Empathic Mirroring Responses to Compensate for AI Chatbot Failures

Jano Jimenez Barreto
Universidad de La Laguna
Grzegorz Kapuscinski
Oxford Brookes University
Natalia Rubio
Autónoma University of Madrid Q2818013A
Corné Dijkmans
Breda University

Acknowledgements:
PID2023-147414OB-I00; TED2021-129513B-C22

Cite as:
Jimenez Barreto Jano, Kapuscinski Grzegorz, Rubio Natalia, Dijkmans Corné (2025), Empathic Mirroring Responses to Compensate for AI Chatbot Failures. *Proceedings of the European Marketing Academy*, 54th, (126229)

Paper from the 54th Annual EMAC Conference, Madrid, Spain, May 25-30, 2025

Empathic Mirroring Responses to Compensate for AI Chatbot Failures

Abstract

In the face of uncertainties where responsibility for AI failures cannot be attributed to a specific market actor, this research explores how companies can employ persuasive tactics to manage corporate crises caused by AI chatbot failures. Specifically, we provide evidence regarding the effectiveness of different formats of a "mirroring strategy," which involves denying responsibility by attributing blame to external factors, such as consumer misuse of technology. Our investigation reveals that employing the mirroring strategy, combined with displays of empathy toward consumers, can significantly reduce firms' reputational damage during AI chatbot crises, regardless of the pre-existing narratives shaped by news media in their reporting.

Keywords: AI failure, attributed responsibility, reputation crisis
Track: Digital Marketing & Social Media

1. Introduction

Artificial intelligence (AI) has emerged as a game-changing innovation with vast implications for business and society (Hartmann et al., 2024; Yalcin et al., 2022). Although marketers strive to create positive impacts by implementing this technology, AI failures are almost inevitable (Pavone et al., 2023; Srinivasan and Drial-Abi, 2021). Several high-profile cases have triggered public outrage and caused significant damage to firms' reputations. In some cases, firms attribute responsibility to others (e.g., consumers, AI-based agents, or AI developers) for erroneous outcomes, resulting in direct confrontations with consumers regarding who is accountable for the AI failure (Yagoda, 2024). Although the crisis communication literature provides multiple strategies for responding to crises, AI failures pose unique challenges for marketers. Traditionally, firms address reputational crises by identifying the attribution of responsibility, conceptualized through two dimensions: the locus of causality (i.e., whether internal or external factors caused the issue) and controllability (i.e., the extent to which the service provider could control the failure; Coombs, 2007). However, in the context of AI, determining the locus of causality and controllability is more complex than in human-made crises (Prahl and Goh, 2021). Consequently, key stakeholders may experience inherent uncertainty about their responsibilities for AI failures (Lee et al., 2021).

In this uncertain context represented by AI failures, we propose that firms can apply persuasive tactics to mitigate corporate crises and, in turn, maintain positive relationships with consumers. This involves what Prahl and Goh (2021) refer to as the "mirroring strategy." Mirroring is a form of denial through scapegoating, transferring the blame for AI failures to public misuse of technology. For instance, firms might argue that AI failures, such as those involving chatbots, result from inappropriate user behavior, such as malicious prompt injections—hacker-generated prompts designed to manipulate AI systems (Liu et al., 2023). However, there is limited evidence regarding the effectiveness of the mirroring strategy and its impact on firms' reputations. This research addresses this gap through a series of studies.

2. Theoretical Background

During crises, the public relies on information from various sources to make sense of the situation and assign responsibility (Coombs, 2007). News media content about AI failures shapes consumer responses, including negative emotions (Kim and Cameron, 2011),

responsibility attributions (Cho and Gower, 2006), and organizational reputation (Mason, 2019). According to Jin et al. (2020), crisis responsibility information is often perceived as ambiguous, particularly when multiple sources present conflicting accounts of the causes. AI failures are novel and complex phenomena involving many actors that influence service outcomes (Syed, 2023), making attribution of responsibility more challenging. A study of European news coverage of AI failures (Barassi et al., 2022) revealed that journalists offer diverse explanations for AI errors, attributing responsibility inconsistently. In this context, four actors may emerge in media narratives about AI failures: the organization implementing AI, the AI technology (e.g., chatbots), software developers, and consumers.

