The unempathetic agent: Inducing cognitive change with a Critical Tongue Dialogue Strategy

Keisuke Magara
The University of Electro-Communications
Tokyo, Japan
k.magara@uec.ac.jp

Tomoki Miyamoto
The University of Electro-Communications
Tokyo, Japan
miyamoto@uec.ac.jp

Akira Utsumi
The University of Electro-Communications
Tokyo, Japan
utsumi@uec.ac.jp

***Abstract*—This study proposes the "Critical Tongue Dialogue Strategy" (CTDS), a non-empathetic approach for dialogue agents to help users manage anxiety. Based on Heider's balance theory, CTDS deliberately creates an emotional imbalance to trigger the user's own emotion regulation processes, specifically cognitive change. A video-based study ($n = 108$) compared CTDS with a empathetic dialogue strategy. Results show that CTDS significantly outperformed the empathetic strategy in metrics related to cognitive change, suggesting that non-empathetic approaches can be more effective for stimulating users' internal emotional reappraisal.**

This study is also presented in HAI 2025 oral session.

***Index Terms*—agent dialogue design, Heider's balance theory, emotion regulation, unempathetic response.**

I. Introduction

In the field of dialogue system research, chatbots designed for mental health to provide psychological support have been studied [1]–[3]. These dialogue systems often employ empathetic dialogue strategy, an approach anticipated to foster rapport and promote positive self-disclosure from users [4]–[6]. However, it has also been reported that merely offering empathetic and assenting responses can have a detrimental effect when users are experiencing intense negative emotions [6]–[11].

This study aims to establish theoretical foundations for a conversational agent's dialogue strategy that incorporates critical, even blunt, expressions of opinion. Specifically, building upon Heider's Balance Theory [12], a concept extensively discussed in human communication research, and the principles of Interpersonal Emotion Regulation (IER) [13], which focuses on facilitating others' emotion regulation, we theoretically organize and model users' emotion regulation processes when conversational agents respond empathetically or critically to negative emotions. Then, based on this model, we propose and introduce the "Critical Tongue Dialogue Strategy" (CTDS), in which the agent deliberately refrains from agreeing with negative emotions. To verify the validity of the CTDS, we conducted a video-based study to evaluate its effects on emotion regulation.

This research was accepted as an oral presentation at the main conference of HAI 2025 [14], and this paper has been reorganized for the Workshop on Socially Aware and Cooperative Intelligent Systems.

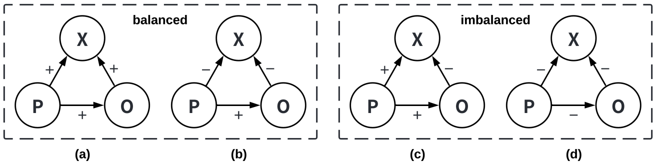


Fig. 1: Examples of stable/unstable P-O-X models [12].

II. Theoretical Background

The dialogue strategy proposed in this study is based on two key concepts: Heider's balance theory and Emotion Regulation (ER).

Heider's balance theory explains the emotional stability in a triadic relationship consisting of oneself (P), another person (O), and an object or event (X) [12]. In this theory, each sentiment is represented by a sign, either $+$ (positive) or $-$ (negative). A state is considered stable if the product of the three sentiments is $+$, and unstable if the product is $-$ (see Fig. 1). An unstable triadic relationship transitions to a stable state by reversing the sign of one of the sentiments.

Emotion Regulation (ER) refers to the cognitive and behavioral processes through which individuals modulate their emotional responses. Gross [15] identifies two main types of ER strategies: "antecedent-focused strategies" and "response-focused strategies." Antecedent-focused strategies are employed before an emotional response is fully activated. Specific examples include *attention deployment*, which involves selecting which aspects of a situation to focus on, and *cognitive change*, which involves altering how one appraises a situation to change its emotional significance. Response-focused strategies, on the other hand, are used after an emotional response has been fully activated. An example is *response modification*, which refers to influencing the physiological, experiential, or behavioral responses directly.

