactance	3.51	0.72	0.53**	1		
3. Willingness to Fact-Check	3.70	0.81	0.46**	0.49**	1	
4. Initial Trust Level	3.40	0.74	-0.11	-0.28**	-0.15	1

** P < 0.01**

8.2 Main Effect Test

With risk perception as the independent variable and willingness to fact-check as the dependent variable, the results showed that risk perception had a significant positive predictive effect on fact-checking intention $\beta=0.46, t=6.11, p<0.001$, explaining 21% of the variance $R^2=0.21$. These results support H1.

8.3 Mediating Effect Test of Psychological Reactance

The mediating effect of psychological reactance was tested by PROCESS Model 4, and the results are shown in Table 2. The positive effect of risk perception on psychological reactance was significant $\beta=0.53, p<0.001$. When risk perception and psychological reactance were included in the model to predict fact-checking willingness, the effect of psychological reactance was significant $\beta=0.34, p<0.01$, while the direct effect of risk perception was no longer significant $\beta=0.18, p=0.082$. Bootstrap analysis showed that the indirect effect of psychological reactance was 0.24, with a 95% confidence interval of [0.12, 0.38], excluding 0. These findings indicate that psychological reactance fully mediates the relationship between risk perception and willingness to fact-check, thereby supporting H2.

Table 2. Mediation Effect Test Results (PROCESS Model 4)

Path	Effect Value	Boot SE	95% CI	Conclusion
Risk Perception → Psychological Reactance (a)	0.53***	0.07	[0.39, 0.67]	Significant
Psychological Reactance → Willingness to Fact-Check (b)	0.34**	0.11	[0.12, 0.56]	Significant
Risk Perception → Willingness to Fact-Check (c‘)	0.18	0.10	[-0.02, 0.38]	Not significant
Indirect Effect (a×b)	0.24	0.07	[0.12, 0.38]	Significant

*** $p < 0.001$, ** $p < 0.01$

8.4 Test of the Moderating Effect of Initial Trust

Using risk perception as the independent variable, initial trust as the moderating variable, and psychological reactance as the dependent variable, the moderating effect was tested with PROCESS Model 1. The results (Table 3) indicate that the interaction between risk perception and initial trust significantly predicted psychological reactance $\beta=0.21, p<0.05$. Simple slope analysis showed that for users with low initial trust (mean −1 SD), the positive effect of risk perception on psychological reactance was weaker $\beta=0.32, p<0.05$. For users with high initial trust (mean +1 SD), the positive effect of risk perception on psychological reactance was stronger $\beta=0.65, p<0.001$. Thus, initial trust positively moderates the relationship between risk perception and psychological reactance, supporting H3.

Table 3. Results of the Moderating Effect Test (PROCESS Model 1)

Predictor Variable	β	SE	t	p
Risk Perception	0.48	0.09	5.33	<0.001
Initial Trust Level	-0.22	0.08	-2.75	<0.01
Risk Perception × Initial Trust Level	0.21	0.09	2.33	0.021

8.5 Supplementary Analysis: The Impact of Demographic Variables

An independent samples t-test was conducted to examine gender differences. The results showed no significant differences between males and females in risk perception $t=0.64, p=0.52$, psychological reactance $t=0.37, p=0.71$, or willingness to fact-check ($t = -0.85$, $p = 0.40$). One-way ANOVA indicated that users with different usage frequencies (1-2 times per week, 3-5 times per week, daily use) also showed no significant differences in the core variables ($p > 0.05$). Due to the limited sample size (especially the male subgroup with only 66 participants), these null results should be interpreted with caution.

8.6 Comprehensive Interpretation of Effect Sizes

The main effect in this study had a Cohen's f^2 of 0.27 and was considered a medium effect ($f^2 \geq 0.02$ small, ≥ 0.15 medium, ≥ 0.35 large). The fully standardised value of the indirect effect was 0.21 and was therefore considered a small-to-medium effect. Therefore, it can be expected that there will be some effect of risk perception on the willingness to fact-check via psychological reactance, and other unobserved factors may still be influencing this phenomenon.