Crisis communication research emphasizes that responsibility attributions link crisis information to post-crisis reputation (Fediuk et al., 2012). Additionally, research highlights that the extent of reputational damage depends on stakeholders' empathy for the organization (Ndone and Park, 2022). Grounded in theories of interpersonal forgiveness, such as the empathy model of forgiveness, Schoofs et al. (2019) demonstrate that empathy towards an organization in crisis can mediate the relationship between crisis information (i.e., victim crisis and apologetic response) and organizational reputation in two ways. First, empathy explains this relationship sequentially through lower perceptions of organizational responsibility, which induce consumer empathy for the organization. Second, independently of responsibility attribution, empathy allows victims to see the crisis from the other's perspective, building a greater understanding of the organization's behavior (Wade et al., 2005). Moreover, empathy establishes a connection between the transgressor and the victim, which can offset the negative feelings resulting from the crisis (Riek and Mania, 2012).

Empathy has cognitive and affective components. Cognitive empathy refers to understanding another person's emotional state, while the affective component is defined as a congruent emotional response to another person's emotions (Blair, 2005). Affective empathy received much more attention in social-psychological research on forgiveness (Fehr et al., 2010) and was shown to mediate the relationship between crisis information and public crisis responses (Schoofs et al., 2019). Similarly, Fannes and Claeys (2022) suggest that in the context of crisis communication, the expression of cognitive empathy is insufficient; therefore, to protect their post-crisis reputation, firms should ensure their responses contain affective empathy.

We expect that frames embedded in crisis news reports attributing responsibility to different actors will result in different degrees of affective empathy towards the firm in crisis during an AI failure. Empathic feelings toward the organization are expected to emerge when the organizations are perceived as victims of the crisis. Although we expect responsibility framing to influence organizational reputation through increased feelings of empathy toward the organization, the impact might be moderated by the crisis response strategy.

Crisis responses are critical to shaping public perceptions post-crisis. While numerous crisis communication studies demonstrate the effectiveness of apologies, some have shown that denial can also be an effective means of protecting reputations by reducing the attribution of responsibility—especially when the crisis emerges in a complex context, such as consumer interaction with AI technology (Page 2020). In light of mixed findings, these studies call for a re-examination of denial's role in crisis recovery (Kim and Sung, 2014). Notably, few of the studies that considered the role of denial differentiated between denial strategies (e.g., scapegoating, attacking the accuser) and rarely studied the effectiveness of scapegoating (Page, 2020).

As the mirroring strategy represents a unique form of scapegoating (i.e., denial strategy) specific to issues of AI, determining its role in public responses is critical for marketers. Moreover, we suggest that while denial can shape perceptions of responsibility, expression of empathy can help observers develop emphatic concern towards firms. In this regard, firms often express emotions when communicating with the public, which may affect the emotions the receiver experiences toward the organization (Van der Meer and Verhoeven, 2014). Therefore, empathy is considered a vital part of crisis communication efforts. Expression of compassion and concern towards the affected victims helps stakeholders to cope with the crisis psychologically (Coombs, 2007), increases public empathy towards the CEO (Schoofs et al., 2019), allows firms to appear more trustworthy, and as such, reduces the negative impact of crises on its reputation (Kiambi and Shafer, 2016).

Given that the mirroring strategy is a form of denial, it is likely to be met with some resistance from the public. According to Coombs (2007), scapegoating tends to be perceived negatively by stakeholders who want organizations to take responsibility rather than shift the blame. Moreover, denial may frustrate people's understanding of what happened (Gillespie et al. 2014). Denial signals no compassion for victims and is ineffective in inducing empathic concern (cognitive and affective empathy) for an organization in crisis (Ndone and Park,

2022; Schoofs et al., 2019). For these reasons, the expression of empathy may be particularly relevant to offset the potentially negative effects of the mirroring strategy, which may be seen as cold and uncaring.

Based on our literature review, we propose that the responsibility frames will result in different degrees of empathy toward the firm in crisis, and this relationship is moderated by a denial response accompanied by the expression of empathy (i.e., toward consumers). In turn, empathy toward the organization stimulates the public to evaluate the firm reputation more positively and be less inclined to engage in negative word-of-mouth. Formerly, our hypothesis considers that when firms respond to an AI crisis in which news reports attribute responsibility for the crisis to consumers (a), the firm (b), developers (c), or the AI chatbot (d), using a denial strategy (mirroring) that features expressions of customer empathy (vs. without empathy expressions) will arouse more empathy toward the organization and subsequently enhance its reputation.

