Theory of Mind and Self-Disclosure to CUIs

Samuel Rhys Cox
srcox@cs.aau.dk
Aalborg University
Aalborg, Denmark

Abstract

Self-disclosure is important to help us feel better, yet is often difficult. This difficulty can arise from how we *think* people are going to react to our self-disclosure. In this workshop paper, we briefly discuss self-disclosure to conversational user interfaces (CUIs) in relation to various social cues. We then, discuss how expressions of uncertainty or representation of a CUI's reasoning could help encourage self-disclosure.

ACM Reference Format:
Samuel Rhys Cox. 2025. Theory of Mind and Self-Disclosure to CUIs. In *Proceedings of ToMinHAI at CUI'2025: Theory of Mind in Human-CUI Interaction (ToMinHAI at CUI'2025).* ACM, New York, NY, USA, 4 pages.

1 Introduction

Self-disclosure (the revealing of personal information such as thoughts, feelings, and opinions [5]) has been shown to benefit both emotional and physical well-being: it can reduce stress, promote self-reflection, and encourage social support [14]. However, we may not always feel confident and comfortable when self-disclosing if we perceive the *risks* of disclosing as outweighing any potential *benefits* of self-disclosure [3, 24].

These perceived risks may stem from how we *think* our interlocutor will react. For example, we may think that they will judge or stereotype us, or that they will not be receptive to our disclosures [30]. These form the *Theory of Mind* (ToM) that we hold for others. That is: our ability to attribute mental states such as beliefs, desires, intentions, and emotions to others.

In this workshop paper, we will discuss how ToM can help us design conversational user interfaces (CUIs) as avenues of self-disclosure. CUIs (such as text-based conversational agents (CAs) or voice-based assistants) are increasingly used by people as a means of companionship and emotional support [11, 17], and in April 2025 Harvard Business Review found therapy and companionship to be the number one use of generative AI [33]. Motivated by this, we discuss prior work investigating self-disclosure to CUIs. Here, we explore how the social cues of a CUI (such as verbal, visual, auditory, or invisible cues [13]) can be varied, and how this may affect people's ToM that results from perceived risks of self-disclosure.

2 CUIs for Self-Disclosure

The social cues of a CUI can be adapted to encourage people's self-disclosure. A common thread among some prior work is that these social cues can be adapted to reduce people's *feelings of judgement* from the CUI. E.g., people's beliefs that:

"*I think that this chatbot will judge me*"

vs.

"*I think that this chatbot will* ***not*** *judge me*"

For example, studies have found that people may feel less judged when disclosing to a CUI compared to a human [16, 26], with common threads of research connecting the level of social presence to feelings of judgement (e.g., people *may* feel more judged with higher levels of social presence, such as more anthropomorphised visual cues [25], or precise recall of past user utterances [9]). Demonstrating this, Pickard et al. found that people were more likely to disclose shame to a disembodied voice agent compared to both an embodied CA or a human interviewer [25]. Similarly, Wang et al. found people were more willing to disclose to an "*algorithm*" than a person or AI assistant [31].

Studies have also investigated how changing verbal cues in both terms of conversational content [8, 9, 16] and conversational style [7, 10] can affect self-disclosure. For example, studies have investigated the effect of reciprocal self-disclosure (motivated by Social Penetration Theory [3]) whereby people disclose more if a chatbot itself discloses to them [1, 18, 21, 27]. Recent work has also investigated the effect of multiple social cues (such as interaction effects of both visual and verbal cues [7]) reflecting the breadth and complication of potential designs.

Finally, the beliefs of the user themselves effect people's likelihood to self-disclose. For example, Sundar and Kim found that as people's belief in the machine heuristic (a belief that machines are more secure and trustworthy than humans) increased, so too did their likelihood to disclose personal information to a CUI [29]. Additionally, our own work has found that people's views of chatbots as either a tool or a conversational partner affects preferred social cues [9]. Further, people's pre-existing beliefs regarding CUI emotional capacity and intelligence impact cue effectiveness [10, 19].

While some results may seem somewhat contradictory (e.g., some studies have conflicting results regarding the impact of a CUI's social presence on self-disclosure) we draw on the common thread that people often do not self-disclose due to perceived risks and feelings of judgement from CUIs, as well as apprehension to self-disclose often being mediated by people's personal beliefs.

