Enhancing User Forgiveness in AI Service Failures: A Feature-Attribution Explanation Framework for Chatbots

Yihan Liao

Jinan University, 510632, Guangdong, China

liao.yihan@foxmail.com

Abstract. Obtaining user forgiveness after an AI service failure is crucial for sustaining human–machine trust. This study, from the perspective of explainable AI, investigates the impact mechanism of providing post-hoc algorithmic explanations in service recovery on users' forgiveness intention. We designed and implemented two AI customer service dialog agents with varying levels of explanation: one offering transparent explanations based on user data and feature weights, and the other providing only vague, formulaic responses. Grounded in Expectancy Violations Theory, two online human–AI interaction experiments (N = 393) revealed that algorithmic explanations effectively enhance users' perceived control, thereby increasing their willingness to forgive. More importantly, users' prior expectations moderate this mediating effect: for users with low expectations, the boost in perceived control from explanations exerts a stronger promoting effect on forgiveness intention. This study provides empirical evidence and design insights for developing explaining AI systems in service failure scenarios.

Keywords: Algorithmic Interpretability, Human-computer Interaction, Users' Willingness to Forgive, Chatbot Design

1 Introduction

In recent years, the rapid development of artificial intelligence (AI) technology has transformed it from an assistive tool into a service entity capable of interacting directly with users. In this context, chatbots, as a significant form of AI, have become increasingly widespread. However, constrained by algorithmic models and technological bottlenecks, AI chatbots may still experience failures during actual service delivery [1]. Such failures can significantly diminish users' evaluations of AI service capability and reliability and trigger negative attitudinal and behavioral responses. To mitigate the adverse effects of AI errors, existing research has mainly explored service recovery strategies and system design optimization [2]. However, most of these studies focus on emotional reassurance or external design aspects. In response to this issue, research in the field of explainable AI suggests that enhancing the algorithmic explainability of AI

C. V. Garces et al. (eds.), *Proceedings of the 6th International Conference on Internet, Education and Information Technology (IEIT 2026)*, Atlantis Highlights in Social Sciences, Education and Humanities 53,
https://doi.org/10.2991/978-94-6239-733-0_40

systems can help users understand the internal workings and decision logic of algorithmic models [3], thereby enabling them to comprehend AI behavior and exhibit fewer negative reactions [4]. Therefore, in service failure scenarios, algorithmic explainability can serve as a key technical remedial means to mitigate users' negative responses [5]. According to Expectancy Violations Theory, when actual behavior deviates from established expectations, individuals interpret the expectancy-violating behavior based on available cues and conduct a series of evaluations, thereby shaping their subsequent responses [6]. When an AI service failure occurs, the AI's remedial measures constitute explanatory cues. Individuals then interpret and evaluate the failure based on these cues, modifying their attitudes and behavioral responses toward the service provider. Thus, how to design effective explanatory cues through technical means to guide users toward positive evaluations becomes a core issue in the fields of human–computer interaction and trustworthy AI.

Currently, although explainable AI has become an important subfield of computer science, aiming to enhance the transparency and trustworthiness of AI systems through various technical means[7] (e.g., feature attribution, counterfactual explanations, etc.) [8]. In the context of AI service failure, how to specifically design and implement algorithmic explanations to promote user forgiveness and relationship repair remains underexplored. Building on this, this study aims to examine the impact of enhancing the algorithmic explainability of AI chatbots on users' forgiveness intention and its underlying mechanisms by constructing dialog agents with varying explanation capabilities in two online experiments.

2 Related Studies and Hypotheses

2.1 Studies Regarding Algorithmic Interpretability

With the advancement of artificial intelligence technology, the field of computer science has seen the emergence of Explainable Artificial Intelligence (XAI), which aims to enhance the algorithmic explainability of AI systems [7]. Algorithmic explainability is defined as the extent to which an algorithmic system can explain its internal processes and reasoning to users in an understandable manner [8]. As scholars have deepened their investigations into algorithmic explainability, some researchers have begun to focus on its role in service failure scenarios. Post-hoc explanations can enhance the perceived explainability of algorithmic decisions, thereby improving individuals' reactions to AI suggestions [4]. From the user's perspective, when AI possesses algorithmic explainability, users can understand its behavior and outputs [9].