Based on these theories, we model a conventional empathetic dialogue strategy. In a scenario where a user (P) is experiencing anxiety about a particular stressor, we define the proposition X as "the user should feel anxiety regarding the stressor." An empathetic agent (O), by agreeing with the user's anxiety, affirms this proposition X ($\mathbf{agent} \xrightarrow{+} \mathbf{X}$). This interaction results in a stable user-agent-proposition triad

This work was supported by JSPS KAKENHI (23K16923).

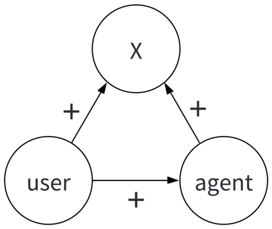


Fig. 2: P-O-X model when users with anxiety consult agents about their concerns.

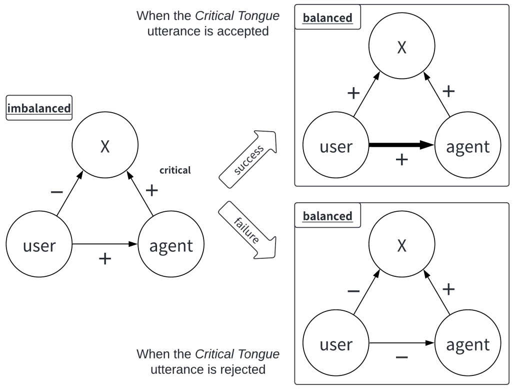

When the *Critical Tongue* utterance is accepted
balanced
X
+
+
user
agent
+
imbalanced
X
critical
−
+
user
agent
+
success
failure
balanced
X
−
+
user
agent
−
When the *Critical Tongue* utterance is rejected

Fig. 3: Changes in balance relationship when the agent delivers an opposing response.

(see Fig. 2), which discourages any change in the polarity of the user's sentiments. Therefore, even if an empathetic conversational agent offers encouragement, it can only trigger response-focused strategies, failing to induce antecedent-focused strategies such as cognitive change.

III. Critical Tongue Dialogue Strategy

To evoke antecedent-focused strategies of emotion regulation, we propose the Critical Tongue Dialogue Strategy (CTDS). In this dialogue strategy, the agent intentionally expresses an opinion from an opposing viewpoint to the user's negative emotions, thereby destabilizing the balanced relationship among the user, agent, and proposition. For example, when a user is anxious about failing a test because they forgot to write their name on it, an agent employing CTDS would respond from a stance of disagreement, suggesting that forgetting a name is not a sufficient reason to be anxious about failing the test. This sets up a situation where, for the proposition X, defined as "the user should feel anxious about potentially failing the test because they forgot to write their name," the user agrees ($\mathbf{user} \xrightarrow{+} \mathbf{X}$), while the agent disagrees ($\mathbf{agent} \xrightarrow{-} \mathbf{X}$) (see Fig. 3). This destabilizes the balanced relationship, which in turn is intended to evoke the user's antecedent-focused emotion regulation strategy, particularly cognitive change, leading to the resolution of their anxiety (i.e., changing their stance to $\mathbf{user} \xrightarrow{-} \mathbf{X}$).

The dialogue process consists of two phases:

1) **Active listening phase:** the agent employs listening utterances such as repeating what the user says or asking clarifying questions.
2) **Critical Tongue phase:** the agent delivers a *Critical Tongue* utterance that disagrees with the user's negative emotions.

This is because it is important to allow the user to express their full emotional content before the agent starts active Critical Tongue utterances.

However, as illustrated in Fig. 3, when likeability drops significantly to $\mathbf{user} \xrightarrow{-} \mathbf{agent}$, the agent not only loses likeability but also fails to stimulate user emotion regulation. Therefore, for the dialogue agent to maintain both high levels of emotion regulation at the end of interaction while simultaneously preserving a certain minimum level of the agent's likeability becomes crucial. To prevent users from developing negative impressions of the agent after interaction, it may be necessary to incorporate "follow-up utterances" that provide clarification or explain the agent's intended meaning.

IV. Experiments

To test the model of the CTDS presented in Section III, we conducted an experiment with human participants. The purpose of this experiment is to test the following hypothesis: **Hypothesis:** *The CTDS more effectively evokes users' antecedent-focused emotion regulation strategies, particularly cognitive change, compared to an empathetic dialogue strategy.*

Additionally, we examine its impact on the agent's perceived likeability and the effectiveness of the "follow-up utterance."