3 Expressions of Uncertainty

Prior research and theories have posited that people generally dislike uncertainty in communication, and that more concrete and unambiguous information will drive communication [6]. Connected to this, some barriers to self-disclosure are related to the uncertainty regarding the beliefs of your interlocutor (e.g., "*is my interlocutor likely to stigmatise me?*"). Uncertainty such as this can produce anxiety and disrupt self-disclosure. For example, Sien and McGrener investigated people sharing stories about their mental health within a community of users. They found that people (who initially thought that others would judge or stigmatise them) felt

ToMinHAI at CUI'2025, Waterloo, ON, Canada
2025.

more relieved of anxiety and open to social support after reading concrete stories from others [28].

However, human-CUI interactions (like human-human interactions) may be operating on levels of incomplete information [12]. This could lead either the CUI or the human user to incorporate false beliefs in their ToM. False beliefs in each interlocutor's ToM could then lead to misunderstandings, frustration, and feelings of not being understood or appreciated. In natural language conversations, an interlocutor could simply transparently state that they have a level of uncertainty in communication, and therefore seek clarification. Yet, depending on context, different response styles could lead to potentially negative user perceptions [32]).

In addition to increasingly capable natural language responses (given the capabilities of LLMs), CUIs offer the affordance to reply using a variety of cues not available in normal human-human conversations. For example, uncertainty could be expressed visually via changing background colours of the CUI (e.g., with colour attribution to indicate less confidence and understanding in the CUI's ToM). Additionally, (while perhaps seeming less exciting than a fully open-ended natural language LLM-driven conversation) a CUI could seek clarification or express uncertainty using other means such as multiple choice responses (see Figure 1).

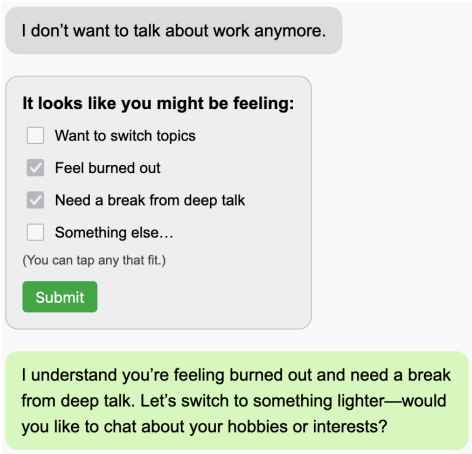

Figure 1: CUIs have different affordances to human-human conversations to express uncertainty. For example, the ability to use full natural language responses, multiple choice responses, or visual cues.

While use of multiple choice responses in Figure 1 may seem like a bit of a "*step back*" for CUIs (with this Figure offered perhaps as more of a provocation), some recent work indicates that people may not want more capable and personalised CUI responses when they are disclosing sensitive emotional information [8, 9, 23]. For example, in our recent work people who talked to chatbot that did not have memory between chatting sessions described not feeling judged during interactions (e.g., one user stated: "*I didn't feel judged* [...] *I felt like I was just journaling with helpful prompts being offered along the way*") [8].

4 Exposing the CUI's Internal ToM

When interacting with CUIs (such as ChatGPT) they may display "reasoning" to users while responses are being processed. Such reasoning can form explanations of a CUI's intent, and could transparently share the CUI's ToM with the user. For example, a CUI could could directly refer to the user's fear of judgement in its reasoning before providing a response:

> "*I think the user may have an emotional need that they wish to discuss. I will not judge them, and do not want them to think I would react negatively to what they will say. When replying to the user, I should...*"

Conversational examples of such reasoning that disclose the CUI's ToM are shown in Figures 2 and 3. Here, the CUI's chain of thought (that includes both ToM attribution, and related plans to respond) could be shared explicitly with users. The example text in both Figures 2 and 3 is the same, with just the visual representation of "thoughts" being adapted between more or less anthropomorphised cues (somewhat similarly to prior work investigating the use of comic style speech bubbles for text-based CUIs [4]). Although these examples use natural language, reasoning could be adapted to include more "machine-like" explanations, such as sentiment scores (e.g., "*My sentiment model scores this as high in shame (0.82)*").

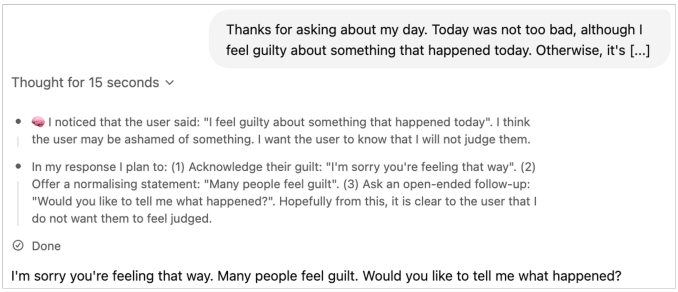

Figure 2: A "standard" reasoning GUI (i.e., mirroring ChatGPT's reasoning).