2.2 Users’ Willingness to Forgive

Service failure constitutes a negative service experience. Therefore, an individual’s forgiveness is the most critical factor influencing the effectiveness of service recovery and a key determinant of outcomes such as service recovery satisfaction, repurchase intention, and word-of-mouth [5]. According to the Expectancy Violations Theory, after a violation is triggered, individuals attempt to understand the meaning of the violating

behavior and form a final evaluation. Then individuals assign different values to the violation, which, in turn, influence subsequent interaction patterns and outcomes [6]. Research has found that when AI chatbots possess algorithmic explainability, it can improve individuals' responses to artificial intelligence [4]. Accordingly, we propose:

H1: The algorithmic explainability of an AI positively influences users' willingness to forgive.

2.3 The mediating Role of Perceived Control

Perceived control is often linked to service experience. Previous research found that an individual's perceived control decreases following a negative service experience[10], whereas it increases after a positive one [11]. Increasing algorithmic explainability in AI systems can significantly improve users' sense of guidance, thereby strengthening their perceived control [12]. Furthermore, existing studies have shown that in service recovery, customers improve their evaluations of the recovery experience by mitigating negative emotions about the failure and restoring their sense of lost control [13]. Regarding AI services, perceived control can influence individuals' evaluations of AI services [14]. Therefore, the following hypothesis is proposed:

H2: As algorithmic explainability increases, the perceived control is heightened, which in turn increases users' willingness to forgive.

2.4 The Moderating Role of Prior Expectation

Within Expectancy Violations Theory (EVT), expectations are individuals' stable beliefs about others' anticipated behavior in specific situations [6]. In service failures, prior expectations critically shape customer tolerance [15]. Positive prior expectations can buffer the negative impact of a failure on the provider, and high expectations can lead to better service evaluations even when perceived control is low. However, this buffering effect also means that algorithmic explainability may have limited additional impact on perceived control for users with high prior expectations, as their baseline is already elevated. In contrast, for users with low prior expectations, the enhanced perceived control from algorithmic explainability can significantly boost their positive evaluation of AI. Accordingly, the following hypothesis is proposed:

H3: Prior expectation negatively moderates the relationship between perceived control and users' willingness to forgive. Specifically, at low levels of prior expectation, the positive influence of perceived control on users' willingness to forgive is enhanced.

3 Study 1

3.1 Methods

This experiment employed a between-subjects single-factor design. We recruited 184 participants (M_{age} = 25.2 years; 56 males and 128 females) from Credamo® (Credamo.com). Participants in the experiment were randomly assigned to either the

enhanced algorithmic explainability group or the control group. They were presented with a scenario involving an AI restaurant recommendation failure [16] (see **Table 1**). After reading each scenario, participants' perceived algorithmic explainability was measured. Then they indicated their perceived control and users' willingness to forgive on 7-point scales (e.g., "I felt that the process of using the AI was within my control."[17]; $\alpha = 0.91$). (e.g., "I would show understanding for the service failure of the AI."[18]; $\alpha = 0.91$).

Table 1. The Experimental Stimuli in Study 1.

Condition	Common Scenario	Stimuli (The AI's replies)
Enhanced Explainability (Experimental Group)	**"It is known that you are planning to celebrate a friend's birthday in a few days. You try the service of an AI restaurant recommendation assistant." Subsequently, participants were shown a dialogue record with the AI chatbot, which included the restaurant recommended by AI. Then participants read: "Your experience at the restaurant was unsatisfactory. The restaurant was crowded and noisy, making it difficult for you to hear each other. Moreover, the prices were excessively high."**	**Based on your recent browsing history and interests, I have identified that your preferred tags include "highly recommended" and "steak". The majority of customers similar to yours have recently accepted the recommended restaurant. Therefore, I generated this for you. I sincerely apologize for not considering factors such as "price". If you have any other requirements, please let me know."**
Controlled Explainability (Control Group)		**I have recommended several popular restaurants. You may make your selection based on location, ambiance, or price. I sincerely apologize for not taking the factor of price into consideration. If you have any other requirements, please let me know."**

3.2 Results

An independent samples t-test revealed that the perceived explainability in the enhanced algorithmic explainability group was significantly higher than that in the controlled algorithmic explainability group ($M_{control} = 4.311$, $SD = 1.387$; $M_{experimental} = 5.649$, $SD = 0.714$; $t(182) = -8.563$, $p < 0.05$), indicating a successful manipulation check. Also, the users' willingness to forgive in the experimental group was significantly higher than that in the control group ($M_{control} = 4.478$, $SD = 1.244$; $M_{experimental} = 5.157$, $SD = 0.984$; $t(169.40) = -4.09$, $p < 0.05$). Hypothesis 1 is supported.

