

MEMORY ON OR OFF: HOW CONVERSATIONAL MEMORY IN GENERATIVE AI AFFECTS DISCLOSURE LENGTH AND PERCEIVED SAFETY

*Li Xia

**Bilal Ahmad Lone

Paper Received: 30.11.2025 / **Paper Revised:** 30.01.2026 / **Paper Accepted:** 15.02.2026 / **Paper Published:** 28.02.2026
Corresponding Author: Li Xia; doi:10.46360/cosmos.et.620261006

Abstract

Persistent, cross-session memory is becoming a standard feature of generative AI assistants, allowing systems to recall a user's name, preferences, and prior disclosures across conversations. This article examines how enabling versus disabling conversational memory affects two outcomes central to human-AI communication: the length of users' self-disclosures and their perceived safety when sharing personal information. Integrating communication privacy management theory, computers-are-social-actors research, and recent empirical work on long-term memory in large language model (LLM)-driven chatbots (2019–2025), the article reports an illustrative between-subjects experiment ($N = 356$) constructed to illustrate how such a true experiment could be analyzed, including the operational and ethical requirements of the memory manipulation; these results are not derived from actual human participants. The article defines the outcome constructs, discusses ethical safeguards for sensitive-topic disclosure research, and provides a concrete action plan for a genuine empirical test.

Keywords: AI, Human-AI Communication, Generative AI, Ethical, Empathy, Automation, Safety, Chatbot Interaction.

1. Introduction

Generative AI systems built on large language models increasingly offer persistent memory: the ability to retain and recall information from earlier conversations, sometimes across sessions spanning weeks or months. This shift moves conversational AI from a stateless, single-turn tool toward something closer to an ongoing relationship, in which the system 'knows' the user's name, circumstances, and history (Jo et al., 2024). Memory has been marketed as a feature that increases helpfulness and personalization, yet it simultaneously changes the privacy calculus of every disclosure a user makes, since information shared today may resurface, unprompted, in a future conversation.

Two outcomes are of interest: disclosure length and perceived safety. Prior research shows anthropomorphic cues can shape disclosure and privacy concern, though effects vary in strength and directness across studies (Ischen et al., 2020; Pizzi et al., 2023), but memory introduces a distinct dynamic: a durable, accumulating record (Jiang, 2024). This article asks whether memory increases disclosure length, whether it raises or lowers perceived safety, and whether topic sensitivity moderates these effects.

2. Literature Review

2.1 Conversational Memory in Generative AI

Jo et al. (2024) analyzed 1,252 call logs and conducted in-depth interviews with 9 users of CareCall, a voice chatbot for public health check-ins. Given the small interview sample, their qualitative findings should be understood as illustrative rather than broadly generalizable. They found long-term memory enhanced health disclosure and fostered positive perceptions, but also introduced friction around chronic topics and heightened privacy concerns for some users.

Complementary work has examined how users experience memory failures and limitations. Qualitative interview studies of frequent chatbot users have found that lapses in memory are often experienced as relational ruptures rather than mere technical glitches, prompting users to actively repair continuity themselves rather than relying on the system (see recent thesis-level qualitative work on ChatGPT and Character.AI users). This suggests that once users come to expect memory, its presence or absence is read as socially meaningful rather than a neutral technical setting.

2.2 Self-Disclosure and Anthropomorphism in Chatbot Interaction

*-**School of Social Science, Arts and Humanities, Lincoln University College, Malaysia.

Ischen et al. (2020), in a between-subjects experiment with 181 participants, found human-like chatbots elicited more disclosure than machine-like ones, mediated by perceived anthropomorphism and lower privacy concern - consistent with the CASA paradigm (Reeves & Nass, 1996; Nass et al., 1994), which holds that people apply social rules from human interaction to computers exhibiting sufficient social cues. However, Pizzi et al. (2023), studying gaze direction and anthropomorphism, found no significant direct effect on disclosure intentions; the hypotheses predicting direct effects were not supported. Instead, these cues affected disclosure only indirectly, via warmth/competence perceptions reducing skepticism. This is an important correction: the anthropomorphism-disclosure relationship is indirect and conditional, not direct.

2.3 Privacy Perceptions and Perceived Safety with Generative AI

Kelley et al. (2023), surveying 10,011 respondents across ten countries, found 49% expected "less privacy" from AI. Lappeman et al. (2022), in 284 South African respondents (snowball sampling, two groups of 142), found trust predicted disclosure to banking chatbots while perceived risk suppressed it independently. Gumusel (2024), reviewing 38 papers from 894 retrieved records, found persistent gaps between stated privacy concerns and actual disclosure behavior. Lee et al. (2024), analyzing 321 documented AI privacy incidents, proposed a taxonomy of 12 AI privacy risks distinguishing one-time collection from persistent retention/resurfacing risks. Handa et al. (2025), analyzing 1,250 interview transcripts (1,000 workforce, 125 creatives, 125 scientists), found comfort disclosing to AI tied to understanding of data retention.