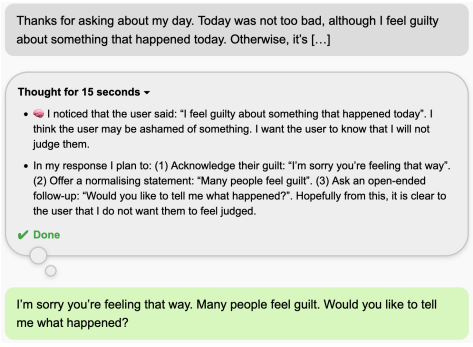

Figure 3: Anthropomorphised "thought-bubble" reasoning GUI.

Such reasoning could also follow different models of representation. For example, perfect memory and recall could be incorporated potentially increasing users' feelings of being understood

and known (e.g., users thinking: "*The CUI really remembers me*"), although perfect recall has also been shown to raise privacy concerns [9]. Alternatively, reasoning (together with a CUI's ToM) could have imperfect memory, and decay in line with people's perceptions of human memories [20].

CUI (or LLM) reasoning could also mirror models of how people perceive reasoning as being structured [22]. Different types and formats of inner dialogue and self-talk could be used to represent both the CUI's reasoning and related ToM. For example, (as discussed by Oleś et al. [22]) Dialogical Self Theory [15] postulates that people's thoughts can be represented by multiple "inner voices", each of which can hold different roles within intrapersonal communication. These different "I-position" roles could form different aspects to represent the CUI's ToM to the user, as well as expressing desires in relation to the user (i.e., "*I will not judge the user*").

4.1 Bidirectional Attribution of ToM

These different modes of expressing ToM (either through potentially more or less levels of personalisation) can then assist in facilitating a continuous bidirectional attribution of ToM (i.e., Mutual Theory of Mind). Rather than a 'one-way' approach to CUIs (where a CUI may "empathy bomb" a user and lead to potentially negative perceptions [2]), a CUI could have an internal ToM pipeline that detects user cues (such as hesitation and sentiment), and chooses the next cue accordingly. By the CUI updating its cues to match its ToM related to the user, the user can then in turn perceive the CUI itself more positively thereby updating their own ToM (i.e., "*This CUI really understands me*", "*This CUI will not judge me*").

5 Limitations

We note that not all studies of self-disclosure to CUIs frame their results around users' perceived risks of disclosure (i.e., their ToM regarding risks, such as whether they *think* their interlocutor will judge them). However, we frame this work within widely accepted theory regarding self-disclosure, where a calculus of perceived risks to benefits is used to determine the breadth and depth of disclosure to an interlocutor [3]. We also appreciate that there can still be conflicting findings within self-disclosure literature (such as conflicting findings regarding the effects of different levels of social presence [9, see Sec.2]). However, (as discussed briefly above) self-disclosure can be affected by user beliefs, social cues, and user context. From this, we focus on perceived *risks* of self-disclosure (such as risks of judgement) that can be applied across these findings.

Acknowledgments

This work is supported by the Carlsberg Foundation, grant CF21-0159.

References

[1] Martin Adam and Johannes Klumpe. 2019. Onboarding with a Chat–The Effects of Message Interactivity and Platform Self-Disclosure on User Disclosure Propensity. *Proceedings of the 27th European Conference on Information Systems (ECIS)* (May 2019). https://aisel.aisnet.org/ecis2019_rp/68

[2] Lize Alberts, Ulrik Lyngs, and Max Van Kleek. 2024. Computers as bad social actors: Dark patterns and anti-patterns in interfaces that act socially. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–25.

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

[4] Toshiki Aoki, Rintaro Chujo, Katsufumi Matsui, Saemi Choi, and Ari Hautasaari. 2022. EmoBalloon - Conveying Emotional Arousal in Text Chats with Speech Balloons. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 527, 16 pages. https://doi.org/10.1145/3491102.3501920

[5] 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. https://doi.org/10.1037/0022-3514.38.1.120

[6] Charles R Berger and Richard J Calabrese. 1974. Some explorations in initial interaction and beyond: Toward a developmental theory of interpersonal communication. *Human Communication Research* 1, 2 (1974), 99–112. https://doi.org/10.1111/j.1468-2958.1975.tb00258.x

[7] Jiahao Chen, Mingming Li, and Jaap Ham. 2024. Different dimensions of anthropomorphic design cues: How visual appearance and conversational style influence users' information disclosure tendency towards chatbots. *International Journal of Human-Computer Studies* (2024), 103320. https://doi.org/10.1016/j.ijhcs.2024.103320

