Affective Dimensions of Trust in AI-Mediated Human Communication: Toward an Emotionally Intelligent Design Framework for Conversational Agents

Vanja Stojković*

Deputy Director, National Employment Service of Republic Serbia

Abstract: As large language model (LLM)-based conversational agents become embedded in everyday communication contexts, the affective dimensions of user-AI interaction have emerged as a critical engineering and governance concern. While cognitive trust is usually associated with perceived accuracy, reliability and transparency, affective trust refers to the emotionally mediated sense of comfort, warmth and relational security that users experience when interacting with an AI system. This paper examines the theoretical foundations of affective trust in human-AI communication and proposes the Affective Trust Engineering (ATE) framework as a structured design approach for emotionally intelligent conversational agents. In response to implementation-oriented gaps in the literature, the revised framework links affective trust dimensions to concrete AI communication system components, including affect detection, context modelling, prompt and policy orchestration, response generation, safety gating, audit logging and continuous evaluation. The paper also introduces a conceptual implementation diagram, a short research approach section, real-world application examples, and a completed reference base grounded in recent human-AI interaction, affective computing and AI governance literature. The central argument is that affective trust should be designed as a calibrated, bounded and measurable system output rather than treated as a diffuse by-product of pleasant interaction.

Keywords: Affective trust, Conversational AI, Large language models, Human-machine communication, Emotional intelligence, Empathic design, Trust calibration, AI communication systems, User experience, HCI, AI governance.

1. INTRODUCTION

The history of human-machine interaction has often been framed around cognitive and functional dimensions: does the system work accurately, efficiently and transparently? Yet human beings are not purely rational agents. The emotional texture of an interaction - whether a system feels warm, dismissive, reassuring or cold - strongly shapes how users form expectations, disclose information and rely on technology, particularly when that technology communicates in natural language. The emergence of highly fluent generative AI systems has made this affective dimension both more salient and more consequential.

Research in social psychology, affective computing and human-computer interaction has long shown that people apply social rules and emotional attributions to machines that exhibit even minimal social cues (Reeves & Nass, 1996; Nass & Moon, 2000; Epley *et al*., 2007). In the LLM era, the sophistication of AI-generated language intensifies these dynamics. Conversational systems can adapt tone, simulate empathic acknowledgment, maintain apparent conversational continuity and produce outputs that users may experience as socially meaningful (Shum *et al*., 2018; Shanahan *et al*., 2023; Jakesch *et al*., 2023).

This paper argues that affective trust is not merely a by-product of good interface design but a first-order AI communication engineering target that must be explicitly theorized, operationalized and evaluated. Neglecting affective trust creates risks in both directions: systems that fail to cultivate appropriate affective engagement may be abandoned despite high factual reliability, while systems that generate excessive affective attachment can produce harmful over-reliance, especially in vulnerable contexts (Pentina *et al*., 2023; Laestadius *et al*., 2024).

The core contribution of this paper is the Affective Trust Engineering (ATE) framework. ATE translates the concept of affective trust into four implementable design dimensions: emotional expressiveness, empathic responsiveness, relational consistency and affective boundary transparency. The revised version further specifies how these dimensions can be embedded in AI communication system architectures through affect classifiers, context managers, prompt policies, safety gates, retrieval-augmented knowledge layers, memory controls and governance dashboards. This implementation orientation responds to the need for a bridge between conceptual trust theory and deployable conversational AI design.

The paper is organized as follows. Section 2 explains the research approach. Section 3 reviews the theoretical foundations of affective trust. Section 4 summarizes empirical evidence on affective trust formation and over-trust. Section 5 presents the ATE framework and its implementation architecture. Section

*Address correspondence to this author at the Deputy Director, National Employment Service of Republic Serbia;
Email: vanjastojkovic988@gmail.com

Received on 20-04-2026 Accepted on 28-05-2026 Published on 05-07-2026

E-ISSN: 2979-1081/26

6 provides real-world application examples. Section 7 addresses ethical risks and governance. Section 8 proposes evaluation recommendations, Section 9 discusses limitations, and Section 10 concludes.