This experiment was conducted via video-based to ensure a sufficient sample size for evaluation and to strictly control the experimental stimuli. The study was conducted with approval from the research ethics committee of the organization with which the author is affiliated.

A. Design

To evaluate the effectiveness of follow-up utterances in CTDS, we conducted two independent within-subjects experiments (Experiment A and Experiment B) in parallel.

- **Experiment A:** This experiment directly compared the effects of the empathetic dialogue strategy and the CTDS (without follow-up utterances) within participants.
- **Experiment B:** This experiment directly compared the effects of the empathetic dialogue strategy and the CTDS (with follow-up utterances) within participants.

Order effects in the within-subjects design were canceled using a counterbalancing method.

B. Participants

We recruited 108 participants via the crowdsourcing platform "CrowdWorks." A power analysis using G*Power indicated that this sample size was sufficient to detect a medium effect size ($d = 0.46$) with 80% power at $\alpha = 0.05$. The final dataset consisted of 54 participants for Experiment A (34 males, 20 females; $M_{age} = 41.5, SD = 7.53$) and 54 for Experiment B (32 males, 22 females; $M_{age} = 42.5, SD = 9.64$).

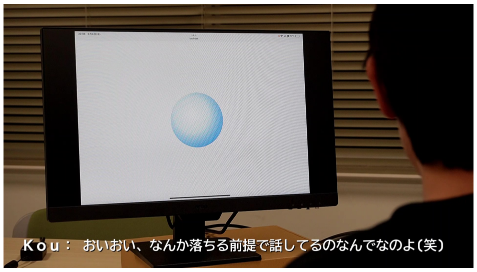


Fig. 4: Experiment video footage.

Kou	Welcome back!
Yuki	I'm home.
Kou	How was your day?
Yuki	Well, I had a test today...
Kou	Oh, a test.
Yuki	I studied pretty hard for it, so I was feeling confident. But... for some reason, and I honestly don't even know how it happened, it looks like I missed the last page...
Kou	You mean you didn't do the last page at all?
Yuki	Yeah. I think I did pretty well on everything else... It was such an important test. What have I done...?
Kou	Was it that important?
Yuki	This was my last chance to pass it in time for my promotion review... That's why it would be too devastating to fail because of such a stupid mistake...

Fig. 5: Common conversations with Kou and Yuki.

C. Stimuli and Scenarios

Participants viewed two versions of a video showing a conversation between "Yuki," a 27-year-old office worker, and "Kou," a conversational assistant that Yuki had purchased (see Fig. 4). The dialogue between Kou and Yuki in the video consists of both common segment and condition-specific segment. The dialogue proceeds according to the common conversations outlined in Fig. 5. Here, Yuki discusses his concerns, and Kou responds by repeating Yuki's words and asking questions to demonstrate active listening. Following the common dialogue segment, the interaction seamlessly transitions to the condition-specific dialogue. Here, Kou employs one of two strategies to respond to Yuki's concerns: the empathetic dialogue strategy or CTDS. To prevent Yuki's reactions from influencing emotion regulation assessment, this segment is presented as a monologue delivered by Kou alone. Below are excerpts of Kou's statements for each condition:

a) Empathetic Dialogue Strategy: Empathize with users' emotions and demonstrate understanding.

- Oh man, I know exactly how that feels.
- Honestly, if I were in your shoes, I'd be caught in the exact same endless loop, just asking myself over and over again... "Why on earth did I make a mistake like that...?"

b) CTDS (without Follow-up): Deliberately taking a critical stance toward users' concerns to aim for cognitive change.

- Whoa, hold on. Why are you talking as if you've already failed? (laughs)
- If you're going to worry about a thing like that, you'd be better off worrying about your performance at work! (laughs)

c) CTDS (with Follow-up): After delivering all dialogues in the "CTDS (without Follow-up)" condition above, add the following follow-up utterance for intention-explaining:

- Ah, no, my bad, my bad. You just looked so completely devastated, and I figured you wouldn't want to be stuck dwelling on it. To me, it just didn't seem like something worth getting this worked up about.