The present research

A multimethod investigation, including two studies, explores whether firms' mirroring empathetic responses to AI failures effectively enhance consumer empathy toward the company and its reputation.

Study 1 evaluates whether consumers agree on who is responsible in a situation where a firm is under media scrutiny due to an AI failure. Using visual elicitation techniques, we presented 58 U.S. consumers (44% female; 32% aged 25-34) with a newspaper report about a failure related to a service company's chatbot. The report included four different arguments regarding who might be responsible for the AI failure: the firm, the consumers, the AI-chatbot itself, or the AI developer. After reading the report, participants evaluated the firm's response to the public, presented as a post on X (formerly Twitter) that either included an empathetic response or did not, depending on the participant's random assignment to this second stimulus. Participants were then asked to share their thoughts on who was responsible and whether the firm responded appropriately to consumers. To analyze the data, we examined whether the actors identified in the news article as potentially responsible for the crisis were reflected in the consumers' opinions. Our findings indicated that, although the media attributed responsibility to the firm, consumers, AI-chatbot, or AI developers regarding the AI failure, there was no clear consensus on who was responsible. This lack of clarity stemmed from the high degree of disparity in participants' narratives about responsibility.

Moreover, participants mentioned that the firm's empathetic response was more effective in addressing the situation and mitigating potential harm to its reputation than the non-empathetic response.

Study 2 manipulates the context regarding an AI failure of a service firm and the subsequent firms' public response. Specifically, Study 2 tests whether the empathetic mirroring response empirically moderates the relationships among an external (vs. internal) attribution of responsibility, consumer empathy toward the firm, and firm reputation. We employed a 2 (Responsibility frames: the firm vs. others, i.e., Consumers, chatbot, developers) x 2 (Firm response: Mirror with vs. without empathy expression) between-subjects experimental design with five hundred and eighty-six U.S participants (59% female; 30% between 25-34). The scenarios introduced a case of a fictitious airline that attracted media attention following an AI chatbot failure. The reputational frame embedded in each of the four news stories attributed the cause of the crisis to one of four actors (i.e., Airline, Chatbot, Developer, or Consumers). Next, respondents read a response from the airline; one condition denied responsibility by transferring blame to consumers with an expression of empathy, and the second denied without an expression of empathy. To check for manipulations, we measured awareness of framing by asking whether the article attributed responsibility to either of the four actors (4-items), and expression of empathy by perceived demonstration of understanding (2-items, $r = .80$; Sacks, 1992). A one-way ANOVA tested the manipulation check for *responsibility framing* and showed a significant effect for recognition of the actor blamed for the crisis in each case as expected; consumers ($F[1, 3] = 45.09$; $p < .001$), airline ($F[1, 3] = 47.65$; $p < .001$), chatbot ($F[1, 3] = 25.11$; $p < .001$) and developer ($F[1, 3] = 26.76$; $p < .001$). Hence, the manipulation checks were successful. The test showed that in the case of a firm response with an expression of empathy, respondents observed more demonstration of understanding ($M = 4.78$; SD =1.48) than those in the scenario with no empathy expression ($M = 2.68$; SD = 1.61; $t[585] = -16.37$, $p < .001$). Next, we measured participants' affective empathy toward the firm (7-items, $\alpha = .97$; adapted from McCullough et al., 2003) and firm reputation (4-items, $\alpha = .92$; adapted from Coombs and Holladay, 2002; seven-point Likert scale, 1 = strongly disagree; 7 = strongly agree).

Further, we conducted a moderated mediation analysis to determine the relationships in our conceptual model (Hayes 2022, Model 7). In this model, the failure responsibility frame (the firm vs. others: consumers, chatbot, AI developer) is the independent variable, the firm response featuring customer empathy (vs. no empathy), the moderator, empathy towards the