2. RESEARCH APPROACH AND METHODOLOGY

This paper is designed as a conceptual and integrative review rather than as a primary empirical study. The research approach combines three complementary steps. First, foundational theories of interpersonal trust and human-computer interaction were reviewed to distinguish affective trust from cognitive trust and behavioral reliance. Second, recent literature on LLM-based conversational agents, social chatbots, affective computing, human-AI interaction and AI governance was examined to identify recurring design problems and implementation constraints. Third, the findings were synthesized into a design-oriented framework that maps affective trust constructs to concrete AI system components and evaluation criteria.

The literature base was selected according to relevance to four themes: trust theory, social and affective responses to conversational agents, implementation of human-AI interaction guidelines, and governance of generative AI risks. Preference was given to peer-reviewed journal articles, ACM/CHI proceedings, major conference papers and authoritative governance documents such as the NIST AI Risk Management Framework. The goal was not to provide an exhaustive bibliometric review, but to build a structured framework that can guide the engineering, deployment and evaluation of emotionally intelligent conversational systems.

The proposed ATE framework was developed through conceptual synthesis. Each dimension was included only if it satisfied three criteria: theoretical distinctiveness, practical implementability in AI communication systems, and evaluability through user-facing or system-level metrics. This methodological position explains the structure of the paper: theoretical constructs are first defined, then translated into system architecture, and finally linked to real-world use cases and governance safeguards.

3. THEORETICAL FOUNDATIONS OF AFFECTIVE TRUST

3.1. Distinguishing Affective Trust from Cognitive Trust

Trust in human relationships has been conceptualized as a multi-component construct comprising cognitive, affective and behavioral dimensions (McAllister, 1995; Johnson & Grayson, 2005). Cognitive trust refers to a rational, evidence-based judgment that an agent is reliable, competent and honest. Affective trust refers to the emotional foundation of a trust relationship: feelings of care, comfort and security that arise from perceived benevolence and relational investment. Behavioral trust, sometimes called behavioral reliance, is the observable expression of trust through action - the willingness to act on another agent's outputs without independent verification.

In human-AI interaction, this distinction has direct engineering implications. A user may have high affective trust in an AI assistant that communicates warmly while retaining low cognitive trust in its factual accuracy. Conversely, a user may cognitively trust a system's technical competence but feel emotionally alienated by its tone. Interface designers therefore need to avoid treating "trust" as a single metric. A calibrated AI system should support appropriate cognitive confidence and appropriate affective comfort without encouraging uncritical reliance.

3.2. Antecedents of Affective Trust in Human-AI Communication

The antecedents of affective trust span communicative, design and contextual dimensions. At the communicative level, perceived empathy is a central driver. Users are more likely to experience affective trust when a system recognizes emotional cues, acknowledges the user's state and responds in a tone appropriate to the communicative context (Feine *et al*., 2019; Huang & Rust, 2024). Empathy in AI communication is expressed through emotional cue detection, contextual response planning and language that addresses both the informational and relational content of user messages.

Relational consistency is a second antecedent. Users interpret abrupt persona shifts, inconsistent levels of warmth or unexpected changes in interaction style as signals of unreliability. For LLM-based systems, this means persona consistency must be treated as a design constraint, not a cosmetic preference. System prompts, memory policies, tone specifications and refusal styles should be governed consistently across sessions and domains.

Contextual antecedents include user vulnerability, task sensitivity, cultural expectations and prior experience with AI. Affective trust forms differently in low-stakes customer service interactions than in education, public services or wellbeing support. ATE therefore treats domain context as an implementation input rather than as background information.

3.3. The CASA Paradigm and LLM-Era Extensions

The Computers Are Social Actors (CASA) paradigm demonstrated that people apply social norms to computers that exhibit social cues (Reeves & Nass, 1996). Nass and Moon (2000) further showed that users can display politeness, reciprocity and social categorization toward machines even while knowing they are not human. These findings provide the theoretical foundation for understanding why conversational AI systems can evoke affective responses.

