Workshop paper presented at Bias4Trust — a CHI 2026 workshop.

How does Chatbot Memory and Framing Affect People's Self-Disclosure Decisions?

SAMUEL RHYS COX, Aalborg University, Denmark
MICHAEL YIN, University of British Columbia, Canada
NIELS VAN BERKEL, Aalborg University, Denmark

1 Introduction

Self-disclosure (the sharing of one's thoughts, feelings, and opinions [2]) is often recommended by mental health experts as it can provide emotional relief, facilitate self-reflection, and encourage social support and bonding [9]. Despite these benefits, we may often prove reluctant to disclose due to perceived risks, such as fear of being judged or stereotyped, or concern that the recipient will not be sufficiently supportive [31]. This reluctance highlights the decision-making processes we follow in determining what to share, with whom to share it, and whether a given context is considered appropriate for sharing.

Calling upon Social Penetration Theory, self-disclosure decisions are commonly described using a risk-reward calculus [1]. For example, we may perceive lower risk when disclosing within established relationships where trust has developed over time, whereas we may perceive greater risk in disclosing to work colleagues if doing so disrupts social norms. Beyond this (perhaps more reflective) risk-reward calculus, self-disclosure has also been attributed to impulsiveness-based decision criteria [22]. Within Human-Computer Interaction (HCI) this has led to exploration of heuristics (i.e., 'rule of thumb' cognitive shortcuts), especially within contexts of sharing information online [12, 29]. For example, John et al. [12] found that visual cues on websites affected disclosure. Participants disclosed more incriminating behaviours when a survey used a playful heading ("*How BAD Are U???*") alongside a cartoon devil than when the survey appeared to be associated with Carnegie Mellon University, suggesting the influence of authority bias.

Conversational AI systems are a particularly salient form of digitally mediated disclosure contexts. As more people turn to chatbots (such as ChatGPT) for emotional support and help-seeking [6, 14], the factors that affect self-disclosure decisions (especially sensitive health and well-being disclosure) are increasingly pertinent. The same calculus and heuristics described above have also been observed within human–chatbot interactions. For instance, our decisions about whether to self-disclose could be influenced by: whether we are talking to a human or chatbot (with a "*machine heuristic*" leading people to disclose more to chatbots) [28]; the conversational style of a chatbot [7]; and the memory capabilities of chatbots, including whether and how prior disclosures are retained and referenced across sessions [4, 5].

Related to this last factor, persistent cross-session memory is emerging as a dominant design paradigm in commercial conversational AI systems, with companies such as OpenAI and Google introducing memory features into their consumer-facing conversational AI platforms. Work has shown that conversational memory can help develop feelings of closeness and foster trust in human–AI relationships [5, 11]. However, work has also found that chatbot memory may also make people feel judged when disclosing emotional concerns [4] or raise privacy concerns when disclosing health behaviours [5, 11]. In light of these alternative paradigms — **persistent versus ephemeral memory** — we consider how memory configuration and its framing to users may affect disclosure decisions. Specifically, we discuss how ephemerality cues may be designed and framed to activate or mitigate cognitive heuristics that shape comfort,

Authors' Contact Information: Samuel Rhys Cox, srcox@cs.aau.dk, Aalborg University, Aalborg, Denmark; Michael Yin, jiyin@cs.ubc.ca, University of British Columbia, Vancouver, Canada; Niels van Berkel, nielsvanberkel@cs.aau.dk, Aalborg University, Aalborg, Denmark.

perceived risk, and self-disclosure in health and well-being contexts. By conceptualising conversational memory as a factor that amplifies or mitigates cognitive biases, we aim to inform bias-aware design strategies for emotionally sensitive human–AI interactions.

2 Persistent Memory and Self-Disclosure

First, we discuss the emerging paradigm of chatbots with memory between sessions, and its potential impact on self-disclosure decisions. Here, a *persistent chatbot* refers to a system with cross-session memory that retains and references prior interactions, and is framed to users as a continuous conversational entity across repeated encounters.

Although often not explored as a comparison of persistent versus ephemeral chatbots, prior work has explored the formation of human–chatbot relationships over longitudinal interactions. For example (while participants criticised "*Replika's poor memory or communication skills*" [26]) Skjuve et al. investigated longitudinal human–chabot relationship development in two studies [26, 27]. Here, both studies found that people can develop a sense of closeness and intimacy with chatbots over time, and that (similarly to human–human relationships [1]) self-disclosure can help human–chatbot relationships develop. However, longitudinal relationship development is not guaranteed. In a pre-LLM chatbot study, Croes and Antheunis [8] found that feelings of closeness decreased over time, which participants attributed to repetitive interactions and diminishing novelty. This suggests that repeated interactions with the same chatbot do not necessarily lead to closeness, particularly when conversational continuity is limited.

