

WittyHead: An Empathetic Architecture for Human-Agent Collaboration

Yuri A. Tijerino

Kwansei Gakuin University, Intelligent Blockchain+ Innovation Research Center
Sanda, Hyogo, Japan, ontologist@kwansei.ac.jp, <https://linkedin.com/ontologist>

Online Presentation

Abstract—Multi-agent systems supporting vulnerable populations require authentic emotional expression beyond conversational capability. This paper presents WittyHead, an anthropomorphic empathetic agent with coordinated facial expressions, gaze, gestures, and voice. WittyHead addresses three MASST priorities: (1) context-aware behavioral guard rails through ontology-driven emotion validation, (2) mutual observability through explainable multi-modal reasoning, and (3) design-time risk mitigation through therapeutic alliance research integration. Based on evidence that empathetic responses require compassionate concern rather than emotional mirroring [1], WittyHead implements asymmetric response mappings coordinating FACS-validated facial expressions, therapeutic eye contact, and prosodic-aligned gestures for privacy-preserving community support through Digital MOAI networks.

Index Terms—multi-agent safety, empathetic agents, emotional expressivity, therapeutic alliance, behavioral guard rails

I. Introduction

Multi-agent systems (MAS) serving vulnerable populations face critical safety requirements: preventing inappropriate emotional responses, maintaining transparent reasoning, and mitigating design-time risks [2]. Current empathetic agents exhibit a fundamental flaw: simple emotional mirroring that can exacerbate user distress [1]. Therapeutic alliance research demonstrates that empathetic responses to distress require compassionate concern, not mirrored negative emotions—smiling at someone experiencing distress is perceived as “invalidating and aversive” [1].

WittyHead is an anthropomorphic multi-modal empathetic agent implementing evidence-based emotional expressivity through coordinated facial expressions (FACS), gaze (therapeutic alliance), gestures (prosodic-aligned), and voice modulation. The system serves as the empathetic interface for Digital MOAI [3], AI-enhanced mutual aid networks for vulnerable populations including foster care youth, elderly, and individuals experiencing mental health challenges.

II. MASST-Aligned Architecture

A. Context-Aware Behavioral Guard Rails

WittyHead implements three layers of behavioral safety:

Empathetic Response Mapping: Asymmetric emotion mappings prevent invalidating mirroring. Table I

shows evidence-based mappings where user distress (sad, angry, fear) triggers compassionate concern, not mirrored distress. This implements Gilbert et al.’s finding that compassionate expressions achieve significantly higher empathy ratings than mirrored negative emotions [1].

TABLE I: Empathetic Response Mapping

User Emotion	Avatar Response	Safety Rationale
Sad	Compassionate Concern	Prevent mirroring
Angry	Calm Concern	De-escalation
Fear	Reassuring	Protective stability
Happy	Happy	Reinforce positive

Ontology-Driven Validation: Medical ontologies block inappropriate responses (e.g., celebratory expressions for serious diagnoses).

Accessibility Guard Rails: WCAG 2.1 AAA compliance [4] ensures expressions accommodate visual, auditory, motor, and cognitive disabilities.

B. Mutual Observability

WittyHead provides explainable multi-modal reasoning enabling human oversight:

Transparency: System logs specify: “User emotion: Sad (0.8 intensity) → Avatar response: Compassionate Concern with 75% eye contact (therapeutic alliance), palm-up gestures (openness), FACS AU4+AU6+AU12 (concern+warmth).”

Multi-Modal Coordination: Six services ensure synchronized expressivity: Emotion Detection, Empathetic Response Orchestrator, Facial Expression Manager (FACS-validated ARKit blendshapes), Gaze Manager (therapeutic eye contact [5]), Gesticulation Manager (prosodic-aligned [6]), Voice Modulation. Central orchestrator ensures 60 FPS synchronization.

Reasoning Traces: Decisions traceable to evidence [1], [5], [7].

C. Design-Time Risk Mitigation

WittyHead integrates therapeutic alliance research preventing empathy failures before deployment:

FACS-Validated Expressions: Facial Action Coding System [8] provides scientific basis for expressions. Compassionate concern implements AU4 (brow lowerer) + AU6 (cheek raiser) + gentle AU12 (lip corner puller), validated by Gilbert et al. [1] as conveying understanding without sharing distress.

Therapeutic Eye Contact: Research demonstrates that high eye contact enhances therapeutic alliance and

empathy [5]. Approach-avoidance theory [9] informs gaze direction for different emotional contexts.

Prosodic Gesture Alignment: Gestures synchronize with prosodic peaks [6]. Palm orientation signals trust vs. dominance [7].

III. Digital MOAI Integration

WittyHead provides the empathetic interface for Digital MOAI, AI-enhanced Okinawan mutual aid networks (摸合, moai) serving vulnerable populations [3].

Privacy-Preserving: Local-first emotion processing on AIngle DLT platform (EU H2020 FASTER [10]) enables real-time expressivity (0.16ms latency).

User-Controlled: Three automation levels provide explicit consent for AI emotional responses. **Safety-Critical:** Targets foster care youth, elderly experiencing social isolation, mental health support [11].