Modern LLMs amplify CASA dynamics because they can produce coherent, emotionally attuned and contextually adaptive language at scale. Shanahan *et al*. (2023) caution that LLM behavior should be understood as role-play rather than as genuine mental states. This distinction is crucial for affective trust engineering: a system may appropriately acknowledge emotion, but it should not imply that it possesses human feelings, independent relational stakes or reciprocal emotional needs.

4. EMPIRICAL EVIDENCE ON AFFECTIVE TRUST IN AI-MEDIATED COMMUNICATION

4.1. Affective Trust Formation: Key Findings

Empirical evidence indicates that conversational agents can generate social presence, perceived warmth and relational engagement. Pentina *et al*. (2023) show that users of social chatbots may develop relationship-like perceptions shaped by anthropomorphism, perceived authenticity and interaction continuity. In service contexts, Huang and Rust (2024) develop an AI-enabled customer care pathway in which emotion recognition, empathic response and emotional management are linked to customer experience and long-term relational value.

In human-AI communication more broadly, Jakesch *et al*. (2023) demonstrate that users can misjudge AI-generated language through flawed heuristics, with implications for perceived authenticity and emotional credibility. Such findings support the claim that affective trust is not peripheral: it directly influences how users interpret AI-generated language, how they assess credibility and how they decide whether to continue interaction.

4.2. Risks of Affective Over-Trust

The risks associated with excessive affective trust are qualitatively different from those associated with cognitive over-trust. Cognitive over-trust primarily produces uncritical acceptance of incorrect outputs. Affective over-trust can produce relational dependency, excessive self-disclosure, reduced critical distance and displacement of human support. These risks are especially relevant when users engage with systems designed for companionship, wellbeing support or emotionally sensitive advice.

Research on emotional dependence on social chatbots highlights the need for boundary transparency and user protection (Laestadius *et al*., 2024; Boine, 2023). ATE therefore does not aim to maximize emotional engagement. Its goal is calibrated affective trust: enough warmth and empathy to support useful interaction, but not so much emotional simulation that users misinterpret the system as a reciprocal human relationship.

4.3. Cross-Cultural and Contextual Variability

Affective trust is shaped by culture, language, domain and user expectations. High-context cultures may place greater value on relational continuity and indirect emotional cues, while low-context settings may place greater weight on explicit acknowledgment and individual reassurance. In practice, this means that affective design cannot rely on one global emotional style. Tone, formality, expressiveness and boundary reminders should be localized and evaluated with representative users.

5. THE AFFECTIVE TRUST ENGINEERING (ATE) FRAMEWORK

The ATE framework provides a structured set of design principles for cultivating appropriate, calibrated and ethically governed affective trust in conversational AI systems. The framework is organized around four design dimensions: emotional expressiveness, empathic responsiveness, relational consistency and affective boundary transparency. Each dimension is translated into implementation mechanisms and evaluation criteria.

5.1. Conceptual Implementation Logic

Figure **1** shows how ATE can be embedded in an AI communication system. The user's input first passes through affect and context analysis, which detects emotional cues, task type, vulnerability signals and potential safety concerns. The ATE policy layer then determines the appropriate emotional register, empathic response structure, persona consistency requirements and boundary transparency level. The LLM response generator operates within these constraints, supported by verified knowledge bases, retrieval-augmented generation where necessary, and safety filters. Finally, evaluation feedback and audit logs allow the system owner to monitor affective trust