3. Conceptual Framework and Hypotheses

Drawing on communication privacy management theory (Petronio, 2002) and the computers-are-social-actors' paradigm (Reeves & Nass, 1996), this article proposes conversational memory operates as a double-edged relational cue: signalling attentiveness (increasing disclosure and personalization) while creating a persistent record that undermines perceived control (reducing perceived safety, especially for sensitive topics).

3.1 Construct Definitions and Measurement Requirements

Disclosure Length: Operationalized as word count of free-text responses; a valid but narrow proxy for depth or quality of disclosure. Future work should supplement word count with independent coding of disclosure intimacy/sensitivity content.

Perceived safety, privacy control, personalization. Measured via single seven-point items in the simulated demonstration; single-item

measures are practical but typically show lower reliability and construct coverage than multi-item scales. A real study should pilot multi-item versions (3–5 items per construct) and report internal consistency and discriminant validity before relying on single items.

Topic sensitivity. Operationalized categorically (low/moderate/high) using researcher-assigned topics; a real study should include a manipulation check confirming participants' own sensitivity ratings align with the intended category, since sensitivity perceptions may vary across individuals and cultures.

(H1) Memory-on participants will produce longer disclosures and report higher perceived personalization than memory-off participants.

(H2) Memory-on participants will report lower perceived privacy control and, for sensitive topics, lower perceived safety.

(H3) The negative effect of memory on perceived safety is strongest for high-sensitivity topics and weakest for low-sensitivity topics.

These hypotheses have not yet been tested with real experimental data; Sections 4–5 demonstrate how such a test could be designed and analysed.

4. Methodology

This article adopts a conceptual and methodological approach to examine how conversational memory in generative AI may influence disclosure behaviour and perceived safety during human-AI interactions. Rather than presenting empirical findings, the purpose of this section is to outline a research design capable of investigating these relationships while addressing the operational and ethical requirements associated with memory manipulation.

The proposed framework is grounded in Communication Privacy Management Theory and the Computers-Are-Social-Actors paradigm, which together suggest that conversational memory functions as both a relational and privacy-relevant cue. Memory can increase personalization and continuity by allowing AI systems to retain information across interactions, while simultaneously influencing users' perceptions of privacy, control, and safety.

Future empirical studies could employ an experimental design that compares interactions with memory-enabled and memory-disabled AI systems. The purpose of such a design would not merely be to compare technological features but to examine how differing levels of perceived information retention influence disclosure decisions and privacy

perceptions. A crucial methodological requirement is that memory status be transparently communicated to participants and accurately reflected in the actual operation of the system. Participants should be informed whether information is retained across interactions and whether previously disclosed information can be recalled by the AI.

4.1 Operational and Ethical Requirements of the Memory Manipulation

The proposed manipulation requires clear operational procedures to ensure that memory-enabled and memory-disabled conditions are meaningfully distinct and ethically defensible.

In a memory-enabled condition, the AI system should be capable of retaining and referencing previously disclosed information within or across interactions. This capability should be explicitly disclosed to users prior to participation. In a memory-disabled condition, the AI system should neither store nor reference previously disclosed information, and participants should be informed of this limitation. Transparency is essential to prevent misleading participants regarding the system's actual data-handling practices.

Future studies should incorporate manipulation checks that assess whether participants perceive the intended difference between memory-enabled and memory-disabled systems. Such checks are necessary to establish that observed outcomes result from perceptions of conversational memory rather than unrelated aspects of the interaction.

Because investigations of disclosure may involve personal, health-related, financial, or otherwise sensitive information, researchers should establish clear policies regarding data protection, retention, deletion, participant withdrawal, and informed consent. These safeguards should be embedded within the research protocol before any data collection is conducted.

5. Results

The proposed framework conceptualizes conversational memory as a double-edged feature of generative AI systems. From a relational perspective, memory may signal attentiveness, continuity, and personalization, thereby encouraging users to engage more deeply with the AI system. From a privacy perspective, however, memory may also increase awareness that disclosed information can be retained and resurfaced in future interactions, potentially reducing users' sense of control over their personal information.

The first theoretical proposition is that conversational memory enhances perceived personalization. When users believe that an AI

system remembers previous interactions, they may interpret the interaction as more attentive, relevant, and individualized. This perception may strengthen relational engagement and encourage more extensive communication.

