Well-Being AI Encouragement in Sustained Interaction: Attitude-Aware Design Principles

Lingxuan Xiang[1], Hideaki Kikuchi[2]

[1]Waseda University, Graduate School of Human Sciences
[2] Waseda University, Faculty of Human Sciences
xianglingxuan@akane.waseda.jp, kikuchi@waseda.jp

Abstract

Encouraging utterances is widely used in supportive dialogue systems; however, how this should be designed for Well-Being AI remains underexplored. The Technology Acceptance Model (TAM) is sometimes interpreted as suggesting a simple, monotonic relationship in which more positive attitudes toward AI lead to higher perceived effectiveness of AI-mediated support. In our preliminary study with older Japanese adults examining AI-delivered encouragement, we measured participants' baseline attitudes toward AI and collected perceived effectiveness ratings of encouragement strategies across multiple strategy types and worry contexts. This study provides preliminary indications that perceived effectiveness may not be well explained well by a simple linear assumption regarding users' attitudes toward AI. Based on these observations, this paper proposes five design principles for encouragement in Well-Being AI, with a focus on adapting encouragement type selection based on users' attitudes toward AI. Our key ideas include reframing "user attitudes toward AI" from a pre- and post-interaction evaluation measure to a dynamic signal that should be continuously recognized and responded to during interaction. Encouragement strategy selection and realization should be guided by users' current attitudes and interaction feedback. In this process, users' evolving attitudes reshape the system's behavior, while the system's responsive support in turn influences users' subsequent attitudes. This bidirectional adaptation informs the design of emotionally supportive interactions and may contribute to the co-evolution of human and machine intelligence in Well-Being AI.

Introduction

As conversational agents and chatbots designed for well-being become increasingly prevalent, these systems can listen to users' worries, provide emotional support to reduce stress and anxiety, and promote well-being (Vaidyam et al. 2019). Emotional support encompasses a range of behaviors, from attentive listening and validation to active encouragement and advice-giving (Cutrona and Suhr 1992). Prior research has shown that attentive listening can reduce stress in supportive interactions (Weger et al. 2014). However, listening and encouragement may differ in how they alleviate distress. Research on social support suggests that effective supportive communication works primarily by facilitating the recipient's cognitive reappraisal of stressful situations, rather than through receptive responses alone (Burleson and Goldsmith 1998), and messages promoting cognitive reappraisal may produce greater reductions in emotional distress than receptive responses (Batenburg and Das 2014). Receptive listening alone may not enable the cognitive reframing, which is central to longer-term recovery (Pauw et al. 2018). Encouragement goes beyond receptive listening by actively reframing concerns and motivating coping behavior, which may enhance well-being in situations of worry and stress (Khan 2013; Xiang et al. 2025). This study focused on verbal encouragement. As a proactive form of emotional support, its effectiveness in AI-mediated settings is particularly sensitive to how users perceive and evaluate the AI delivering it. Therefore, encouragement is one of the core interaction behaviors that Well-Being AI systems should prioritize in their designs.

In AI-mediated encouragement, user responses are shaped not only by the content of supportive utterances and usability, but also by users' attitudes toward the AI and their broader evaluations of the system, such as perceived credibility and trust (Pan et al. 2017). Accordingly, users' thresholds for accepting the same supportive utterance and the ways in which they interpret it may vary as their attitudes fluctuate (Lyons, Hamdan, and Vo 2023). Studies on emotion recognition and empathic dialogue primarily focus on users' affective states during interaction, such as stress, depressive mood, and positive or negative affect (Wang and Beigi 2025; Kim et al. 2025). However, beyond users' momentary emotional states, AI-mediated support also involves how users feel about the AI itself. Few studies have

explicitly modeled such AI-directed affect, including negative attitudes toward AI, as a state that should be recognized and responded to in real time during interaction.