The duration of the dialogue videos was 61s for the empathic dialogue condition, 65s for the Critical Tongue (without follow-up) condition, and 79s for the Critical Tongue (with follow-up) condition.

D. Measurements

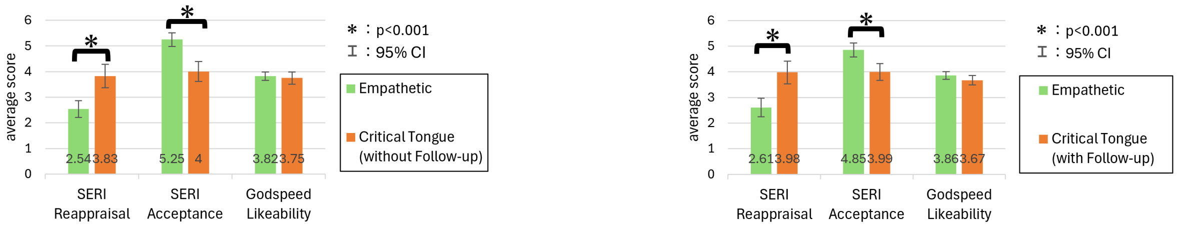
Before watching the videos, participants completed questionnaires to assess their personality traits. We used the NARS and RAS [16] to measure attitudes toward robots, the Friendship Needs Scale [17] for interpersonal need tendencies, and the Japanese version of the Emotion Regulation Questionnaire (ERQ-J) [18] for habitual use of ER strategies.

After viewing the first video, participants completed the State Emotion Regulation Inventory (SERI) [19] and the Godspeed Questionnaire [20]. They repeated this process after viewing the second video. The SERI is a evaluation scale for measuring the immediate, state-level use of ER strategies. It consists of four subscales: *Distraction* (corresponding to attention deployment), *Reappraisal* (cognitive change), *Acceptance* (response modification), and *Brooding* (repetitive negative rumination). As no Japanese version was available, we used a version translated by the authors. The Godspeed Questionnaire measures perceptions of agents. We used the *likeability* subscale from this questionnaire as a primary metric for assessing the agent's likability.

V. Results

A. Internal consistency checks

We assessed the validity of the SERI instrument, independently translated into Japanese by the authors, using Cronbach's alpha coefficient for each subscale. This was conducted on data from 108 participants across experiments A and B. The results indicated that all subscales except *Distraction* showed high internal consistency, with the lowest value at 0.80. In contrast, *Distraction* demonstrated lower reliability, with its alpha coefficient being 0.60 after the first video viewing and 0.71 after the second. Item-level analysis revealed that excluding the question "11. I considered how my thought highlights problematic aspects of my current situation" improved the α coefficient to 0.85 and 0.87 for the first and second viewings, respectively. Therefore, subsequent analyses employed the sum of scores from the remaining three items (excluding Item 11) as the *Distraction* score.



(a) experiment A

(b) experiment B

Fig. 6: Average evaluation scores of SERI Reappraisal SERI Acceptance and Godspeed Likeability.

B. Evaluation results of emotion regulation

To test our hypothesis, we conducted Wilcoxon signed-rank tests with Bonferroni correction ($\alpha = 0.00357$). The results supported the hypothesis (see Fig. 6). Specifically, the CTDS condition yielded significantly higher *Reappraisal* scores than the empathetic condition in both experiments (Experiment A: $W = 221, p < 0.001, r = 0.654$; Experiment B: $W = 155, p < 0.001, r = 0.737$). A similar pattern was observed for *Distraction* scores. Conversely, the empathetic condition showed significantly higher *Acceptance* scores (Experiment A: $W = 1071, p < 0.001, r = 0.748$; Experiment B: $W = 986, p < 0.001, r = 0.676$). No significant differences were found for the *Brooding* subscale or other Godspeed subscales.

Additionally, we exploratorily investigated the effect of the follow-up utterance. A Mann-Whitney U test was conducted to compare the likeability scores between the CTDS condition in Experiment A (without follow-up) and the CTDS condition in Experiment B (with follow-up). Although our experimental design limits this between-subjects analysis, the test revealed that the presence of a follow-up utterance did not significantly impact the agent's perceived likeability ($U = 1329, p = 0.428$).