The second proposition is that conversational memory influences privacy perceptions. While memory can improve continuity and convenience, it may also raise concerns regarding data retention and future use of disclosed information. Users' willingness to disclose information may therefore be shaped not only by the benefits of personalization but also by perceptions of privacy control and informational vulnerability.

The third proposition concerns topic sensitivity. The influence of conversational memory is unlikely to be uniform across all disclosure contexts. Users may respond differently when discussing routine topics compared with highly personal topics involving health, finances, relationships, or emotional wellbeing. As topic sensitivity increases, privacy considerations may become more salient, potentially altering the relationship between memory, disclosure behaviour, and perceived safety.

Collectively, these propositions contribute to emerging discussions regarding trust, privacy, personalization, and human-AI communication. The framework suggests that conversational memory should not be regarded solely as a technological enhancement but rather as a social and psychological feature that simultaneously offers benefits and introduces risks. Future empirical research is required to determine how users balance these competing considerations in real-world interactions.

6. Discussion

If confirmed with real data, these findings would extend the disclosure-enhancing role of memory documented by Jo et al. (2024) - albeit from a 9-person interview sample - to a broader generative AI setting, while isolating a countervailing cost: reduced perceived privacy control. This dual pattern would be consistent with Jiang's (2024) conceptual characterization of AI disclosure as a paradox of trust and vulnerability. It would also partially align with anthropomorphism research, though this connection requires care: Ischen et al. (2020) found attentive cues increase disclosure by lowering privacy concern in the moment, consistent with CASA, but Pizzi et al. (2023) found anthropomorphism affected disclosure only indirectly, with no direct effect. Future work on memory should explicitly test direct versus indirect pathways rather than assuming a direct effect.

The moderating role of topic sensitivity would be instructive if confirmed: for low/moderate-sensitivity topics, memory's relational benefits may outweigh privacy costs, consistent with the resigned-but-tolerant attitudes documented by Kelley et al. (2023, N=10,011). For high-sensitivity topics, the narrowing disclosure gap would suggest users apply more deliberate privacy calculus, echoing Lappeman et al.'s (2022) finding that perceived risk suppresses disclosure independent of trust. These results would also speak to Lee et al.'s (2024) taxonomy distinguishing one-time collection from persistent-retention risks: it may not be disclosure itself that reduces safety, but the anticipation that information will resurface.

7. Practical Implications for Generative AI Design

The conceptual framework presented in this article highlights the need for responsible design of conversational memory features within generative AI systems. Because memory can influence both personalization and privacy perceptions, system designers should carefully consider how information is collected, retained, displayed, and controlled.

First, transparency should be treated as a core design principle. Users should be informed when memory features are active, what information is retained, how it is used, and under what circumstances it may be recalled. Memory status should be communicated clearly throughout the interaction rather than only during initial onboarding or consent procedures.

Second, users should be provided with meaningful control over stored information. Rather than relying solely on a universal memory setting, future systems may benefit from allowing users to determine which categories of information can be remembered, modified, or deleted. Such controls may help balance the benefits of personalization with users' expectations regarding privacy and autonomy.

Third, topic-sensitive memory policies merit consideration. Different types of information may carry different privacy expectations, and designers should recognize that users may be comfortable storing certain forms of information while preferring stricter protections for others. Adaptive memory-management strategies may therefore enhance user trust while reducing concerns about persistent information retention.

Finally, organizations deploying memory-enabled AI systems should develop governance frameworks that define acceptable memory use, retention practices, data-protection procedures, and accountability mechanisms. Responsible deployment requires balancing personalization

benefits with user privacy, transparency, and informed consent.

7.1 Ethical Safeguards for Sensitive-Topic Disclosure Research

Because conversational memory may affect disclosures involving sensitive personal information, ethical considerations should occupy a central role in both research and system design. Researchers and developers should avoid creating situations in which users feel pressured to reveal more information than they are comfortable sharing. Users should always be able to decline, skip, shorten, modify, or withdraw disclosures without negative consequences.

Transparency regarding data handling is equally important. Individuals should be informed how their disclosures are stored, how long the information is retained, and the procedures governing deletion or withdrawal requests. Clear communication of these processes promotes informed participation and supports user autonomy.

For studies involving potentially vulnerable individuals or highly sensitive topics, predefined procedures should exist to address disclosures indicating risk of harm. Such procedures should be integrated into ethics protocols before implementation and should be supported by qualified professionals where appropriate.