In this paper, we argue that AI-delivered encouragement can more effectively promote well-being when the selection of encouragement strategies is informed by users' current attitudes toward AI. Unlike conventional approaches that select strategies based primarily on users' emotional states or predefined rules, we propose that users' attitudes toward AI should be treated as a key input for strategy adaptation. Users' attitudes toward AI may shape how supportive utterances are interpreted. Therefore, attending to this dimension should be one of the core design considerations for AI systems that aim to support well-being, whether implemented as virtual agents or physically embodied robots.

Meanwhile, research on trust calibration and automation bias (Tatasciore and Loft 2025) indicates that affective support systems should manage dependency risks while promoting well-being. This includes mitigating avoidance or rejection driven by under-trust, as well as blind compliance and over-reliance driven by over-trust (Romeo and Conti 2025). A methodological gap also remains in the evaluation of sustained interactions. Widely used questionnaires, such as the Godspeed Questionnaire Series (Bartneck et al. 2009), Robotic Social Attributes Scale (RoSAS) (Pan et al. 2017), and Negative Attitudes toward Robots Scale (NARS) (Nomura et al. 2006), are useful for pre/post-assessment and cross-sectional comparisons. However, their reliability, sensitivity, and respondent burden under long-term repeated administration may remain insufficiently validated (Matheus et al. 2025).

Therefore, this paper advances a design perspective for Well-Being AI that treats users' attitudes toward AI as an important signal in emotionally supportive interaction. The Technology Acceptance Model (TAM) posits that user acceptance of a technology is primarily determined by its perceived usefulness and perceived ease of use, with attitude towards the technology serving as a mediating variable (Davis 1989). The TAM has been extended to explain the acceptance of social and embodied AI systems (e.g., Heerink et al. 2010), and is sometimes interpreted as implying that more positive attitudes toward AI may yield higher perceived effectiveness of supportive interactions. However, encouraging utterances should not rely on the assumption that more positive attitudes always yield better outcomes. Instead, users' attitudes toward AI should be treated as dynamic affective signals that can fluctuate during interaction and thus require continuous recognition and responsive adaptation. Building on this, we propose actionable design principles and evaluation recommendations for emotional support and dependency risk management in sustained interactions.

From a co-evolutionary perspective, this process is not unidirectional: as the system adapts its encouragement strategies to users' evolving attitudes, users in turn develop new expectations and reliance patterns that reshape subsequent interactions. This bidirectional adaptation constitutes a form of human–AI co-evolution in emotionally supportive contexts.

Preliminary Study

Building on the theoretical background above, we conducted an exploratory online survey with older Japanese adults to investigate the perceived effectiveness of different AI-delivered encouragement strategies across worry contexts, and how users' baseline attitudes toward AI may relate to these perceptions. This study analyzed the AI-delivered encouragement conditions of 31 older Japanese older adults (aged 65–74 years; 12 females, 19 males). The participants were first asked to describe an everyday worry. Based on a prior classification framework (Tanaka 2015), encouragement was categorized into five types: (1) "Reassurance and affirmation" (affirming with positive praise), (2) "Expressing concern" (showing sympathy by asking questions), (3) "Encouragement to take action" (motivating the person to act), (4) "Offering specific actions" (offering concrete help), and (5) "Distraction" (shifting focus away from problems). For each strategy type, participants were presented with its definition. After understanding each definition, they rated the perceived effectiveness of that encouragement type on a scale of 0 to 100. Each participant provided ratings for both psychological and physical worry contexts. Participants' baseline attitudes toward the AI were measured using the Negative Attitudes toward Robots Scale (NARS) (Nomura et al. 2006).