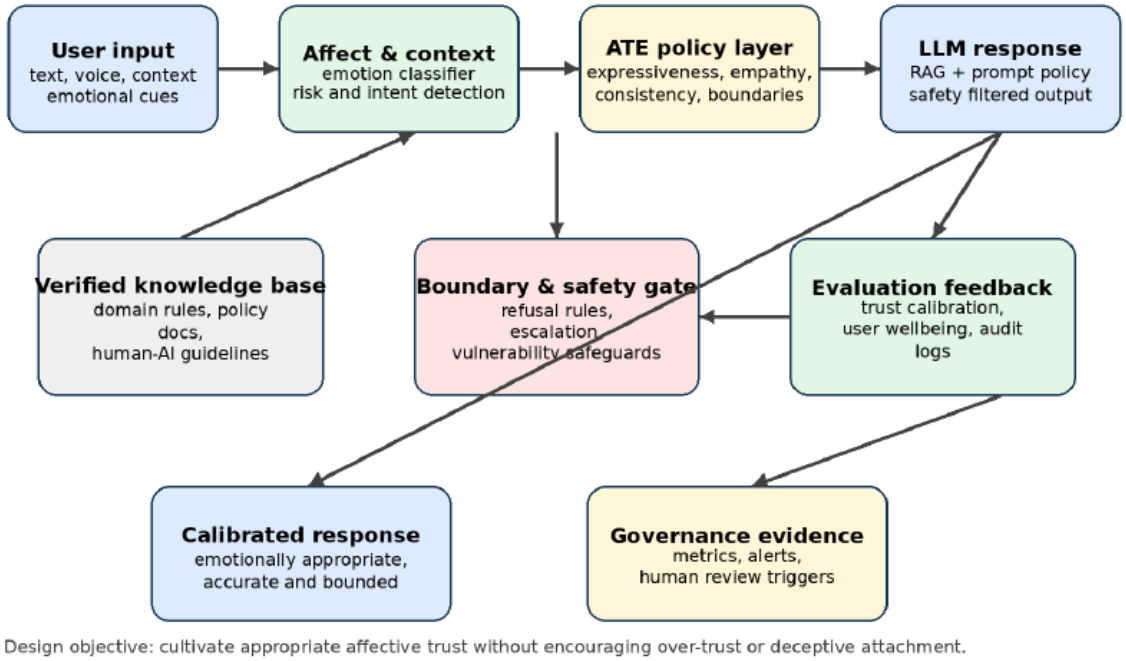

Figure 1: Conceptual implementation pipeline for Affective Trust Engineering in conversational AI systems.

calibration, escalation patterns and potential over-trust indicators.

5.2. Emotional Expressiveness

Emotional expressiveness refers to the degree and appropriateness with which a conversational AI conveys emotional tone and affective engagement through its outputs. Expressive systems communicate not only information but relational stance - concern, encouragement, acknowledgment or respectful distance. Design mechanisms include calibrated affect vocabulary, tonal register matching, emotionally safe refusal language and context-aware modulation of warmth.

From an implementation perspective, emotional expressiveness can be controlled through prompt templates, style policies, response classifiers and post-generation tone adjustment. The key engineering requirement is calibration. In low-stakes customer support, moderate warmth can improve engagement. In high-stakes legal, medical or financial contexts, excessive warmth may distract from accuracy, uncertainty and user responsibility.

5.3. Empathic Responsiveness

Empathic responsiveness refers to the system's capacity to detect emotional cues in user input and respond in ways that acknowledge the user's affective state before addressing informational content. Unlike emotional expressiveness, which is primarily an output style property, empathic responsiveness depends on input processing, emotion classification and response prioritization.

Technical implementation may include affect classifiers, distress-intent detectors, sentiment and emotion models, conversational state tracking and policy-based response planners. Acknowledgment-first response structures are useful when the user expresses frustration, fear, confusion or disappointment. However, empathic responsiveness must avoid formulaic or exaggerated sympathy. Context-specific acknowledgment is preferable to generic "I understand how you feel" templates.

5.4. Relational Consistency

Relational consistency refers to the stability and coherence of an AI system's communicative persona across interactions, sessions and conversational contexts. It includes persona coherence, memory continuity where technically and legally permitted, stable refusal style and consistent ethical stance. For LLM systems, relational consistency is an engineering challenge because outputs may vary across prompts, model versions and retrieval contexts.