Empirical studies of chatbots with cross-session memory suggest a trade-off: retaining prior interactions can increase perceived intelligence, support accountability in behaviour change contexts [5], and foster familiarity and sharing [11], while also raising privacy concerns among users [5, 11]. Cox et al. have explored user perceptions in explicit comparisons of persistent versus ephemeral chatbots [4, 5]. Here, they found that if the development of a human–chatbot relationship does not follow socially expected progress it can lead to expectancy violations. In particular, when a chatbot was framed as a persistent "*companion*", participants reported lower comfort with disclosure when emotionally sensitive questions were asked during the initial interaction [4]. This early elicitation of emotional disclosure was perceived as premature, and inconsistent with socially expected patterns of relationship development, where disclosure typically progresses from superficial to more personal topics over time, as described by Social Penetration Theory [1]. However, when the first interaction focused on more factual or impersonal questions, followed by emotionally oriented questions in a subsequent session, participants reported greater comfort, enjoyment, and willingness to continue interacting with the chatbot. In this case, the perceived discrepancy between persistent and ephemeral chatbots diminished, suggesting that adherence to expected patterns of relational progression may play an important role in shaping disclosure comfort in persistent conversational systems. These findings suggest that persistent memory may activate relational scripts and expectancy-based heuristics, where users anticipate a gradual progression of disclosure, and deviations from these expectations can reduce comfort and willingness to share.

It should be noted, however, that these studies (i.e., [4, 5]) primarily examined perceived comfort and privacy concerns related to self-disclosure, alongside social qualities of the chatbot, rather than directly measuring the breadth or depth of participants' self-disclosure, as has been done in other conversational disclosure research [7, 10]. This highlights an important area for future work in understanding how persistent memory influences not only perceptions of disclosure comfort, but disclosure behaviour itself. Notably, participants interacting with ephemeral chatbots in Cox et al. [4] also qualitatively described feeling less judged when sharing emotional concerns, suggesting that persistent memory may introduce additional perceived evaluative pressure during disclosure.

More broadly, these findings suggest that persistent memory may alter disclosure decisions by creating a sense of conversational continuity, which can increase both relational attachment and sensitivity to how one's disclosures are remembered and referenced over time.

3 Ephemeral Memory and Self-Disclosure

In contrast to persistent chatbots, we now consider chatbots with ephemeral memory, and how the absence of cross-session retention may influence self-disclosure decisions. Here, an *ephemeral chatbot* refers to a system in which conversational content is not retained across sessions, and interactions are framed as discrete, moment-bound exchanges rather than part of an ongoing conversational history.

Complementing the findings described in the previous section, Cox et al. also observed distinct disclosure experiences when chatbots were framed as ephemeral rather than persistent [4]. Participants frequently described ephemeral chatbots using metaphors such as diary-keeping or journaling, suggesting that interactions were perceived as more private and reflective. In this framing, participants also reported feeling less judged by the chatbot, feelings of anonymity, and greater comfort when emotionally sensitive questions were asked during the initial interaction. These findings indicate that ephemerality may reduce perceived evaluative pressure and enable earlier or more exploratory forms of emotional disclosure. Similar effects have been observed in other disclosure contexts. For example, Park et al. found that when disclosing traumatic experiences, people felt less writing difficulty and there was greater emotional disclosure when there was *not* a responsive follow-up based on user input [23].

Although chatbots explicitly framed as ephemeral have been less widely explored, prior work on ephemeral social media sharing (e.g., Instagram Stories and Snapchat) has found that ephemerality lowers inhibition in sharing, as people feel less pressure to maintain a consistent self-presentation and greater freedom to share more authentic experiences [3, 33]. Feelings of anonymity and lessened self-awareness can encourage self-disclosure [13], and feelings of anonymity on social media have also been found to affect *what* we share, with Ma et al. finding that people share more negative-valence experiences, such as bad experiences in romantic relationships [17].

Ephemeral memory therefore appears to shift disclosure towards more exploratory and less self-curated expression by reducing perceived audience persistence and evaluative pressure, enabling users to engage in disclosure with less concern for long-term interpretation or identity consistency.