Overall, the perceived effectiveness varied across strategy types and worry contexts. Most linear associations between NARS subscale scores and perceived effectiveness were small. However, several strategy-specific and context-dependent patterns emerged. In physical worry contexts, "Offering specific actions" showed nonlinear (quadratic) associations with NARS-S1 and NARS-S2, exhibiting an inverted U-shaped pattern and a positive linear association with NARS-S3; other strategies did not show clear linear or nonlinear trends. Additionally, "Expressing concern" showed lower perceived effectiveness among participants with higher NARS-S2 scores. In psychological worry contexts, "Expressing concern" showed a positive association with NARS-S3, while other strategies did not show clear trends. These observations suggest that the attitude–effectiveness relationship is strategy-specific and context-dependent, rather than being well captured by a single linear assumption. Accordingly, attitudes toward AI may shape how users interpret supportive messages and where they

draw acceptability boundaries, which in turn may influence how encouragement strategies are perceived across different contexts. Detailed statistical analyses will be reported separately.

This was an exploratory and preliminary study. It featured a relatively small sample limited to older adults in Japan and a single-session online survey design. This initial investigation focused specifically on verbal encouragement, which is just one form of emotional support. This study did not consider broader types of support, such as informational or instrumental, nor did it include a listening-only control condition. Additionally, finer-grained subtypes of worries and detailed individual contextual factors were not distinguished. Moreover, attitudes toward AI were measured only as a pre-interaction baseline using a self-report scale. Therefore, the potential dynamic evolution of such attitudes across sustained, multi-turn interactions remains unexamined, as does the feasibility of recognizing them in real time with low-burden measures.

Notwithstanding these limitations, the pilot observations motivate a set of design principles that treat user attitudes as dynamic signals requiring continuous recognition and response during interactions.

Guiding Design Principles

Our preliminary study provided design-relevant indications that perceived effectiveness of AI-delivered encouragement may vary across strategy types and worry contexts, and that users' attitudes toward AI may relate to these perceptions in context-dependent ways. Motivated by these observations and previous literature, this paper proposes design principles for delivering encouragement in AI systems that support sustained interaction and promote user well-being.

Our central claim is that users' attitudes toward AI are not merely explanatory factors of acceptance or satisfaction. Rather, they can shape how users interpret supportive utterances and where they draw acceptability boundaries during interactions. Consequently, users' attitudes toward AI should inform both the selection of encouragement strategies and how these strategies are expressed. We propose the following five principles as actionable design hypotheses to guide subsequent system design and evaluation.

P1. Treat Attitudes as Dynamic State Signals During Interaction (AI-directed Affect)

Users' attitudes toward AI (such as AI-directed unease or negative evaluations) should not be treated as static, one-time assessments made before or after an interaction. Instead, they can be conceptualized and modeled as dynamic state signals that fluctuate during interaction. Therefore, systems can benefit from continuously tracking this attitudinal signal, both within and across sessions, and using it as a key input for both selecting encouragement strategies and shaping the linguistic style of their utterances.

P2. Offer Choice and Controllability to Safeguard Autonomy

Emotional support systems should provide users with explicit controls over both the style and intensity of encouragement. These controls should be readily accessible, allowing users to adjust or disable their encouragement at any time. This flexibility can reduce the perceived intrusiveness and the loss of agency that arises when the system overrides users' judgment. More importantly, it preserves users' fundamental ability to decline support and regulate their own engagement. Such designs are therefore important for aligning AI systems that aim to support well-being with the ethical principle of user autonomy.

P3. Promote Trust Calibration and Appropriate Reliance

To prevent over-reliance, the system should support trust calibration. This involves routinely communicating its capability boundaries and the uncertainty of its recommendations, while avoiding design cues that inflate perceived capability or certainty. The overarching goal should shift from maximizing trust to ensuring that user reliance remains appropriate to the system's actual capabilities. Accordingly, when signs of high dependence or reduced vigilance are detected, the system should provide brief prompts to help users preserve their capacity to question and make choices.