8.7 Path Diagram (Standardized Coefficients)

The results of the path analysis are shown in Fig. 1 below.

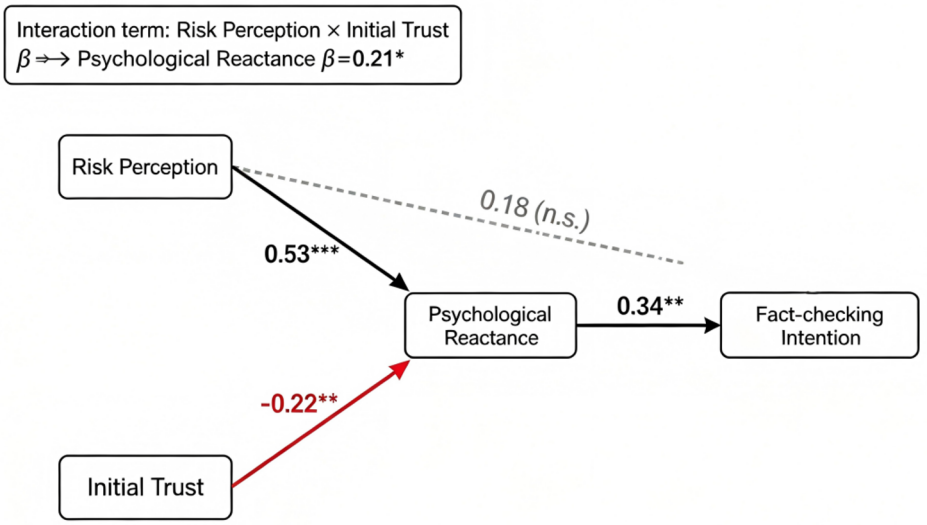

Fig. 1. Path Diagram with Standardized Coefficients (Photo credit:Original)

The above are the main results.

Main path: The positive effect of risk perception on psychological reactance is significant ($\beta = 0.53$, $p < 0.001$), and the negative effect of psychological reactance on the willingness to fact-check is also significant ($\beta = -0.22$, $p < 0.001$); thus, risk perception reduces users' willingness to fact-check through psychological reactance.

Direct effect: The direct path from risk perception to the willingness to fact-check is not statistically significant ($\beta = 0.18$, $p > 0.05$), and thus psychological reactance likely acts fully as a mediator.

Moderating effect: The interaction term between risk perception and initial trust has a significant effect on psychological reactance ($\beta = 0.21$, $p < 0.05$); that is to say, initial trust positively moderates the relationship between risk perception and psychological reactance. The first stage of trust has been shown to increase people's willingness to resist behaviours negatively.

9 Discussion

9.1 Discussion of Results and Comparison with Literature

Explanation of Main Effects and Mediation Effects. This study shows that users' sense of risk for AI hallucinated information is positively correlated with how willing they are to fact-check. This result is consistent with the conclusions of previous research that risk perception drives the behaviour of information verification [5,13]. Psychological reactance also provided a full mediator for the above two. Therefore, the users' intention to fact-check after encountering a risk of AI hallucination is not directly correlated with the actual risk. Risk perception leads to a sense of psychological reactance in users;

thus, their perceived autonomy has been reduced and they feel that their cognitive freedom has been violated. Brehm's theory of psychological reactance is modified in this paper to extend its application to the field of human-computer interaction and AI misinformation scenarios [14], departing from the traditional topics of persuasive communication and health intervention. Notably, this study identifies risk perception as a new type of antecedent to psychological reactance and thus extends the theoretical system on the cognitive conditions that trigger reactance.

In line with the view of Li and Li that AI hallucinations may alter the public's cognitive system via algorithmic dissemination [1], this study's empirical data also show that users are not simply affected passively but are motivated to correct the errors by actively confirming the information. Pei et al. have shown that information literacy can motivate people to verify the accuracy of information [5]; however, they have not explored the role of specific psychological states in the impact of risk perception on this behaviour. By demonstrating that psychological reactance fully mediates the connection between risk perception and the willingness to check facts, this study adds to the body of knowledge by revealing a previously unknown route of risk perception → psychological reactance → verification behaviour.