4 Memory as a Design Continuum

Taking both our prior discussions of persistence and ephemerality together, conversational memory can be understood as a design space spanning a continuum between persistent and ephemeral interaction. That is to say, rather than holding a strict binary view of persistent or ephemeral interactions, conversational memory can be designed along a continuum. Systems may selectively retain, abstract, omit, or visually represent conversational history in ways that influence feelings of continuity and comfort in self-disclosing.

Selective and User-Controlled Memory: Along this continuum, users could be given greater control over which conversational information is retained. For example, users could selectively save specific disclosures for future reference while allowing other interactions to remain ephemeral, or configure systems to avoid retaining sensitive topics such as emotional or health-related concerns. Such selective retention may increase users' sense of agency and predictability, reducing uncertainty about how disclosures will be remembered and used. This approach may be particularly relevant in contexts such as LLM-driven journaling and reflective writing [15, 20], where users may benefit from preserving

insights over time while maintaining privacy over more exploratory thoughts. More broadly, prior work in HCI has shown that providing users with control over data retention and visibility can reduce privacy concerns and increase comfort with interactive systems [16, 24, 30]. However, there is potential for interactions to feel less 'organic' if user choice of memories is seen an mechanistic.

Memory of Conversation Timing, but Not Content: A sense of continuity could be preserved by remembering when interactions occurred, without retaining the content of the conversations themselves. For example, in a conversational system designed for people disclosing emotional challenges, users may prefer that conversational content remains ephemeral. By keeping a log of when users complete conversations (such as in a calendar view) a sense of progress and commitment can be maintained while still allowing chatbot interactions to remain ephemeral.

Models of Memory and Decay: Chatbot memory could follow different models of memory in contrast to a persistent memory where all conversations are retained. This could take influence from prior discussions within social media literature regarding whether our personal data should "*decay*" over time [18, 32], as well as literature describing "human-like" models of forgetting [19]. Visual cues can also be manipulated to represent decay. For example, recent work has explored how the ephemerality of disclosures can be represented via animated visual cues, such as messages fading gradually over time [34], or being visually 'torn' up on screen [21].

5 Implications for Trust, Disclosure, and Bias-Aware Memory Design

The examples in § 4 highlight that conversational agents (CAs) have the affordance to 'shift' between persistent and ephemeral forms of interaction. For example, a CA could have an 'off-the-record' conversation with users, where the same established persistent personality and references to prior conversations is used, but the ultimate contents of the conversation is not retained.

Additionally, we highlight that users' loss aversion can be influenced by how ephemerality is framed. Specifically, whether the absence of memory is presented as a user benefit or as a loss of chatbot capability may shape user perceptions and behaviour. For instance, when ephemerality was framed as beneficial, users reported feeling more comfortable and less judged when discussing emotional concerns [4]. In contrast, when no benefits or rationale were provided, users perceived an ephemeral chatbot as less intelligent and capable [5].

Conversational memory also influences what users attend to and reflect upon during interaction. Persistent conversational histories may reinforce prior disclosures by making them visible and accessible, potentially encouraging reflection, accountability, or behaviour change. Conversely, ephemeral interactions may reduce users' attention to past disclosures, enabling more exploratory and less constrained forms of expression. Prior work in cognitive bias modification has demonstrated that directing attention toward or away from certain cues can influence behaviour and self-perception [25].

Finally, ephemeral memory may increase users' comfort with disclosure by reducing perceived persistence, yet conversational data may still be processed or retained in ways not visible to users. If systems create a perception of ephemerality without clearly communicating actual data practices, users may disclose sensitive information under assumptions that do not reflect system behaviour. Such mismatches between perceived and actual memory introduce risks of dark patterns, underscoring the importance of transparent memory design that supports informed trust and disclosure decisions.

Acknowledgments

This work was supported by a CAISA Fellowship from the National Center for AI in Society (CAISA; Det Nationale Center for AI i Samfundet).

References

[1] Irwin Altman and Dalmas Taylor. 1973. Social penetration: The development of interpersonal relationships. (01 1973). https://psycnet.apa.org/record/1973-28661-000

[2] Richard L Archer and Joseph A Burleson. 1980. The Effects of Timing of Self-Disclosure on Attraction and Reciprocity. *Journal of Personality and Social Psychology* 38, 1 (1980), 120. doi:10.1037/0022-3514.38.1.120

[3] Joseph B Bayer, Nicole B Ellison, Sarita Y Schoenebeck, and Emily B Falk. 2016. Sharing the small moments: ephemeral social interaction on Snapchat. *Information, Communication & Society* 19, 7 (2016), 956–977. doi:10.1080/1369118X.2015.1084349