P4. Dynamically Adapt Support Using Attitudinal and Behavioral Cues

When users exhibit strong negative attitudes, the system should adopt a more deferential and less directive approach. This entails avoiding overly intimate framing and refraining from decision substitution, minimizing highly directive language, and explicitly emphasizing user autonomy through phrases such as "it is up to you." A key implication is that support strategies should not be assumed to work uniformly in AI-mediated interaction; instead, their effectiveness should be validated and adapted to the user's context and feedback. Building on this, the system should implement a continuous adaptation loop that updates strategy selection and message formulation during interaction by integrating inferred attitudinal states with immediate behavioral feedback (e.g., acceptance, rejection, and continued engagement).

P5. Maintain Support Diversity and Acknowledge the Heterogeneity of Strategies and Contexts

Given that encouragement effectiveness is strategy-specific and context-dependent, AI systems that aim to promote user

well-being should avoid pursuing a single, universal strategy. Instead, they should maintain a diverse repertoire of support options spanning multiple strategies and expressive styles to accommodate the heterogeneity in users' attitudes, states, and interaction contexts. Accordingly, strategy selection should be adaptive, balancing supportive experience (e.g., engagement and comfort) with risk management (e.g., preventing over-reliance and preserving autonomy). This repertoire should also include the option of attentive listening without active encouragement, as some users' stress may be alleviated through receptive support alone, and recognizing when not to encourage is itself part of an attitude-aware design.

Integrated Mechanism

The five principles above are not independent guidelines but are components of a closed-loop interaction mechanism. In practice, they operate as follows:

First, the system continuously estimates the user's current attitudinal state through low-burden measures and behavioral cues (**P1**). Based on this estimation, the system selects an appropriate encouragement type and expressive style from a diverse strategy repertoire (**P5**), adapting its approach to the user's inferred attitude (**P4**). For example, when strong negative attitudes are detected, the system avoids highly directive strategies and shifts toward more autonomy-respecting expressions. During delivery, the system provides the user with options to adjust or decline encouragement (**P2**). Finally, the user's responses (e.g., acceptance, rejection, and continued engagement) are fed back into the sensing stage, updating the attitude estimation and initiating the next cycle (**P1, P4**). Throughout this loop, the system periodically communicates its capability boundaries and inserts calibration prompts when signs of over-reliance are detected (**P3**), ensuring that trust calibration operates as a continuous safeguard across all stages. This continuous loop ensures that all five principles work together as an integrated adaptation mechanism rather than as a static checklist. The mechanism is shown in Figure 1.

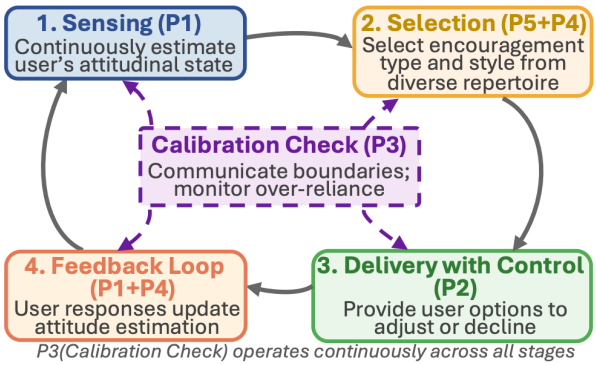

Figure 1: Integrated Mechanism: Attitude-Aware Adaptation Loop

The principles proposed in this paper are formulated at a modality-independent level and are intended for application to both virtual agents and physically embodied robots. However, embodiment may influence users' attitudes toward AI in important ways. Physical robots elicit a stronger social presence and may amplify both positive engagement and negative reactions, such as uncanny valley effects (Mori et al. 2012), which could make attitude dynamics more pronounced and continuous attitude sensing more critical. Simultaneously, embodied systems offer richer multimodal cues (e.g., gaze, gesture, and proximity) that can serve as additional behavioral indicators for real-time attitude estimation. Future studies should examine how embodiment modulates the attitude-aware adaptation proposed in this paper.

Conclusion

This study examined encouraging utterances as a representative emotional support scenario and advanced a design perspective in which users' attitudes toward AI are treated not only as acceptance measures assessed before or after interaction but also as dynamic affective signals that require continuous recognition and response during interactions.