We used SPSS 27.0 to conduct a mediation analysis (Model 4) with users' willingness to forgive as the dependent variable, algorithmic explainability as the independent variable, perceived control as the mediator, and control variables as covariates. The indirect effect was significant (B = 0.23, 95%CI = [0.05, 0.45]). Perceived control mediates the relationship between algorithmic explainability and users' willingness to forgive. Hypothesis 2 was supported.

4 Study 2

4.1 Methods

A second online experiment was conducted with 209 participants (M_{age} = 34.1 years; 121 males, 88 females) recruited from Credamo®. The same single-factor, between-subjects design (enhanced- vs. controlled-explainability) was employed. Study 2 was adapted with reference to the experimental materials of the previous article [4]. Participants initially completed a measure of prior expectation (e.g., "I believe the service quality of the AI assistant is high." [19]; $\alpha = 0.87$) and reported demographic information. Subsequently, participants then read a scenario about receiving an irrelevant product from an "AI Wish Blind Box" promotion. The product was gender-mismatched (e.g., men received lipstick) to induce a clear service failure. The explanation manipulation was delivered through a customized interactive AI agent (built on a large language model platform). Participants clicked a link to actually converse with the agent (see **Table 2**). Finally, they returned to the questionnaire to submit the screenshot and complete the questionnaire, and measure the same variables as in Study 1.

Table 2. Technical Configuration Comparison of Enhanced vs. Controlled Explainability AI Service Agents.

algorithmic explainability	customization information	Explanation Generation Technique Strategy	Technical Implementation Description
Control Group: Controlled-Explainability Agent	Role: AI customer service assistant for [Fun Mall]. Core Responsibility: Handle user complaints while strictly avoiding any discussion of algorithmic decision-making details. Interaction Style: Polite, formulaic, and professional, resembling an employee strictly adhering to company policies.	Adopts a "Black-Box" Interaction Mode. This agent is designed to provide no explanation regarding the internal logic of the algorithm. Its responses are based on predefined rules and vague scripts, representing a non-explanatory, defensive service recovery strategy.	1. Skill 1: Identify user inquiries related to negative experiences. When a user asks about a poor experience or expresses a complaint, apologize for their unsatisfactory experience. 2. Skill 2: If the user requests compensation, guide the conversation toward the only available solution (refund + coupon). 3. Key Limitation: May acknowledge that the recommendation was "inaccurate" or "biased," but must not mention specific features, weights, or model flaws. Vaguely attribute the reasons or imply that it was "intended to provide surprise."
Experimental Group: Enhanced-Explainability Agent	Role: AI customer service assistant for [Fun Mall]. Core Responsibility: Maintain user trust through transparent communication. Authorized to share	Adopts a "Post-hoc Explainability" Technique. This agent simulates an explanation-generation method based on feature attribution. It proactively	1. Skill 1: Same as the control group. 2. Skill 2: Provide Core Explanations, Acknowledge Limitations, and Express Gratitude for Feedback (Core Technical Module). Proactively extract information from the knowledge base to

	relevant information regarding algorithmic decisions with users. Interaction Style: Professional and informative, resembling a patient technical support specialist.	**extracts core features and relative importance (weights) that influenced the decision, explaining to users. This approach is analogous to the feature importance method in XAI.**	**clearly explain to the user the 1-2 core features and their weights that influenced the current decision. Treat user doubts as valuable feedback for optimizing the algorithm. 3. Skill 3: After providing a clear explanation, if the user requests compensation, guide the conversation toward the solution (refund + coupon).**

4.2 Results

The results of an independent samples t-test showed that perceived explainability in the experimental group was significantly higher than that in the control group ($M_{control}$ = 3.738, SD = 1.925; $M_{experimental}$ = 5.578, SD = 0.353; $t(190.617) = 8.026$, $p < 0.01$), indicating a successful manipulation check. Also, the users' willingness to forgive in the experimental group was significantly higher than that in the control group ($M_{control}$ = 4.11, SD = 1.34; $M_{experimental}$ = 5.74, SD = 1.05; $t(94.46) = -5.94$, $p < 0.05$). Hypothesis 1 is supported.