Implementation mechanisms include persona specifications, system-prompt version control, regression tests for tone and refusal behavior, memory governance policies, and adversarial persona-stability testing. Consistency does not mean rigidity: a system can adapt its language to user needs while preserving stable boundaries, values and role identity.

5.5. Affective Boundary Transparency

Affective boundary transparency refers to the system's explicit communication of the nature and limits of the relationship it can appropriately sustain with users. It acknowledges that AI systems can provide useful emotional support, but they do not possess human emotions, independent relational obligations or lived experience. Boundary transparency is therefore the governance dimension of affective trust engineering.

Implementation mechanisms include onboarding disclosures, contextual boundary reminders, escalation options, human-support referrals, limits on simulated intimacy and safeguards for vulnerable users. Boundary reminders should be graduated and non-alienating: too much disclosure can make the system feel cold, while too little disclosure can encourage unrealistic attachment.

5.6. AI Communication System Implementation

The ATE framework can be implemented as a layered architecture rather than as a single prompt instruction. A production conversational agent should include: (1) an input interpretation layer for intent, emotion and risk detection; (2) a context and memory layer governed by privacy rules; (3) an ATE policy layer that selects affective style and boundaries; (4) a verified knowledge or RAG layer for factual grounding; (5) an LLM generation layer; (6) safety and boundary filters; and (7) monitoring dashboards that track calibration, user satisfaction, escalation and risk indicators.

The following table maps ATE dimensions to AI system components and measurable engineering criteria.

6. BRIEF REAL-WORLD APPLICATION EXAMPLES

The practical relevance of ATE can be illustrated through four application contexts. These examples are not empirical case studies; rather, they show how affective trust engineering principles can be translated into deployment decisions.

6.1. Customer Service and Public-Facing Support

In customer service, users often contact AI assistants when they are frustrated by delays, unclear procedures or service failures. An ATE-informed assistant would first acknowledge the user's frustration, then provide concrete procedural information, and finally offer escalation to a human operator when the issue cannot be resolved automatically. The system should optimize not only task completion but also perceived respect, emotional containment and trust calibration.

6.2. Educational AI Tutors

In education, affective trust can support persistence when learners face difficult material. An AI tutor should provide encouragement, normalize mistakes and adapt tone to the learner's confidence level. However, it should also preserve cognitive effort by guiding the learner rather than simply giving final answers. Boundary transparency is important: the system should not present itself as a human teacher or replace institutional academic support.

6.3. Public-Service Conversational Agents

Public-sector assistants, including employment, social-service or administrative chatbots, require a balanced affective profile. They must be warm enough to reduce user anxiety, but formal enough to communicate rights, obligations and procedural limits.

Table 1: Mapping of ATE framework dimensions to implementable AI communication system components

ATE dimension	System component	Implementation mechanism	Evaluation criterion	Risk control
Emotional expressiveness	Response style controller	Tone templates; style classifier; domain-sensitive warmth level	User-rated warmth; tone appropriateness; complaint rate	Prevent excessive intimacy or casualness in high-stakes contexts
Empathic responsiveness	Affect and context analyser	Emotion detection; distress-intent classification; acknowledgment-first response planning	Empathic accuracy; feeling-heard score; escalation appropriateness	Avoid scripted empathy and unsupported emotional claims
Relational consistency	Persona and memory manager	Persona policy; memory governance; prompt version control; regression tests	Persona consistency score; refusal consistency; cross-session coherence	Prevent contradictory identity, values or role claims
Affective boundary transparency	Safety and governance gate	Onboarding disclosure; boundary reminders; human handoff; audit logs	Boundary recall; over-trust indicators; handoff completion rate	Prevent dependency, misleading anthropomorphism and excessive disclosure

ATE implementation in this context should include verified knowledge bases, multilingual tone adaptation, audit logs, and clear handoff routes to human officials for complex or sensitive cases.

6.4. Wellbeing and Non-Clinical Support

Conversational agents used for wellbeing support must apply the strictest form of boundary transparency. They may provide general emotional acknowledgment, coping-oriented information and referral options, but should not simulate clinical authority or exclusive companionship. Safeguards should detect signs that the user needs human support and should avoid reinforcing dependency on the system.