Explanation of the Moderating Effect. Based on the above results, initial trust can enhance the impact of perceived risk on psychological reactance. Users with a high degree of trust were more psychologically reactance-prone when facing the risk of AI hallucinations. The above results support H3 and are consistent with the analysis by Zhang and Shen on the sense of trust betrayal for high-trust users misled by AI [2]. Low-trust users are naturally more skeptical about the information provided by AI and have lower expectations for the accuracy of AI output. Even when they are aware of the risk of hallucination, their expectation gap has not increased significantly from the previous level. Therefore, the additional effect of risk perception on psychological reactance is weaker. Therefore, a person's original trust in AI is a reason for emotion in a particular case.

Findings Do Not Meet Expectations. Gender and use frequency were not significantly different in the main attributes of this study. Not all studies of social media have found differences in behaviour by gender. It is plausible that the application scenarios for AI tools are relatively neutral (e.g., information retrieval, content generation) and do not involve gender-sensitive issues such as appearance comparison on social media. The number of men in this study was also relatively small (66), so statistical power may have been reduced, and some differences could not be found.

9.2 Practical Recommendations

Based on the results above, this paper puts forward the following practical suggestions. Because users with a high sense of trust are more likely to experience psychological reactance when perceiving risks, AI developers should avoid using an all-knowing authority-type output style. They should indicate how confident they are of the information in their answers and provide supporting source links. Reduce the high level of distrust among users due to a loss of faith.

Second, literacy education for users. Literacy education should teach fact-checking

abilities and help people learn to be more critical of the information they encounter through scenario-based teaching. Users should be able to recognise the different kinds of risks in advance and not have a general fear of AI. Thus, these individuals can trigger psychological reactance in response to AI hallucinations and thus exhibit verification behavior. Intervention for high-trust users should focus on preventing excessive reactance due to a loss of trust. Inform the students about the shortcomings of AI beforehand and reduce their expectations accordingly. For low-trust users, the intervention should help them improve their accurate risk perception and increase the efficiency of transforming verification motivation to prevent information-avoidance behaviour due to generalised suspicion.

Thirdly, platform governance. The platform can build a dynamic prompt system that uses user behaviour signals, such as repeatedly viewing the same content, typing skeptical keywords like "Is this reliable?" or "Really?", or staying on the generated result page for an unusually long time, as proxy indicators of high-risk perception. At this time, trigger fact-checking tools and other sources promptly to take remedial measures. This will be a human-AI collaborative verification system to reduce the spread of false information.

9.3 Research Limitations and Future Directions

Research Limitations. First, with 85.9 per cent of the sample aged 18-35, the generalizability of the results is restricted. Caution should be exercised in the extension to the middle-aged and older groups. Second, this study is a cross-sectional survey that cannot determine the direction of causality among the variables and thus is unable to exclude the problem of reverse causality (for example, people who are more willing to verify may also be more risk-sensitive). In the future, laboratory experiments can be conducted to randomly divide participants into a high-risk perception group (which reads cases of AI hallucinations causing serious consequences) and a low-risk perception group (which reads cases of normal use). The above experiments will examine whether there are differences in psychological reactance and the will to fact-check in the following tasks, thereby testing the causal relationships proposed by the current model.

Third, the single-case situation material is the remarriage controversy of Li Qingzhao. Different kinds of AI hallucinations (e.g., factual hallucinations and faithfulness hallucinations) may lead to various degrees of psychological reactance, but this study did not distinguish among them. Fourthly, although the reliability of the two-item scale was within an acceptable range, it was relatively low (initial trust $\alpha = 0.76$), and thus might not be accurate.