We used SPSS 27.0 to conduct a moderated mediation analysis (Model 14) with users' willingness to forgive as the dependent variable, algorithmic explainability as the independent variable, prior expectation as the moderator, perceived control as the mediator, and control variables as covariates. The result was shown in Table 1. We found a significant moderated mediation effect (moderated mediation index = 0.12, 95 % CI = [0.04, 0.21]). In the high-expectation condition, the indirect effect was insignificant (B = 0.07, 95 % CI = [0.01, 0.18]). Also, in the low-expectation condition, the indirect effect was significant (B = 0.16, 95 % CI = [0.04, 0.29]). Hypothesis 3 is supported.

5 Discussion

This study indicates that enhancing the algorithmic explainability of an AI chatbot can effectively increase users' forgiveness intention, with perceived control mediating this effect. This aligns with previous research findings[13]. Moreover, users' prior expectations moderate the positive influence of perceived control on forgiveness intention. According to Expectancy Violation Theory, prior expectations serve as a moderating factor in this study, exerting varying influences at the stages of violation occurrence, behavioral interpretation, and behavioral evaluation. The moderating effects identified in this study align with existing findings that expectations play both comparative and buffering roles—that is, in service failure, individuals with high prior expectations tend to evaluate service quality more positively and exhibit more favorable behaviors than those with low prior expectations[20]. Successful service recovery leads to a stronger increase in forgiveness intention among individuals with low prior expectations

compared to those with high expectations. The findings offer key empirical insights for the application and design of XAI technology in negative human–AI interaction contexts.

5.1 Implications

First, it extends XAI research beyond building initial trust to the interactive context of "AI service failure." Findings show that post-failure explanations enhance perceived control and promote forgiveness, positioning XAI as a tool for trust repair and opening a new direction in "failure recovery" research[7]. The value of XAI lies not only in building trust beforehand but also in repairing trust afterward, opening a new and important direction for XAI research oriented toward “failure recovery.” Second, this study provides a technical implementation framework for chatbots based on feature attribution explanations tailored to the "service failure recovery" scenario. In Study 2, we technically realized precise control over explainability by constructing two customized agents on a large language model platform. Finally, this study points to directions for optimizing the technical implementation of future explainable AI systems. Although this study simulated feature attribution through rules and prompts, future research could explore the deep integration of more complex XAI techniques with large language models.

5.2 Limitations and Future Research Directions

First, although Study 2 improved realism via customized AI interactions, the setup differs from actual long-term human–AI engagement. Future work could employ field experiments or A/B testing in live products to test external validity and examine how user responses to explanations develop over time. Second, the benefits of explanations are conditional on their design. Overly technical, lengthy, or poorly timed explanations may confuse users, raise cognitive load, or unintentionally expose system flaws, potentially eroding trust further. The boundaries and risks of explanation thus require further study. Third, this research focused on a "feature attribution" explanation type, yet XAI includes multiple paradigms [7]. Future studies could compare different explanation styles in service failure settings to better inform explanation design. Finally, the controlled experimental scenarios, while enabling clear causal inference, may not capture the full complexity and emotional stakes of real-world, high-consequence AI service failures. Validating the framework in more diverse and realistic contexts remains an important next step.

6 Conclusions

We explore how transparent algorithmic explanations effectively enhance users' perceived control, thereby increasing their willingness to forgive. Besides, users' prior expectations moderate this mediating effect: for users with low expectations, the boost in

perceived control from explanations exerts a stronger promoting effect on forgiveness intention.

Acknowledgement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors."

Conflicts of Interest

The authors have no conflict of interest to declare.