firm the mediator, and firm reputation the dependent variable. The indexes of moderated mediation showed a significant overall effect (.48; CI [.02 to .93]) from failure responsibility frames to firm reputation via the mediation of empathy toward the company, moderated by the firm response. This indirect effect is significant when the denial response features no empathy (*SE* = -.55; 95% CI [-.90 to -.21]), but the effect disappears when denial features empathy (*SE* = -.06; 95% CI [-.39 to .24]). The firm response that features customer empathy leads to similarly favorable empathy towards the firm (consumers frame $M = 3.67$ vs. others frame $M = 3.67$; $F[1, 585] = .20$; $p > .05$) and, subsequently, similar reputation scores (consumers frame $M = 4.28$ vs. others frame $M = 4.29$; $F[1, 585] = .00$; $p > .05$). By contrast, when the mirroring response demonstrates no customer empathy, the public evaluation of the firm is less favorable, and the group differences are more pronounced in empathy toward the firm (consumers frame $M = 3.27$ vs. others frame $M = 2.48$; $F[1, 585] = 13.1$; $p < .001$) and its reputation (consumers frame $M = 3.96$ vs. others frame $M = 3.46$; $F[1, 587] = 5.2$; $p < .05$ value).

General conclusion

Our investigation shows that when firms address a reputational crisis involving an AI chatbot failure, employing a mirroring strategy that conveys empathy toward consumers can significantly reduce reputational damage, regardless of the attribution of responsibility mentioned by the media. When empathy is absent in the mirroring response, the outcome is less favorable, and the effectiveness of restoring the firm's reputation is lowest when responding to the media frame that attributes blame to the firm, chatbot, or developers. Taken together, this study addresses the need for further research on the effectiveness of denial as a crisis response strategy in marketing and communication (Page, 2020; Pavone et al., 2023; Srinivasan & Drial-Abi, 2021) and adds to the growing body of literature that highlights the importance of expressing empathy in post-crisis reputation recovery (e.g., Fannes & Claeys, 2022). Additionally, our findings demonstrate that evoking empathy toward the firm in public helps explain the relationship between crisis information and reputational outcomes. In this context, we contribute to the literature exploring the mediating role of empathy in crisis communication (Ndone & Park, 2022; Schoofs et al., 2019).

Reference

Barassi, V., Scharenberg, A., Poux-Berthe, M., Patra, R. & Di Salvo, P. (2022). AI Errors and the Profiling of Humans: Mapping the Debate in European News Media. *Report by The Human Error Project. St. Gallen, Switzerland: Universität St. Gallen*.

Blair, R.J.R. (2005). Responding to the emotions of others: Dissociating forms of empathy through the study of typical and psychiatric populations. *Consciousness and Cognition*, *14*(4), 698–718.

Cho, S.H. & Gower, K.K. (2006). Framing effect on the public's response to crisis: Human interest frame and crisis type influencing responsibility and blame. *Public Relations Review*, *32*(4), 420–422.

Coombs, W.T. & Holladay, S.J. (2002). Helping crisis managers protect reputational assets: Initial tests of the situational crisis communication theory. *Management Communication Quarterly*, *16*(2), 165–186.

Coombs, W.T., (2007). Protecting organization reputations during a crisis: The development and application of situational crisis communication theory. *Corporate Reputation Review*, *10*, 163–176.

Fannes, G. & Claeys, A.S., (2022). Putting empathic feelings into words during times of crisis: The impact of differential verbal empathy expressions on organizational reputation. *Public Relations Review*, *48*(2), 102183.

Fediuk, T. A., Coombs, W. T., & Botero, I. C. (2012). Exploring crisis from a receiver perspective: Understanding stakeholder reactions during crisis events. In W. T. Coombs & S. J. Holladay (eds), *The handbook of crisis communication* (pp. 635–656). Chichester: Wiley-Blackwell.

Fehr, R., Gelfand, M.J. & Nag, M., (2010). The road to forgiveness: a meta-analytic synthesis of its situational and dispositional correlates. *Psychological Bulletin*, *136*(5), 894–914.

Gillespie, N., Dietz, G. & Lockey, S., (2014). Organizational reintegration and trust repair after an integrity violation: A case study. *Business Ethics Quarterly*, *24*(3), 371–410.

Hartmann, J., Exner, Y., & Domdey, S. (2024). The power of generative marketing: Can generative AI create superhuman visual marketing content? *International Journal of Research in Marketing* [In press] doi: https://doi.org/10.1016/j.ijresmar.2024.09.002

Hayes, A.F. (2022) *Introduction to mediation, moderation, and conditional process analysis, third edition: A regression-based approach*. Guilford Publications.

Jin, Y., van der Meer, T.G.L.A., Lee, Y.I. & Lu, X. (2020). The effects of corrective communication and employee backup on the effectiveness of fighting crisis misinformation. *Public Relations Review*, 46, 101910.