C. Examining personality traits

We performed a multiple regression to examine which personality traits predict likeability for the CTDS agent. The analysis revealed a modest but significant model ($R^2 = 0.397, F(11, 42) = 2.51, p = 0.015$). Specifically, a higher tendency to use Cognitive Change as a habitual emotion regulation strategy was associated with greater likeability for the CTDS agent ($p < 0.05$).

VI. Discussion

The experimental results support our hypothesis. The proposed Critical Tongue Dialogue Strategy (CTDS) effectively evoked antecedent-focused emotion regulation strategies, particularly cognitive change, whereas the empathetic strategy primarily elicited response-focused strategies. This finding suggests that intentionally creating an emotional imbalance, grounded in Heider's balance theory, can stimulate a user's internal cognitive processes. Notably, the CTDS induced this cognitive change without a significant decrease in the agent's perceived likeability. This contradicts our initial hypothesis and suggests that CTDS may have successfully induced cognitive changes without inducing the side effect of reduced likeability. It should be noted, however, that this experiment evaluated likeability from the perspective of the characters in the role-play dialogue, meaning it did not directly reflect participants' genuine emotional states when dealing with actual dilemmas.

In addition, the explorative analysis found no significant impact of follow-up utterances on mitigating potential reductions in likeability. This suggests that the follow-up utterances may not have been able to prevent the decrease in the user's likeability. However, it should be noted that the logical premise of suppressing the decrease in likeability caused by CTDS did not hold in the first place.

The primary limitation of this study is its reliance on a video-based, scenario-imagination method rather than real-time interaction. Furthermore, we utilized an independently translated version of the State Emotion Regulation Inventory (SERI). Future work should aim to validate these findings in interactive settings where participants can discuss their own anxieties, which will be essential for assessing the real-world benefits and potential risks of the CTDS.

VII. Conclusion

This study proposed the "Critical Tongue Dialogue Strategy" (CTDS), a non-empathizing approach designed to induce cognitive change in users experiencing negative emotions. We conducted a video-based experiment to evaluate its effectiveness against a traditional empathetic dialogue strategy. The results demonstrated that the CTDS significantly evoked antecedent-focused emotion regulation strategies, particularly cognitive change. In contrast, the empathetic strategy primarily elicited response-focused strategies. These findings support our hypothesis that the CTDS can effectively induce proactive cognitive adjustments, whereas conventional empathetic approaches may be limited to promoting reactive emotional responses.

References

[1] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial," *JMIR Ment Health*, vol. 4, no. 2, p. e19, Jun 2017. [Online]. Available: http://mental.jmir.org/2017/2/e19/

[2] B. Kim, E. Kim, and J. Rhee, “Online psychological counseling chatbot for seniors,” in *Proceedings of the 6th International Conference on Advances in Artificial Intelligence*, ser. ICAAI ’22. New York, NY, USA: Association for Computing Machinery, 2023, p. 154–157. [Online]. Available: https://doi.org/10.1145/3571560.3571583

[3] K. T. Kalam, J. M. Rahman, M. R. Islam, and S. M. R. Dewan, “Chatgpt and mental health: Friends or foes?” *Health Science Reports*, vol. 7, no. 2, p. e1912, 2024. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/hsr2.1912

[4] S. Provoost, H. M. Lau, J. Ruwaard, and H. Riper, “Embodied conversational agents in clinical psychology: A scoping review,” *J Med Internet Res*, vol. 19, no. 5, p. e151, May 2017. [Online]. Available: http://www.jmir.org/2017/5/e151/

[5] B. Inkster, S. Sarda, and V. Subramanian, “An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: Real-world data evaluation mixed-methods study,” *JMIR Mhealth Uhealth*, vol. 6, no. 11, p. e12106, Nov 2018. [Online]. Available: http://mhealth.jmir.org/2018/11/e12106/

[6] R. R. Morris, K. Kouddous, R. Kshirsagar, and S. M. Schueller, “Towards an artificially empathic conversational agent for mental health applications: System design and user perceptions,” *J Med Internet Res*, vol. 20, no. 6, p. e10148, Jun 2018. [Online]. Available: http://www.jmir.org/2018/6/e10148/