[4] Samuel Rhys Cox, Rune Møberg Jacobsen, and Niels van Berkel. 2025. The Impact of a Chatbot's Ephemerality-Framing on Self-Disclosure Perceptions. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25)*. Association for Computing Machinery, New York, NY, USA, Article 60, 17 pages. doi:10.1145/3719160.3736617

[5] Samuel Rhys Cox, Yi-Chieh Lee, and Wei Tsang Ooi. 2023. Comparing How a Chatbot References User Utterances from Previous Chatting Sessions: An Investigation of Users' Privacy Concerns and Perceptions. In *Proceedings of the 11th International Conference on Human-Agent Interaction* (Gothenburg, Sweden) *(HAI '23)*. Association for Computing Machinery, New York, NY, USA, 105–114. doi:10.1145/3623809.3623875

[6] Samuel Rhys Cox, Jade Martin-Lise, Simo Hosio, and Niels van Berkel. 2026. Watching AI Think: User Perceptions of Visible Thinking in Chatbots. *arXiv preprint arXiv:2601.16720* (2026). doi:10.48550/arXiv.2601.16720

[7] Samuel Rhys Cox and Wei Tsang Ooi. 2022. Does Chatbot Language Formality Affect Users' Self-Disclosure?. In *Proceedings of the 4th Conference on Conversational User Interfaces* (Glasgow, United Kingdom) *(CUI '22)*. Association for Computing Machinery, New York, NY, USA, Article 1, 13 pages. doi:10.1145/3543829.3543831

[8] Emmelyn AJ Croes and Marjolijn L Antheunis. 2021. Can we be friends with Mitsuku? A longitudinal study on the process of relationship formation between humans and a social chatbot. *Journal of Social and Personal Relationships* 38, 1 (2021), 279–300. doi:10.1177/0265407520959463

[9] Kathryn Greene, Valerian J Derlega, and Alicia Mathews. 2006. Self-Disclosure in Personal Relationships. *The Cambridge Handbook of Personal Relationships* 409 (2006), 427. doi:10.1017/CBO9780511606632.023

[10] Rune Møberg Jacobsen, Samuel Rhys Cox, Carla F. Griggio, and Niels van Berkel. 2025. Chatbots for Data Collection in Surveys: A Comparison of Four Theory-Based Interview Probes. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 228, 21 pages. doi:10.1145/3706598.3714128

[11] Eunkyung Jo, Yuin Jeong, Sohyun Park, Daniel A. Epstein, and Young-Ho Kim. 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 on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 440, 21 pages. doi:10.1145/3613904.3642420

[12] Leslie K John, Alessandro Acquisti, and George Loewenstein. 2011. Strangers on a Plane: Context-Dependent Willingness to Divulge Sensitive Information Purchased. *Journal of Consumer Research* 37, 5 (2011), 858–873. doi:10.1086/656423

[13] Adam N Joinson. 2001. Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity. *European Journal of Social Psychology* 31, 2 (2001), 177–192. https://doi.org/10.1002/ejsp.36

[14] Kyuha Jung, Gyuho Lee, Yuanhui Huang, and Yunan Chen. 2025. "I've Talked to ChatGPT About My Issues Last Night.": Examining Mental Health Conversations with Large Language Models Through Reddit Analysis. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW356 (Oct. 2025), 25 pages. doi:10.1145/3757537

[15] Taewan Kim, Donghoon Shin, Young-Ho Kim, and Hwajung Hong. 2024. DiaryMate: Understanding User Perceptions and Experience in Human-AI Collaboration for Personal Journaling. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1046, 15 pages. doi:10.1145/3613904.3642693

[16] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, Are You Listening?: Privacy Perceptions, Concerns and Privacy-seeking Behaviors with Smart Speakers. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–31. doi:10.1145/3274371

[17] Xiao Ma, Jeff Hancock, and Mor Naaman. 2016. Anonymity, Intimacy and Self-Disclosure in Social Media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 3857–3869. doi:10.1145/2858036.2858414

[18] Reham Mohamed, Paulina Chametka, and Sonia Chiasson. 2020. The Influence of Decaying the Representation of Older Social Media Content on Simulated Hiring Decisions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–19. doi:10.1145/3313831.3376346

[19] Reham Ebada Mohamed and Sonia Chiasson. 2018. Online privacy and aging of digital artifacts. In *Proceedings of the Fourteenth USENIX Conference on Usable Privacy and Security* (Baltimore, MD, USA) *(SOUPS '18)*. USENIX Association, USA, 177–195. https://dl.acm.org/doi/10.5555/3291228.3291243