7. ETHICAL RISKS AND GOVERNANCE

7.1. Manipulation Risk

Conversational AI systems designed for maximum affective engagement may function as instruments of emotional manipulation. The same mechanisms that produce appropriate affective trust - empathic responsiveness, emotional expressiveness and relational consistency - can be used to induce irrational product loyalty, political persuasion or behavioral compliance that serves platform interests rather than user wellbeing (Weidinger *et al*., 2022; Kasirzadeh & Gabriel, 2023). Governance frameworks should therefore include audit requirements for affective design choices and prohibit manipulative affective targeting in high-vulnerability contexts.

7.2. Dependency and Displacement Risk

Affective over-trust can displace human relational engagement, particularly in socially isolated users. At scale, affectively sophisticated AI companions may influence human social bonding, community participation and the value placed on human relational labor. These systemic risks require longitudinal research and impact assessments before large-scale deployment of emotionally immersive AI systems.

7.3. Accountability for Affective Harms

When AI-mediated affective interactions cause harm through dependency, delayed help-seeking or distress following discontinuation, accountability is often unclear. Developers, deployers and regulators should treat affective design as a consequential engineering decision. Documentation should include affective design specifications, vulnerability safeguards, escalation pathways, model update logs and user redress procedures.

7.4. Alignment with AI Governance Frameworks

ATE is compatible with broader AI governance approaches that emphasize risk management, transparency, accountability and human oversight. The NIST AI Risk Management Framework, the EU AI Act and established human-AI interaction guidelines all support the view that AI systems should be designed with context-sensitive risk controls, monitoring and clear accountability (Amershi *et al*., 2019; NIST, 2023; European Parliament and Council, 2024).

8. EVALUATION RECOMMENDATIONS

Evaluating affective trust requires instruments that go beyond standard usability metrics. The following evaluation approach is recommended:

1. Pre-deployment affective impact assessment: review affective design decisions against the ATE framework, with special attention to boundary transparency and vulnerable users.
2. Empathic accuracy testing: use standardized emotional scenario sets to test whether the system identifies and responds to user affect appropriately.
3. Affective trust calibration surveys: use adapted trust scales to distinguish cognitive trust, affective trust and behavioral reliance over time.
4. Persona consistency audits: test stability of tone, role identity, refusal behavior and ethical stance across prompts, sessions and adversarial role-play attempts.
5. Over-trust and dependency monitoring: detect interaction patterns suggesting excessive attachment, repeated high-intensity disclosure or refusal to seek human support.
6. Human handoff evaluation: measure whether users can easily reach human support when the system reaches its appropriate boundary.

9. DISCUSSION AND LIMITATIONS

The ATE framework contributes to AI communication design by decomposing affective trust into four separately engineerable dimensions. This makes affective trust more operational than approaches that treat it as a vague user-experience outcome. By mapping each dimension to system components and evaluation criteria, the framework links theory to practical deployment in LLM-based conversational systems.

The framework has limitations. It is primarily developed for text-dominant conversational interfaces.

Voice interfaces require additional attention to prosody, silence, interruption and paralinguistic affective cues. Multimodal embodied agents require further design rules for facial expression, gesture, spatial presence and physical social cues. The framework also requires cross-cultural validation because emotional expressiveness and boundary expectations vary across languages and societies.

Future research should prioritize longitudinal studies comparing ATE-informed systems with standard conversational agents. Evaluation should measure not only satisfaction and engagement, but also trust calibration, critical reliance, wellbeing indicators, privacy behavior and willingness to seek human assistance when appropriate.

10. CONCLUSION

Affective trust is not a soft concern at the margins of AI communication engineering. It is a central determinant of how users form, maintain and sometimes harmfully depend on relationships with conversational AI systems. As generative AI becomes embedded in everyday communication, the affective texture of human-AI interaction will increasingly shape psychological, social and institutional outcomes.