A preliminary study with older Japanese adults provided design-relevant indications that perceived effectiveness may vary across strategy types and worry contexts and that the relationship between users' attitudes toward AI and perceived effectiveness may be context-dependent rather than linear. These observations motivate a design implication: encouragement design should consider the interplay among users' attitudes toward AI, strategy choice, and worry context.

Building on these observations and previous literature, this paper proposed five design principles for delivering encouragement in AI systems that support sustained interaction and promote user well-being. These principles emphasize treating attitudes as dynamic signals to inform strategy selection and linguistic realization within and across sessions; protecting user autonomy through choice and controllability; promoting trust calibration and appropriate reliance by communicating system boundaries and recommendation uncertainty; dynamically adapting strategies based on attitudinal and behavioral cues; and maintaining support diversity, including the option of attentive listening without active encouragement. An integrated mechanism illustrating how the five principles operate as a closed-loop adaptation cycle was also presented.

By treating users' attitudes as dynamic signals that both shape and are shaped by the system's supportive behavior, the proposed framework may provide a broader discussion of human–AI co-evolution: humans' evolving attitudes reshape the system's behavior, while the system's responsive adaptation influences humans' subsequent experience and

attitudes. This bidirectional process aligns with the goals of Well-Being AI to co-evolve human and machine intelligence.

Future work should develop low-burden, repeatable methods to assess attitudinal states during interaction over longer time spans and across broader populations, for example, by combining short scales with behavioral cues or multimodal signals. These approaches should be evaluated in real-world deployments and within existing support settings. Further research should assess the efficacy of strategy adaptation based on dynamic attitude sensing in improving well-being outcomes while reducing over-reliance and preserving users' ability to question and choose. Such evaluations should also compare the effects of active encouragement against attentive listening alone to clarify the incremental contribution of encouragement. Additionally, research should examine escalation or referral mechanisms to determine whether guiding users to more appropriate support when elevated negative attitudes or high dependency risks are detected can further improve outcomes. Ultimately, implementing and evaluating the proposed principles in a working system is essential to validate whether attitude-aware adaptation improves well-being in sustained interactions.

Acknowledgments

This work was supported by JST SPRING, Grant Number JPMJSP2128.

References

Bartneck, C.; Kulić, D.; Croft, E.; and Zoghbi, S. 2008. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics* 1(1): 71–81. doi.org/10.1007/s12369-008-0001-3.

Batenburg, A.; and Das, E. 2014. An Experimental Study on the Effectiveness of Disclosing Stressful Life Events and Support Messages: When Cognitive Reappraisal Support Decreases Emotional Distress, and Emotional Support Is Like Saying Nothing at All. *PLoS ONE* 9(12): e114169. doi.org/10.1371/journal.pone.0114169.

Burleson, B. R.; and Goldsmith, D. J. 1996. How the Comforting Process Works: Alleviating Emotional Distress through Conversationally Induced Reappraisals. In Handbook of Communication and Emotion, 245–280. San Diego: Academic Press.

Cutrona, C. E.; and Suhr, J. A. 1992. Controllability of Stressful Events and Satisfaction With Spouse Support Behaviors. *Communication Research* 19(2): 154–174. doi.org/10.1177/009365092019002002.

Davis, F. D. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly* 13(3): 319–339. doi.org/10.2307/249008.

Heerink, M.; Kröse, B.; Evers, V.; and Wielinga, B. 2010. Assessing acceptance of assistive social agent technology by older adults: The Almere model. *International Journal of Social Robotics* 2: 361–375. doi.org/10.1007/s12369-010-0068-5.

Khan, A. 2013. Predictors of Positive Psychological Strengths and Subjective Well-Being among North Indian Adolescents: Role of Mentoring and Educational Encouragement. *Social Indicators Research* 114(3): 1285–1293. doi.org/10.1007/s11205-012-0202-x.