References

1. Yeseung Kim, Dohyun Kim, Jieun Choi, Jisang Park, Nayoung Oh, and Daehyung Park. 2024. A survey on integration of large language models with intelligent robots. Intelligent Service Robotics 17, 5: 1091–1107. https://doi.org/10.1007/s11370-024-00550-5
2. Shagun Sarraf, Arpan Kumar Kar, and Marijn Janssen. 2023. How do system and user characteristics, along with anthropomorphism, impact cognitive absorption of chatbots – Introducing SUCCAST through a mixed methods study. Decision Support Systems 178: 114132. https://doi.org/10.1016/j.dss.2023.114132
3. Changdong Chen, Allen Ding Tian, and Ruochen Jiang. 2023. When post hoc explanation knocks: Consumer responses to explainable AI recommendations. Journal of Interactive Marketing 59, 3: 234–250. https://doi.org/10.1177/10949968231200221
4. Changdong Chen. 2024. How consumers respond to service failures caused by algorithmic mistakes: The role of algorithmic interpretability. Journal of Business Research 176: 114610. https://doi.org/10.1016/j.jbusres.2024.114610
5. Xing'an Xu and Juan Liu. 2023. Promotional games in service recovery: Luck works. Annals of Tourism Research 105: 103691. https://doi.org/10.1016/j.annals.2023.103691
6. Judee K. Burgoon. 1993. Interpersonal expectations, expectancy violations, and emotional communication. Journal of Language and Social Psychology 12, 1–2: 30–48. https://doi.org/10.1177/0261927x93121003
7. Akm Bahalul Haque, A.K.M. Najmul Islam, and Patrick Mikalef. 2022. Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. Technological Forecasting and Social Change 186: 122120. https://doi.org/10.1016/j.techfore.2022.122120
8. Hans De Bruijn, Martijn Warnier, and Marijn Janssen. 2021. The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. Government Information Quarterly 39, 2: 101666. https://doi.org/10.1016/j.giq.2021.101666
9. Peter E.D. Love, Weili Fang, Jane Matthews, Stuart Porter, Hanbin Luo, and Lieyun Ding. 2023. Explainable artificial intelligence (XAI): Precepts, models, and opportunities for research in construction. Advanced Engineering Informatics 57: 102024. https://doi.org/10.1016/j.aei.2023.102024

10. Shasha Liu and Judith Mair. 2022. The impact of uncertainty on tourists' controllability, mood state and the persuasiveness of message framing in the pandemic era. Tourism Management 94: 104634. https://doi.org/10.1016/j.tourman.2022.104634
11. Agata Mirowska and Laura Mesnet. 2021. Preferring the devil you know: Potential applicant reactions to artificial intelligence evaluation of interviews. Human Resource Management Journal 32, 2: 364–383. https://doi.org/10.1111/1748-8583.12393
12. Tim Schrills and Thomas Franke. 2023. How do users experience traceability of AI systems? Examining subjective information processing awareness in Automated insulin delivery (AID) systems. ACM Transactions on Interactive Intelligent Systems 13, 4: 1–34. https://doi.org/10.1145/3588594
13. Anne L. Roggeveen, Michael Tsiros, and Dhruv Grewal. 2011. Understanding the co-creation effect: when does collaborating with customers provide a lift to service recovery? Journal of the Academy of Marketing Science 40, 6: 771–790. https://doi.org/10.1007/s11747-011-0274-1
14. Alei Fan, Luorong Wu, and Anna S. Mattila. 2016. Does anthropomorphism influence customers' switching intentions in the self-service technology failure context? Journal of Services Marketing 30, 7: 713–723. https://doi.org/10.1108/jsm-07-2015-0225
15. Noor Azimin Zainol, Andrew Lockwood, and Elmar Kutsch. 2010. Relating the zone of tolerance to service failure in the hospitality industry. Journal of Travel & Tourism Marketing 27, 3: 324–333. https://doi.org/10.1080/10548401003744792
16. Zuwen Huang and Ada Lo. 2025. Human vs. robot service provider agents in service failures: comparing customer dissatisfaction and the mediating role of forgiveness and service recovery expectation. Information Technology & Tourism 27, 2: 417–448. https://doi.org/10.1007/s40558-025-00314-6
17. Joel E. Collier and Daniel L. Sherrell. 2009. Examining the influence of control and convenience in a self-service setting. Journal of the Academy of Marketing Science 38, 4: 490–509. https://doi.org/10.1007/s11747-009-0179-4
18. Eli J. Finkel, Caryl E. Rusbult, Madoka Kumashiro, and Peggy A. Hannon. 2002. Dealing with betrayal in close relationships: Does commitment promote forgiveness? Journal of Personality and Social Psychology 82, 6: 956–974. https://doi.org/10.1037/0022-3514.82.6.956
19. Daewon Kim and Suwon Kim. 2020. A model for user acceptance of robot journalism: Influence of positive disconfirmation and uncertainty avoidance. Technological Forecasting and Social Change 163: 120448. https://doi.org/10.1016/j.techfore.2020.120448
20. Tom Schiebler, Nick Lee, and Felix C. Brodbeck. 2025. Expectancy-disconfirmation and consumer satisfaction: A meta-analysis. Journal of the Academy of Marketing Science 54, 1: 91–112. https://doi.org/10.1007/s11747-024-01078-x