[8] 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 Conference on Conversational User Interfaces (CUI '25)*. https://doi.org/10.48550/arXiv.2505.20464

[9] 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*. https://doi.org/10.1145/3623809.3623875

[10] 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*. 1–13. https://doi.org/10.1145/3543829.3543831

[11] 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. https://doi.org/10.1177/0265407520959463

[12] Harmen De Weerd, Rineke Verbrugge, and Bart Verheij. 2017. Negotiating with other minds: The role of recursive theory of mind in negotiation with incomplete information. *Autonomous Agents and Multi-Agent Systems* 31 (2017), 250–287. https://doi.org/10.1007/s10458-015-9317-1

[13] Jasper Feine, Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2019. A Taxonomy of Social Cues for Conversational Agents. *International Journal of Human-Computer Studies* 132 (2019), 138–161. https://doi.org/10.1016/j.ijhcs.2019.07.009

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

[15] Hubert JM Hermans and Giancarlo Dimaggio. 2007. Self, identity, and globalization in times of uncertainty: A dialogical analysis. *Review of general psychology* 11, 1 (2007), 31–61.

[16] 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*. https://doi.org/10.1145/3706598.3714128

[17] 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*. https://doi.org/10.1145/3613904.3642420

[18] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. "I Hear You, I Feel You": Encouraging Deep Self-disclosure through a Chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12. https://doi.org/10.1145/3313831.3376175

[19] Bingjie Liu and S Shyam Sundar. 2018. Should Machines Express Sympathy and Empathy? Experiments with a Health Advice Chatbot. *Cyberpsychology, Behavior, and Social Networking* 21, 10 (2018), 625–636. https://doi.org/10.1089/cyber.2018.0110

[20] 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

[21] Youngme Moon. 2000. Intimate Exchanges: Using Computers to Elicit Self-Disclosure from Consumers. *Journal of Consumer Research* 26, 4 (2000), 323–339. https://doi.org/10.1086/209566

[22] Piotr K Oleś, Thomas M Brinthaupt, Rachel Dier, and Dominika Polak. 2020. Types of inner dialogues and functions of self-talk: Comparisons and implications. *Frontiers in Psychology* 11 (2020), 486136. https://doi.org/10.3389/fpsyg.2020.

00227

[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. https://doi.org/10.1145/3461778.3462143

[24] Daniel Perlman and Purushottam Joshi. 1987. The revelation of loneliness. *Journal of Social Behavior & Personality* (1987), 63–76. https://psycnet.apa.org/record/1988-26551-001

[25] Matthew D. Pickard and Catherine A. Roster. 2020. Using computer automated systems to conduct personal interviews: Does the mere presence of a human face inhibit disclosure? *Computers in Human Behavior* 105 (2020), 106197. https://doi.org/10.1016/j.chb.2019.106197

[26] Matthew D Pickard, Catherine A Roster, and Yixing Chen. 2016. Revealing sensitive information in personal interviews: Is self-disclosure easier with humans or avatars and under what conditions? *Computers in Human Behavior* 65 (2016), 23–30. https://doi.org/10.1016/j.chb.2016.08.004

[27] Abhilasha Ravichander and Alan W Black. 2018. An Empirical Study of Self-Disclosure in Spoken Dialogue Systems. In *Proceedings of the 19th annual SIGdial meeting on discourse and dialogue*. 253–263.

[28] Sang-Wha Sien and Joanna McGrenere. 2025. A Gentle Introduction to Mental Health Through Storytelling: Design and Evaluation of Digital Human Library. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW005 (May 2025), 34 pages. https://doi.org/10.1145/3710903

[29] 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. https://doi.org/10.1145/3290605.3300768

[30] 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. https://doi.org/10.1037/0022-0167.50.3.351

[31] Yanyun (Mia) Wang, Weizi Liu, and Mike Yao. 0. Which recommendation system do you trust the most? Exploring the impact of perceived anthropomorphism on recommendation system trust, choice confidence, and information disclosure. *New Media & Society* 0, 0 (0), 14614448231223517. https://doi.org/10.1177/14614448231223517 arXiv:https://doi.org/10.1177/14614448231223517

[32] Joel Wester, Tim Schrills, Henning Pohl, and Niels van Berkel. 2024. "As an AI language model, I cannot": Investigating LLM Denials of User Requests. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–14.

[33] Marc Zao-Sanders. 2025. How People Are Really Using Gen AI in 2025. *OpenAI* (9th April 2025). https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025