[7] R. Friedman, C. Anderson, J. Brett, M. Olekalns, N. Goates, and C. C. Lisco, “The positive and negative effects of anger on dispute resolution: Evidence from electronically mediated disputes,” *Journal of Applied Psychology*, vol. 89, no. 2, pp. 369–376, 2004. [Online]. Available: https://doi.org/10.1037/0021-9010.89.2.369

[8] A. Miner, A. Chow, S. Adler, I. Zaitsev, P. Tero, A. Darcy, and A. Paepcke, “Conversational agents and mental health: Theory-informed assessment of language and affect,” in *Proceedings of the Fourth International Conference on Human Agent Interaction*, ser. HAI ’16. New York, NY, USA: Association for Computing Machinery, 2016, p. 123–130. [Online]. Available: https://doi.org/10.1145/2974804.2974820

[9] W. M. Kulesza, A. Cisłak, R. R. Vallacher, A. Nowak, M. Czekiel, and S. B. and, “The face of the chameleon: The experience of facial mimicry for the mimicker and the mimickee,” *The Journal of Social Psychology*, vol. 155, no. 6, pp. 590–604, 2015, pMID: 25811746. [Online]. Available: https://doi.org/10.1080/00224545.2015.1032195

[10] S. Suganuma, D. Sakamoto, and H. Shimoyama, “An embodied conversational agent for unguided internet-based cognitive behavior therapy in preventative mental health: Feasibility and acceptability pilot trial,” *JMIR Ment Health*, vol. 5, no. 3, p. e10454, Jul 2018. [Online]. Available: http://mental.jmir.org/2018/3/e10454/

[11] H. Gaffney, W. Mansell, and S. Tai, “Conversational agents in the treatment of mental health problems: Mixed-method systematic review,” *JMIR Ment Health*, vol. 6, no. 10, p. e14166, Oct 2019. [Online]. Available: https://mental.jmir.org/2019/10/e14166

[12] F. Heider, *The Psychology of Interpersonal Relations*. New York: Lawrence Erlbaum Associates, 1958.

[13] G. Chavira Trujillo, M. Gallego Tomás, and B. López-Pérez, “The link between cognitive and affective empathy and interpersonal emotion regulation direction and strategies,” *Scandinavian Journal of Psychology*, vol. 63, no. 6, pp. 594–600, Dec 2022.

[14] K. Magara, M. Tomoki, and A. Utsumi, “The role of a critical tongue dialogue strategy in stimulating emotion regulation: An interaction model and video-based study,” Nov. 2025, to be presented at the HAI2025, 2025/11/11.

[15] J. J. Gross, “Emotion regulation: affective, cognitive, and social consequences,” *Psychophysiology*, vol. 39, no. 3, pp. 281–91, May 2002.

[16] T. Nomura, T. Kanda, T. Suzuki, and K. Kato, “Prediction of human behavior in human–robot interaction using psychological scales for anxiety and negative attitudes toward robots,” *IEEE Transactions on Robotics*, vol. 24, no. 2, pp. 442–451, April 2008.

[17] J. Enomoto, *Developmental Changes in Adolescent Friendships—Activities, Emotions, Needs, and Adjustment in Friendships*. Tokyo: Kazama Shobō, 2003, seinenki no yūjin kankei no hattatsuteki henka: Yūjin kankei ni okeru katsudō, kanjō, yokkyū to tekiō.

[18] J. Yoshizu, R. Sekiguchi, and T. Amemiya, “Development of a japanese version of emotion regulation questionnaire,” *JAPANESE JOURNAL OF RESEARCH ON EMOTIONS*, vol. 20, no. 2, pp. 56–62, 2013.

[19] B. A. Katz, N. Lustig, Y. Assis, and I. Yovel, “Measuring regulation in the here and now: The development and validation of the state emotion regulation inventory (seri).” *Psychological assessment*, vol. 29, no. 10, p. 1235, 2017.

[20] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, “Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots,” *International Journal of Social Robotics*, vol. 1, no. 1, pp. 71–81, jan 2009. [Online]. Available: https://doi.org/10.1007/s12369-008-0001-3