[20] Subigya Nepal, Arvind Pillai, William Campbell, Talie Massachi, Michael V. Heinz, Ashmita Kunwar, Eunsol Soul Choi, Xuhai Xu, Joanna Kuc, Jeremy F. Huckins, Jason Holden, Sarah M. Preum, Colin Depp, Nicholas Jacobson, Mary P. Czerwinski, Eric Granholm, and Andrew T. Campbell. 2024. MindScape Study: Integrating LLM and Behavioral Sensing for Personalized AI-Driven Journaling Experiences. *Proc. ACM Interact. Mob.*

Wearable Ubiquitous Technol. 8, 4, Article 186 (Nov. 2024), 44 pages. doi:10.1145/3699761

[21] Shunpei Norihama, Shixian Geng, Kakeru Miyazaki, Arissa J. Sato, Mari Hirano, Simo Hosio, and Koji Yatani. 2025. Examining Input Modalities and Visual Feedback Designs in Mobile Expressive Writing. *Proc. ACM Hum.-Comput. Interact.* 9, 5, Article MHCI009 (Sept. 2025), 28 pages. doi:10.1145/3743723

[22] Sina Ostendorf, Yannic Meier, and Matthias Brand. 2022. Self-Disclosure on Social Networks: More Than a Rational Decision-Making Process. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 16, 4 (2022). doi:10.5817/CP2022-4-2

[23] SoHyun Park, Anja Thieme, Jeongyun Han, Sungwoo Lee, Wonjong Rhee, and Bongwon Suh. 2021. "I wrote as if I were telling a story to someone I knew.": Designing Chatbot Interactions for Expressive Writing in Mental Health. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference* (Virtual Event, USA) *(DIS '21)*. Association for Computing Machinery, New York, NY, USA, 926–941. doi:10.1145/3461778.3462143

[24] Rachel Phinnemore, Mohi Reza, Blaine Lewis, Karthik Mahadevan, Bryan Wang, Michelle Annett, and Daniel Wigdor. 2023. Creepy Assistant: Development and Validation of a Scale to Measure the Perceived Creepiness of Voice Assistants. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18. doi:10.1145/3544548.3581346

[25] Charlie Pinder, Jo Vermeulen, Benjamin R. Cowan, and Russell Beale. 2018. Digital Behaviour Change Interventions to Break and Form Habits. *ACM Trans. Comput.-Hum. Interact.* 25, 3, Article 15 (June 2018), 66 pages. doi:10.1145/3196830

[26] Marita Skjuve, Asbjørn Følstad, and Petter Bae Brandtzæg. 2023. A Longitudinal Study of Self-Disclosure in Human–Chatbot Relationships. *Interacting with Computers* 35, 1 (2023), 24–39. https://doi.org/10.1093/iwc/iwad022

[27] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2022. A longitudinal study of human–chatbot relationships. *International Journal of Human-Computer Studies* 168 (2022), 102903. doi:10.1016/j.ijhcs.2022.102903

[28] S. Shyam Sundar and Jinyoung Kim. 2019. Machine Heuristic: When We Trust Computers More than Humans with Our Personal Information. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–9. doi:10.1145/3290605.3300768

[29] S. Shyam Sundar, Jinyoung Kim, Mary Beth Rosson, and Maria D. Molina. 2020. Online Privacy Heuristics that Predict Information Disclosure. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3313831.3376854

[30] Alice Thudt, Dominikus Baur, Samuel Huron, and Sheelagh Carpendale. 2015. Visual Mementos: Reflecting Memories with Personal Data. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2015), 369–378. doi:10.1109/TVCG.2015.2467831

[31] David L Vogel and Stephen R Wester. 2003. To Seek Help or Not to Seek Help: The Risks of Self-Disclosure. *Journal of Counseling Psychology* 50, 3 (2003), 351. doi:10.1037/0022-0167.50.3.351

[32] Toshihiko Yamakami. 2010. We should forget the search results from far past: A computer ethical view in the era of search engines. In *6th International Conference on Digital Content, Multimedia Technology and its Applications*. IEEE, 198–202. https://ieeexplore.ieee.org/abstract/document/5568706

[33] Yueyang Yao, Samuel Hardman Taylor, and Sarah Leiser Ransom. 2024. Who's Viewing My Post? Extending the Imagined Audience Process Model Toward Affordances and Self-Disclosure Goals on Social Media. *Social Media+ Society* 10, 1 (2024). doi:10.1177/20563051231224271

[34] Michael Yin and Robert Xiao. 2026. The Words That Can't Be Shared: Exploring the Design of Unsent Messages. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. doi:10.1145/3772318.3790639