Kiambi, D.M. & Shafer, A. (2016). Corporate crisis communication: Examining the interplay of reputation and crisis response strategies. *Mass Communication and Society*, *19*(2), 127–148.

Kim, H.J. & Cameron, G.T. (2011). Emotions matter in crisis: The role of anger and sadness in the publics' response to crisis news framing and corporate crisis response. *Communication Research*, *38*(6), 826–855.

Kim, S. & Sung, K.H. (2014). Revisiting the effectiveness of base crisis response strategies in comparison of reputation management crisis responses. *Journal of Public Relations Research*, *26*(1), 62–78.

Lee, Y.I., Lu, X. & Jin, Y. (2021). Uncertainty management in organizational crisis communication: the impact of crisis responsibility uncertainty and attribution-based emotions on publics further crisis information seeking. *Journal of Communication Management*, *25*(4), 437–453.

Liu, Y, Deng, G, Li, Y, Wang, Y, Zhang, T., Liu, Y., Wang, H., Zheng, Zheng, Y, & Liu, Y. (2023). Prompt injection attack against LLM-integrated applications, arXiv preprint arXiv:2306.05499.

Mason, A., (2019). Media frames and crisis events: Understanding the impact on corporate reputations, responsibility attributions, and negative affect. *International Journal of Business Communication*, *56*(3), 414–431.

McCullough, M.E., Fincham, F.D. & Tsang, J.A., (2003). Forgiveness, forbearance, and time: the temporal unfolding of transgression-related interpersonal motivations. *Journal of Personality and Social Psychology*, *84*(3), 540–557.

Ndone, J. & Park, J., (2022). Crisis communication: The mediating role of cognitive and affective empathy in the relationship between crisis type and crisis response strategy on post-crisis reputation and forgiveness. *Public Relations Review*, *48*(1), 102136.

Page, T.G. (2020). I didn't do it: comparing denial posture crisis strategies between government and business. *Corporate Reputation Review*, *23*, 24–41.

Pavone, G., Meyer-Waarden, L., & Munzel, A. (2023). Rage Against the Machine: Experimental Insights into Customers' Negative Emotional Responses, Attributions of Responsibility, and Coping Strategies in Artificial Intelligence–Based Service Failures. *Journal of Interactive Marketing*, *58*(1), 52–71.

Prahl, A. & Goh, W.W.P., (2021). "Rogue machines" and crisis communication: When AI fails, how do companies publicly respond? *Public Relations Review*, *47*(4), 102077.

Riek, B.M. & Mania, E.W., (2012). The antecedents and consequences of interpersonal forgiveness: A meta-analytic review. *Personal Relationships*, *19*(2), 304–325.

Sacks, H. (1992). In G. Jefferson (Ed.), *Lectures on conversation: Vol II*. Blackwell.

Schoofs, L., Claeys, A.S., De Waele, A. & Cauberghe, V. (2019). The role of empathy in crisis communication: Providing a deeper understanding of how organizational crises and crisis communication affect reputation. *Public Relations Review*, *45*(5), 101851.

Srinivasan, R., & Sarial-Abi, G. (2021). When algorithms fail: Consumers' responses to brand harm crises caused by algorithm errors. *Journal of Marketing*, 85(5), 74–91.

Syed, R. (2023) *So sue me: Who should be held liable when AI makes mistakes?*, *Monash Lens*. Available at: https://lens.monash.edu/@politics-society/2023/03/29/1385545/so-sue-me-wholl-be-held-liable-when-ai-makes-mistakes (Accessed: 22 January 2024).

Van der Meer, T.G. & Verhoeven, J.W. (2014). Emotional crisis communication. *Public Relations Review*, *40*(3), 526–536.

Wade, N.G., Worthington Jr, E.L. & Meyer, J.E., (2005). But do they work? A meta-analysis of group interventions to promote forgiveness. *Handbook of forgiveness*, 423–440.

Yagoda, M. (2024) Airline held liable for its chatbot giving passenger bad advice - what this means for travellers. Available at: https://www.bbc.com/travel/article/20240222-air-canada-chatbot-misinformation-what-travellers-should-know (Accessed: 1 May 2024).

Yalcin, G., Lim, S., Puntoni, S., & van Osselaer, S. M. (2022). Thumbs up or down: Consumer reactions to decisions by algorithms versus humans. *Journal of Marketing Research*, 59(4), 696–717.