Future Directions. First, future studies should employ a longitudinal design or an experimental method to verify the direction of causality more strictly. Second, expand the sample coverage to include people of all ages and walks of life. Thirdly, we will also study the difference in psychological reactance under various situations of AI hallucination. Fourthly, other mediating variables such as cognitive dissonance and information insecurity should be examined in connection with psychological reactance, and competitive or chain mediation effects are also to be explored. Fifth, more

comprehensive multi-dimensional scales need to be developed to improve the reliability of the core constructs.

10 Conclusion

This study empirically examined the mediating mechanism by which risk perception affects willingness to fact-check through psychological reactance, using a questionnaire survey N=141 and path analysis. The results indicate that users‘ risk perception of AI hallucinations positively influences their willingness to fact-check through the full mediating effect of psychological reactance. Initial trust positively moderates the effect of risk perception on psychological reactance. This finding reveals a general principle: in contexts of AI misinformation, users do not passively accept false information but transform risk perception into proactive verification behavior through the motivational state of psychological reactance. Psychological reactance plays the role of a cognitive alarm – risk perception triggers users' vigilance toward threats to cognitive freedom, which in turn drives them to take counteractive actions.

Therefore, the following are the practical suggestions proposed in this paper. AI product design should appropriately indicate the confidence of the information and avoid overly authoritative responses to prevent trust breakdown among users who are already highly trusting. User literacy education should focus on promoting the construction of cognitive vigilance and psychological reactance. The platform will issue prompts for human-machine collaborative verification of facts dynamically.

Provide empirical support for understanding the psychological mechanisms of users' responses to AI hallucinations and extend the scope of application for psychological reactance theory in human-computer interaction. Future research can use longitudinal studies or experimental methods to further investigate the causality of the phenomena; increase the size of the sample to enhance generalizability outside the original population; and explore other mediators such as cognitive dissonance and information insecurity for comparison or chain effects. Compare the psychological reactions to different kinds of AI hallucinations and offer targeted interventions based on these comparisons.

References

1. Li J K, Li Y: Trust crisis and governance paths of information ecology driven by AI hallucination. Issues Studies, (7) (2025)
2. Zhang M H, Shen W Q: Misaligned escalation and structural disorder: Information fog risks dominated by AI hallucination and the possibility of classified governance. Media Forum, (10), 53–63 (2025)
3. Huang L et al: A survey on hallucination in large language models. ACM Transactions on Information Systems, **43**(2), 1–55 (2025)
4. Alansari A, Luqman H: Large language models hallucination: A comprehensive survey. arXiv, 2510.06265 (2025)

5. Pei H J, Guo R Y, Yang Z H: Information hallucination in AI language generation for libraries: Risk avoidance strategies from the perspective of information literacy. Library and Information Guide, **9**(12), 36–43 (2024)
6. Zhou T, Zhang C L, Deng S L: Research on intermittent interruption behavior of AIGC users based on C-A-C. Modern Information, (3) (2025)
7. Blom J D: A dictionary of hallucinations. Springer, Berlin (2009)
8. Baker S, Kanade T: Hallucinating faces. Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition, 83–88. IEEE, Piscataway (2000)
9. Maleki N, Padmanabhan B, Dutta K: AI hallucinations: A misnomer worth clarifying. 2024 IEEE Conference on Artificial Intelligence, 1–6. IEEE, Piscataway (2024)
10. Ji Z, Lee N, Frieske R et al: Survey of hallucination in natural language generation. ACM Computing Surveys, **55**(12), 1–38 (2023)
11. Liu Z Y, Wang P J, Song X B et al: A review of hallucination problems in large language models. Journal of Software, **36**(3), 1152–1185 (2025)
12. Slovic P: Perception of risk. Science, **236**(4799), 280–285 (1987)
13. Rahman M M et al: Hallucination to truth: A review of fact-checking in LLMs. Artificial Intelligence Review, **59**, 70 (2026)
14. Brehm J W: A theory of psychological reactance. Academic Press, London (1966)
15. Rogers R W: A protection motivation theory of fear appeals and attitude change. Journal of Psychology, **91**(1), 93–114 (1975)
16. Zhang H Z, Ren W J: Beyond the “second self”: Human-AI dialogue based on trust in large language model applications. Journalism University, (3) (2024)