Kim, J.; Mok, C.; Lee, J.; Kim, H. S.; and Jo, Y. 2025. Dialogue Systems for Emotional Support via Value Reinforcement. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 28733–28766, Vienna, Austria. Association for Computational Linguistics.

Lyons, J. B.; Hamdan, I. A.; and Vo, T. Q. 2023. Explanations and trust: What happens to trust when a robot partner does something unexpected? *Computers in Human Behavior* 138: 1–11. doi.org/10.1016/j.chb.2022.107473.

Matheus, K.; Ramnauth, R.; Scassellati, B.; and Salomons, N. 2025. Long-Term Interactions with Social Robots: Trends, Insights, and Recommendations. *ACM Transactions on Human-Robot Interaction* 14(3): 1–42. doi.org/10.1145/3729539.

Mori, M.; MacDorman, K. F.; and Kageki, N. 2012. The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine* 19(2): 98–100. doi.org/10.1109/MRA.2012.2192811.

Nomura, T.; Suzuki, T.; Kanda, T.; and Kato, K. 2006. Altered Attitudes of People toward Robots: Investigation through the Negative Attitudes toward Robots Scale. In Proceedings of the AAAI-06 Workshop on Human Implications of Human-Robot Interaction, 29–35. Palo Alto, CA: AAAI Press.

Pan, M. K. X. J.; Croft, E. A.; and Niemeyer, G. 2017. Validation of the Robot Social Attributes Scale (RoSAS) for Human-Robot Interaction through a Human-to-Robot Handover Use Case. Paper presented at the IROS 2017 Workshop on Human-Robot Interaction in Collaborative Manufacturing Environments, September 24.

Pauw, L. S.; Sauter, D. A.; van Kleef, G. A.; and Fischer, A. H. 2018. Sense or Sensibility? Social Sharers' Evaluations of Socio-Affective vs. Cognitive Support in Response to Negative Emotions. *Cognition and Emotion* 32(6): 1247–1264. doi.org/10.1080/02699931.2017.1400949.

Romeo, G.; and Conti, D. 2025. Exploring Automation Bias in Human–AI Collaboration: A Review and Implications for Explainable AI. *AI & Society* 41: 259–278. doi.org/10.1007/s00146-025-02422-7.

Tanaka, M. 2015. Various Aspects of 'Encouraging Utterances' Found in Drama Scripts. *Keio University Japan Language and Culture Education Center Research Bulletin* 43: 19–35.

Tatasciore, M.; and Loft, S. 2025. Calibrating Reliance on Automated Advice: Transparency and Trust Calibration Feedback. *International Journal of Human–Computer Interaction* 41(23): 14723–14733. doi.org/10.1080/10447318.2025.2487861.

Vaidyam, A. N.; Wisniewski, H.; Halamka, J. D.; Kashavan, M. S.; and Torous, J. B. 2019. Chatbots and Conversational

Agents in Mental Health: A Review of the Psychiatric Landscape. *Canadian Journal of Psychiatry* 64(7): 456–464. doi.org/10.1177/0706743719828977.

Wang, Z.; and Beigi, H. 2025. Quality-Controlled Multimodal Emotion Recognition in Conversations with Identity-Based Transfer Learning and MAMBA Fusion. arXiv:2511.14969.

Weger, H., Jr.; Castle Bell, G.; Minei, E. M.; and Robinson, M. C. 2014. The Relative Effectiveness of Active Listening in Initial Interactions. *International Journal of Listening* 28(1): 13–31. doi.org/10.1080/10904018.2013.813234.

Xiang, L.; Kikuchi, H.; Yang, J.; and Kikuchi, H. 2025. Investigating the Impact of Encouraging Utterances by Conversational Robots on Subjective Well-Being: A 15-Day Sustained Interaction. In Social Robotics, Lecture Notes in Computer Science 15562: 115–127. Cham, Switzerland: Springer.