Conference Name: MBP 2024 Tokyo International Conference on Management & Business Practices, 18-19 January
Conference Dates: 18-Jan- 2024 to 19-Jan- 2024
Conference Venue: TKP Ichigaya Conference Center, Building 2F, 8 Ichigaya Hachiman-cho, Shinjuku-ku, Tokyo 162-0844
Appears in: PEOPLE: International Journal of Social Sciences (ISSN 2454-5899)
Publication year: 2024

Kim et. al., 2024

Volume 2024, pp. 92-101

DOI- https://doi.org/10.20319/icssh.2024.92101

This paper can be cited as: Kim, H and Lee, S.W. (2024). Investigating the Effects of Generative-AI Responses on User Experience After AI Hallucination. MBP 2024 Tokyo International Conference on Management & Business Practices, 18-19 January, 2024. Proceedings of Social Science and Humanities Research Association (SSHRA), 2024, 92-101.

INVESTIGATING THE EFFECTS OF GENERATIVE-AI RESPONSES ON USER EXPERIENCE AFTER AI HALLUCINATION

Hayoen Kim
Yonsei University, Seoul, South Korea
hy1107@yonsei.ac.kr

Sang Woo Lee
Yonsei University, Seoul, South Korea
leesw726@yonsei.ac.kr

Abstract

The integration of generative artificial intelligence (GenAI) systems into our daily lives has led to the phenomenon of "AI hallucination," where AI produces convincing yet incorrect information, undermining both user experience and system credibility. This study investigates the impact of AI's responses, specifically appreciation and apology, on user perception and trust following AI errors. Utilizing attribution theory, we explore whether users prefer AI systems that attribute errors internally or externally and how these attributions affect user satisfaction. A qualitative methodology, featuring interviews with individuals aged 20 to 30 who have experience with conversational AI, has been employed. Respondents preferred AI to apologize in hallucination situations and to attribute the responsibility for the error to the outside world. Results show that *transparency in error communication is essential for maintaining user trust, with detailed explanations. The research contributes to the understanding of how politeness and attribution strategies can influence user engagement with AI and has significant implications for AI development, emphasizing the need for error communication strategies that balance transparency and user experience.*

Keywords

Generative Artificial Intelligence, Hallucination, Politeness Strategy, Attribution Theory

1. Introduction

As generative AI(GenAI) becomes increasingly integrated into daily life, the phenomenon known as "AI hallucination," where AI generates plausible yet incorrect or misleading information, has emerged as a critical concern (Athaluri et al., 2023). This issue not only poses challenges to the credibility of AI systems but also to the user experience, making the exploration of effective communication strategies to address AI errors imperative. The sophistication of conversational GenAI has advanced to the point where these systems are no longer mere tools but are perceived as social entities that participate in dialogues, necessitating adherence to social norms and politeness strategies (Nißen et al., 2022). The efficacy of these strategies, particularly in service failure scenarios, is paramount in maintaining user satisfaction and trust (Song et al., 2023). Moreover, attribution theory offers a lens through which to understand user responses to AI-generated information. It posits that the way individuals infer causes of events, including errors made by AI, impacts their subsequent behavior and interaction with the system (Heider, 1958; Weiner, B., 1994) As such, whether an AI attributes errors to internal limitations or external factors can alter the user's trust and the AI's perceived reliability. This research delves into user preferences and perceptions regarding AI's error management, with a focus on GenAI systems like ChatGPT. By examining these dynamics, the study aims to provide insights that could guide the development of more effective and trustworthy AI systems. As GenAI technology continues to evolve, understanding and addressing the human factors influencing the acceptance and use of these systems remain of utmost importance.

2. Literature Review

2.1. AI Hallucination

In the context of GenAI, 'AI hallucination' refers to the phenomenon where an AI system produces answers that appear convincing yet are entirely fabricated (Athaluri et al., 2023). Even though 'generative' chatbots like OpenAI's ChatGPT, Microsoft's Bing, and Google's Bard have made significant strides in their capabilities over the past year, a major and critical flaw persists they often generate fabricated information (De Vynck, 2023). With the increasing accessibility of AI-generated content online, the issue of "AI hallucination" – the generation of misleading or false information by GenAI – is anticipated to worsen, presenting fresh challenges for ensuring the truthfulness of information in the digital era.

2.2. Politeness Strategy (Appreciation vs. Apology)

Politeness strategies are linguistic tactics employed to preserve the dignity or 'face' of another individual. Brown & Levinson (1987) have identified these strategies as crucial in maintaining 'face,' distinguishing between positive and negative politeness. Positive politeness involves direct actions that acknowledge the other person, aiming to make them feel appreciated and valued, often through expressions of appreciation and compliments. On the other hand, negative politeness is characterized by a more passive approach, seeking to avoid intruding on the other person, typically manifesting in the form of apologies. Historically the focus of interpersonal communication research, politeness theory is increasingly relevant in Computer-Mediated Communication (CMC) as interactions with chatbots become commonplace. As chatbots evolve to mimic human communication more closely, they are being recognized as proficient social entities (Nißen et al., 2022). In instances of service failure, it is beneficial for chatbots to employ polite service recovery strategies to alleviate the adverse effects (Song et al., 2023). Properly addressing issues with dissatisfied users through polite chatbot interactions can lead to increased satisfaction, potentially exceeding the levels before the service error (Hart et al., 1990). This suggests that the implementation of politeness strategies in AI communication is not only a matter of preserving face but also an operational necessity for enhancing user experience.