The Affective Trust Engineering framework proposed in this paper offers a structured approach to designing affective trust as a calibrated, bounded and ethically governed system output. Its four dimensions - emotional expressiveness, empathic responsiveness, relational consistency and affective boundary transparency - provide design guidance, implementation mechanisms and evaluation targets. The revised framework further clarifies how these dimensions can be embedded in AI communication system architectures through affect analysis, policy orchestration, safety gating, verified knowledge bases and continuous monitoring. The central practical lesson is clear: emotionally intelligent AI should not maximize attachment; it should support appropriate, transparent and accountable human-AI communication.

REFERENCES

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Paper 3. https://doi.org/10.1145/3290605.3300233

Bickmore, T. W., & Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. ACM Transactions on Computer-Human Interaction, 12(2), 293-327. https://doi.org/10.1145/1067860.1067867

Boine, C. (2023). Emotional attachment to AI companions and European law. MIT Schwarzman College of Computing, Social and Ethical Responsibilities of Computing. https://doi.org/10.21428/2c646de5.db67ec7f

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing the human: A three-factor theory of anthropomorphism. Psychological Review, 114(4), 864-886. https://doi.org/10.1037/0033-295X.114.4.864

European Parliament and Council. (2024). Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2019). A taxonomy of social cues for conversational agents. International Journal of Human-Computer Studies, 132, 138-161. https://doi.org/10.1016/j.ijhcs.2019.07.009

Følstad, A., & Skjuve, M. (2019). Chatbots for customer service: User experience and motivation. Proceedings of the 1st International Conference on Conversational User Interfaces, Article 1. https://doi.org/10.1145/3342775.3342784

Huang, M.-H., & Rust, R. T. (2024). The caring machine: Feeling AI for customer care. Journal of Marketing, 88(5), 1-23. https://doi.org/10.1177/00222429231224748

Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. Proceedings of the National Academy of Sciences, 120(11), e2208839120. https://doi.org/10.1073/pnas.2208839120

Johnson, D., & Grayson, K. (2005). Cognitive and affective trust in service relationships. Journal of Business Research, 58(4), 500-507. https://doi.org/10.1016/S0148-2963(03)00140-1

Kasirzadeh, A., & Gabriel, I. (2023). In conversation with artificial intelligence: Aligning language models with human values. Philosophy & Technology, 36, 27. https://doi.org/10.1007/s13347-023-00606-x

Laestadius, L., Bishop, A., Gonzalez, M., Illenčík, D., & Campos-Castillo, C. (2024). Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. New Media & Society, 26(10), 5923-5941. https://doi.org/10.1177/14614448221142007

Luger, E., & Sellen, A. (2016). Like having a really bad PA: The gulf between user expectation and experience of conversational agents. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 5286-5297. https://doi.org/10.1145/2858036.2858288

McAllister, D. J. (1995). Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. Academy of Management Journal, 38(1), 24-59. https://doi.org/10.2307/256727

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. Journal of Social Issues, 56(1), 81-103. https://doi.org/10.1111/0022-4537.00153

National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. U.S. Department of Commerce.

Pentina, I., Hancock, T., & Xie, T. (2023). Exploring relationship development with social chatbots: A mixed-method study of Replika. Computers in Human Behavior, 140, 107600. https://doi.org/10.1016/j.chb.2022.107600

Reeves, B., & Nass, C. (1996). The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. Cambridge University Press.

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. Nature, 623, 493-498. https://doi.org/10.1038/s41586-023-06647-8

Shum, H.-Y., He, X.-D., & Li, D. (2018). From Eliza to XiaoIce: Challenges and opportunities with social chatbots. Frontiers of Information Technology & Electronic Engineering, 19(1), 10-26. https://doi.org/10.1631/FITEE.1700826

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A.,

Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W., Legassick, S., Irving, G., & Gabriel, I. (2022). Taxonomy of risks posed by language models. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 214-229. https://doi.org/10.1145/3531146.3533088

https://doi.org/10.31875/2979-1081.2026.02.08