IV. Implementation

Compassionate Concern Expression: ARKit blend-shapes implement validated FACS pattern: AU4 (brow lowerer), AU6 (cheek raiser), gentle AU12 (lip corner puller) [1].

Multi-Modal Coordination: Direct gaze enhances approach emotions [9]. Beat gestures align with pitch peaks [6]. Eye contact marks turn-yielding [12].

V. Discussion

A. MASST Initiative Contributions

WittyHead demonstrates three safety mechanisms for empathetic MAS:

Behavioral Guard Rails: Asymmetric emotion mappings prevent invalidating mirroring [1]. Ontology-driven validation blocks inappropriate responses. Accessibility guard rails ensure inclusive support.

Mutual Observability: Explainable reasoning enables oversight. Decisions traceable to research [1], [5].

Design-Time Mitigation: Integration of therapeutic alliance research [1], [5], FACS [8], gesture-prosody coordination [6] prevents empathy failures before deployment.

B. Implications for Empathetic MAS

Beyond emotional mirroring paradigm: Empathy requires different strategies for positive vs. negative emotions. Mirrored distress undermines trust with vulnerable populations [1]. WittyHead demonstrates multi-modal coordination can implement therapeutic alliance principles for safe, effective empathetic agents.

Digital Therapeutic Alliance (DTA) extension: While DTA research focuses on text-based chatbots [13], [14], WittyHead extends to multi-modal embodied agents. Coordinated nonverbal cues (facial, gaze, gesture) grounded in face-to-face therapeutic research enable anthropomorphic empathetic expressivity beyond conversational agents.

VI. Conclusion

WittyHead demonstrates that safe empathetic multi-agent systems require: (1) behavioral guard rails preventing

invalidating emotional mirroring, (2) explainable multi-modal reasoning enabling human oversight, and (3) design-time risk mitigation through therapeutic alliance research integration. Evidence that empathy requires compassionate concern rather than mirroring [1] informs asymmetric response mappings coordinating FACS-validated expressions, therapeutic eye contact, and prosodic gestures. Integration with Digital MOAI demonstrates privacy-preserving community support for vulnerable populations. Future work will validate through human subjects research (JSPS KAKENHI Grant JP23K01882) employing DTA measurement instruments [13] to assess anthropomorphic avatar alliance beyond text-based agents.

Acknowledgments

JSPS KAKENHI Grant JP23K01882 (PI: K. Kotoku). EU H2020 FASTER (Grant 833507).

References

- [1] P. Gilbert, C. McEwan, R. Matos, and A. Rivis, "Compassionate faces: Evidence for distinctive facial expressions associated with specific prosocial motivations," *PLOS ONE*, vol. 14, no. 1, p. e0210283, 2019.
- [2] J. M. Bradshaw and M. Mahmud, "First international MASST initiative workshop: Multi-agent system safety and teamwork," in *IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology*, 2025.
- [3] D. Buettner, *The Blue Zones: Lessons for Living Longer from the People Who've Lived the Longest*. National Geographic Society, 2008.
- [4] World Wide Web Consortium (W3C), "Web content accessibility guidelines (wcag) 2.1," 2018, w3C Recommendation, June 2018. [Online]. Available: <https://www.w3.org/TR/WCAG21/>
- [5] N. M. Dowell and J. S. Berman, "Therapist nonverbal behavior and perceptions of empathy, alliance, and treatment credibility," *Journal of Psychotherapy Integration*, vol. 23, no. 2, pp. 158–165, 2013.
- [6] D. P. Loehr, "Temporal, structural, and pragmatic synchrony between intonation and gesture," *Laboratory Phonology*, vol. 3, no. 1, pp. 71–89, 2012.
- [7] A. Pease and B. Pease, *The Definitive Book of Body Language*. Bantam, 2006.
- [8] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto: Consulting Psychologists Press, 1978.
- [9] R. B. Adams and R. E. Kleck, "Effects of direct and averted gaze on the perception of facially communicated emotion," *Emotion*, vol. 5, no. 1, pp. 3–11, 2005.
- [10] "First responder advanced technologies for safe and efficient emergency response," EU H2020 FASTER Project, grant Agreement No. 833507, 2019-2022. [Online]. Available: <https://www.faster-project.eu>
- [11] K. Kotoku and Y. A. Tijerino, "Artificial intelligence (ai) in nursing practice: Current status and challenges," *Regional Caring (Chiiki Caring)*, vol. 23, no. 4, pp. 39–45, 2021, in Japanese.
- [12] R. Vertegaal, R. Slagter, G. van der Veer, and A. Nijholt, "Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes," in *Proceedings of ACM CHI*, 2001, pp. 301–308.
- [13] C. Beatty, T. Malik, V. Meheli, and S. Sinha, "Evaluating the therapeutic alliance with a free-text CBT conversational agent (wysa): A mixed-methods study," *Frontiers in Digital Health*, vol. 4, p. 847991, 2022.
- [14] S. D'Alfonso, O. Santesteban-Echarri, S. Rice *et al.*, "The digital therapeutic alliance and human-computer interaction," *JMIR Mental Health*, vol. 7, no. 11, p. e21895, 2020.