RQ1. Is Generative AI's politeness strategy (Appreciation vs. Apology) helpful for Users' experience when errors occur?

2.3. Attribution Theory (Internal vs. External)

Attribution Theory is a concept from social psychology that explains how individuals infer the causes of events, others' behavior, and their behavior. It was initially proposed by Heider (1958) and further developed by Weiner (1994). People make attributions to understand their world and to seek reasons for certain outcomes. According to Weiner (1994), causal attributions affect future behaviors, especially after experiences of failure. In Attribution Theory, as explained by Heider (1958), individuals interpret events based on whether they perceive their causes as external or internal. External attribution occurs when an individual ascribes the cause of an event to factors in the external environment, such as when an AI says, "I misunderstood because the question provided was too ambiguous or broad." This type of attribution externalizes the source of the error. In contrast, internal attribution occurs when a person believes that they are to blame for an event, as in the admission that "I generated incorrect information due to my limitations." This form of attribution internalizes the responsibility for the outcome.

RQ2. Is Generative AI's attribution strategy helpful for Users' experience when errors occur?

3. Method

3.1. Study Design

To investigate the users' perceptions of AI hallucination and user response strategies, we designed a qualitative study utilizing user interviews. Our participants consist of five individuals aged between 20 to 30 years who have prior experience with conversational generative AI systems. Participants were exposed to four communication stimuli (2 [Appreciation vs. Apology] X 2 [Internal vs. External]) of AI after reading the AI hallucination scenario.

Table 1. *Participants of the Study*

Participants	Age	Gender	Occupation
P1	24	Female	Students
P2	35	Male	Office-worker
P3	26	Female	Students
P4	24	Female	Students
P5	27	Female	Students

3.2. Stimuli

Exploring the interplay of politeness and attribution strategy in AI, this study assessed how AI responses affect user experience during instances of AI hallucination. Table 2 presents the various communication strategies that AI might employ under such conditions. We crafted four distinct responses as stimuli, presented them to participants, and then conducted interviews to gather their perceptions of each response.

Table 2. *Communication Strategies of GenAI*

Theory		**Attribution strategy**	
		External	Internal
Politeness strategy	Appreciation	Thank you for pointing out the error. There was incorrect information in the external data I referenced (e.g. website, newspaper article, report, etc.).	Thank you for pointing out the error. I think there was a mistake during data processing due to the limitations of my algorithm.
	Apology	Sorry for the misinformation, there was misinformation in the external data I referenced (e.g. website, newspaper article, report, etc.).	Sorry for the misinformation, I think there was a mistake during data processing due to the limitations of my algorithm.

4. Results

The Significance of Transparency in Error Disclosure

The way AI communicates its errors is pivotal in influencing users' trust and satisfaction. Our study highlights a demand for transparency: users expect detailed explanations when errors occur, without which there is a notable decrease in trust.

"The lack of a detailed explanation about the error in the answer leads to a loss of trust" (P5)

“The feeling is that they're glossing over the issue with a hasty apology without specifying the cause of the error, which gives the impression that future responses may also contain inaccurate information" (P3)

Appreciation vs. Apology

Participants preferred expression of appreciation over apology, which reflects the results of previous studies. Song et al. (2023) found that in service failure scenarios, customers' satisfaction after recovery is heightened more effectively through an appreciation strategy rather than by acknowledging the chatbot's limitations via an apology. Lv et al. (2021) observed that when AI devices face service failures, chatbots that express gratitude rather than an apology are more likely to secure consumer forgiveness. In this study, most participants said AI's expression of appreciation makes them feel self-efficacy. For this reason, respondents preferred expressions of appreciation over apology.

Internal vs. External

Our results showed a clear preference among respondents for AI to attribute responsibility for errors to external factors rather than to its internal limitations. This inclination aligns with the principle of locus of control, where users find it less concerning and maintain their trust in the AI when the source of error is perceived as external and beyond the AI’s control. Such a stance seems to be less damaging to the AI's perceived reliability, as the AI is not seen as inherently flawed but rather as impacted by external circumstances.

“I like the way AI properly explains for the transmission of error information due to external data problems” (P1)

"I think external attribution is a softer way to handle errors, and from the perspective of users who use ChatGPT to obtain information, it enables them to maintain their trust in ChatGPT" (P4)

Impact of Deflecting Blame for External Data Errors with Apology

The participants showed a preference for AI errors to be attributed to external factors. However, apologies for errors ascribed to external data were often viewed negatively, being seen as the AI shifting blame rather than accepting responsibility. This perception of evasion led to adverse feelings and diminished trust.