Ultimately, the ethical management of conversational memory requires attention to transparency, privacy, user autonomy, and accountability. These safeguards are not ancillary considerations but essential requirements for any future empirical investigation or real-world deployment involving memory-enabled generative AI systems

8. Limitations and Future Research

Sections 4–5 present simulated, not real, data; no findings can yet be drawn about real users. Single-item measures of perceived safety, privacy control, and personalization limit construct coverage and should be replaced with validated multi-item scales (Section 3.1). A controlled experimental setting may not capture the cumulative, longitudinal dynamics of memory accumulation described in naturalistic studies (Jo et al., 2024), especially given that study's small interview sample (n=9). The topic-sensitivity manipulation, while grounded in prior research, is coarse relative to individuals' fine-grained boundary rules and should be validated via participant-rated sensitivity checks (Section 3.1). Future research should track disclosure and perceived safety longitudinally across many sessions, examine individual differences in privacy orientation, and test topic-specific memory controls via field

experiments with deployed assistants. A further lesson from this revision is that internal numerical consistency (e.g., between t-statistics, sample sizes, and effect sizes) must be independently re-derived and checked before results are reported, a step that should be a standing part of the analysis workflow, not an afterthought.

9. Conclusion

Conversational memory is reshaping how users may disclose to generative AI systems, plausibly encouraging longer, more personalized exchanges while potentially eroding users' sense of control over their own information. Designers and researchers should treat memory not as an unambiguous improvement but as a disclosure–privacy tradeoff whose balance shifts with topic sensitivity. These propositions require empirical testing with real, ethically collected data - gathered using the operational safeguards in Section 4.1 and the ethical safeguards in Section 7.1 - before they can inform product design or policy with confidence.

Conflict of Interest

There is no conflict of interest between the authors in this manuscript. Results have been deduced objectively; nothing influenced this result by any individual, monetary, or institutional interests. In effect, utmost endeavors were in order to conduct a bias-free operation of gathering and analyzing the data into interpretation and producing a study based on merit, away from research inconsistencies.

References

1. Gumusel, E. (2024). A literature review of user privacy concerns in conversational chatbots: A social informatics approach. *Journal of the Association for Information Science and Technology*, 76, 121–154.
2. Handa, K., Stern, M., Huang, S., Hong, J., Durmus, E., McCain, M., Yun, G., Alt, A., Millar, T., Tamkin, A., Leibrock, J., Ritchie, S., & Ganguli, D. (2025). Introducing Anthropic Interviewer: What 1,250 professionals told us about working with AI. Anthropic.
3. Ischen, C., Araujo, T., Voorveld, H., van Noort, G., & Smit, E. (2020). Privacy concerns in chatbot interactions. In *Chatbot Research and Design* (pp. 34–48). Springer.
4. Jiang, Z. Z. (2024). Self-disclosure to AI: The paradox of trust and vulnerability in human-machine interactions. arXiv preprint, arXiv:2412.20564.
5. Jo, E., Jeong, Y., Park, S., Epstein, D. A., & Kim, Y.-H. (2024). Understanding the impact of long-term memory on self-disclosure with large language model-driven chatbots for public health intervention. In *Proceedings of the 2024 CHI Conference*.
6. Kelley, P. G., Cornejo, C., Hayes, L., Jin, E. S., Sedley, A., Thomas, K., Yang, Y., & Woodruff, A. (2023). "There will be less privacy, of course": How and why people in 10 countries expect AI will affect privacy in the future. *SOUPS 2023*.
7. Lappeman, J., Marlie, S., Johnson, T., & Poggenpoel, S. (2022). Trust and digital privacy: Willingness to disclose personal information to banking chatbot services. *Journal of Financial Services Marketing*, 28(2), 337–357.
8. Lee, H.-P., Yang, Y.-J., von Davier, T. S., Forlizzi, J., & Das, S. (2024). Deepfakes, phrenology, surveillance, and more: A taxonomy of AI privacy risks. In *Proceedings of the 2024 CHI Conference*.
9. Maeda, T., & Quan-Haase, A. (2024). When human-AI interactions become parasocial: Agency and anthropomorphism in affective design. In *Proceedings of the 2024 ACM FAccT Conference* (pp. 1068–1077).
10. Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 72–78). ACM.
11. Petronio, S. (2002). *Boundaries of privacy: Dialectics of disclosure*. State University of New York Press.
12. Pizzi, G., Vannucci, V., Mazzoli, V., & Donvito, R. (2023). "I, chatbot!" The impact of anthropomorphism and gaze direction on willingness to disclose personal information and behavioral intentions. *Psychology & Marketing*, 40(7), 1372–1391.
13. Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.