"It feels like AI blaming others when they apologize, which is a negative feeling, unlike internal data error notice." (P3)
"An expression implying a problem with external data casts doubt on the source and accuracy of ChatGPT data. It is also dissatisfied because they dismiss the problem as external." (P2)

Positive Responses to Appreciation for Identifying External Data Errors

In contrast, expressions of appreciation related to external data were met with a more positive reception. These responses made users feel valued and acknowledged for contributing to the AI’s improvement.

"Despite shifting the blame to the outside world, the expression of appreciation didn't give me the impression of avoiding responsibility, so I had a little less doubt about ChatGPT performance" (P3)
"It felt good to know that I was informing ChatGPT of new information when I heard this response. I felt like I discovered an error in external data and contributed to the improvement of ChatGPT." (P5)

Mixed Reactions to Admissions of Internal Limitations

The study also found mixed responses to the AI's acknowledgment of its internal limitations by offering an apology. Participants showed both positive and negative attitudes when AI apologized while acknowledging its internal limitations. While some participants valued AI's honesty and transparency, viewing it as an opportunity to engage with the AI and aid its improvement, others perceived it as a diminishment of the AI’s reliability. This dichotomy suggests that while transparency is appreciated, the framing of such disclosures is critical. AI developers may need to find a balance between candidness and maintaining the AI's image as a competent tool. The impact of the AI admitting internal limitations with an apology showed mixed reactions.

"I am satisfied that it specifies that it is the result of a problem with its algorithm or learning data and that it contains the exact content of the apology. Given that the error was caused by internal data, providing feedback gives the impression that the function will improve, making users want to provide more feedback." (P3)

"It was good that you explained why the incorrect answer occurred. However, the admission of 'internal limit' through apology feels like the ChatGPT itself was insufficient and provided incorrect information." (P4)

The impact of the AI admitting internal limitations with appreciation was unsatisfying overall. Some respondents, like P1, expressed dissatisfaction, feeling that the AI’s response was excessively submissive, which could be perceived as a lack of confidence in its own abilities. P5 shared a sentiment of inadvertent guilt as if the AI’s acknowledgment of its internal limitations made them feel somewhat responsible for highlighting the AI's shortcomings. P2 found the overly polite expression uncomfortable, suggesting that an overly deferential response like appreciation might not always be well-received by users.

5. Conclusion

Our research demonstrates that users appreciate transparency in AI communication, particularly when it involves explicit and detailed explanations of errors. Such clarity was crucial for maintaining trust. Significantly, the study revealed a user preference for appreciation over apology when generative AI systems like ChatGPT dealt with errors. This aligns with the finding that expressions of gratitude by the AI, rather than apologies, are more effective in enhancing user satisfaction, especially in the context of service recovery. Users felt more empowered and engaged when AI systems showed appreciation for their role in error identification, particularly with external data errors. Additionally, attributions of errors to external sources rather than internal AI limitations were preferred, as they were less damaging to the AI's perceived competence. Overall, the study emphasizes the importance of carefully crafted AI responses that prioritize transparent acknowledgments and user recognition to foster trust and user satisfaction.

The findings from this study underscore the nuanced dynamics of user experience in the event of AI errors, highlighting the strategic advantage of appreciation over apology in error communication. Such insights are invaluable for AI developers, suggesting that incorporating acknowledgment strategies that lean towards gratitude can significantly enhance user trust and satisfaction. In a follow-up study, it will be essential to explore how these preferences impact long-term user engagement with AI systems and to investigate strategies that can effectively balance transparency, responsibility, and user experience. Further research could also examine the role of cultural differences in user responses to AI error management.

6. Acknowledgements

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2023-2020-0-01749) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

REFERENCES

Athaluri, S. A., Manthena, S. V., Kesapragada, V. K. M., Yarlagadda, V., Dave, T., & Duddumpudi, R. T. S. (2023). Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus*, *15*(4).

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Vol. 4). Cambridge university press.

De Vynck, G. (2023). Forget AI. For a moment Silicon Valley was obsessed with floating rocks. The Washington Post. Retrieved from https://www.washingtonpost.com/technology/2023/08/11/superconductors-hype-lk99-silicon-valley/

Hart, C. W., Heskett, J. L., & Sasser Jr, W. E. (1990). The profitable art of service recovery. *Harvard business review*, *68*(4), 148-156.

Heider, F. (1958). The psychology of interpersonal relations. Wiley.

Lv, X., Liu, Y., Luo, J., Liu, Y., & Li, C. (2021). Does a cute artificial intelligence assistant soften the blow? The impact of cuteness on customer tolerance of assistant service failure. *Annals of Tourism Research*, *87*, 103114.

Nißen, M., Selimi, D., Janssen, A., Cardona, D. R., Breitner, M. H., Kowatsch, T., & von Wangenheim, F. (2022). See you soon again, chatbot? A design taxonomy to characterize user-chatbot relationships with different time horizons. *Computers in Human Behavior, 127*, 107043.

Song, M., Zhang, H., Xing, X., & Duan, Y. (2023). Appreciation vs. apology: Research on the influence mechanism of chatbot service recovery based on politeness theory. *Journal of Retailing and Consumer Services*, *73*, 103323.

Weiner, B. (1994). Integrating social and personal theories of achievement striving. *Review of Educational research*, *64*(4), 557-573.