Human–Machine Communication in the Age of Generative AI: Engineering Trust, Interpretability, and Interaction Efficiency in Intelligent Systems

Vanja Stojković*

Deputy Director, National Employment Service of Republic Serbia

Abstract: The proliferation of generative artificial intelligence (GenAI) systems has fundamentally transformed the landscape of human–machine communication (HMC). This paper examines three critical engineering dimensions of this transformation: trust calibration, interpretability mechanisms, and interaction efficiency. Drawing on recent advances in large language models (LLMs), conversational agents, and human–computer interaction (HCI) research, we analyze how these dimensions interrelate and affect overall system usability. We propose a conceptual framework—the TIE (Trust–Interpretability–Efficiency) model—that explicitly builds upon and extends established HCI usability models (ISO 9241-11; Nielsen, 1994) and AI trust taxonomies (Lee & See, 2004; Hoff & Bashir, 2015) to provide actionable design principles for GenAI communication systems. Novelty lies in the tripartite synthesis and the explicit modeling of interdependencies among cognitive trust, affective trust, behavioral reliance, multi-level interpretability (model-, system-, and interface-level), and operationally defined efficiency metrics. Our analysis indicates that trust and interpretability share a bidirectional dependency, while efficiency gains are contingent on users developing accurate mental models of AI capabilities. Ethical and governance dimensions—including accountability and misuse prevention—are integrated as co-equal design concerns. Applicability is primarily scoped to LLM-based conversational systems, with acknowledged limitations for multimodal and embodied AI. The paper concludes with a structured evaluation checklist and metric table to support practitioner utility.

Keywords: Generative AI, Human–machine communication, Trust calibration, Cognitive trust, Affective trust, Behavioral reliance, Interpretability, Explainability, Interaction efficiency, Large language models, Conversational agents, HCI usability, AI governance.

1. INTRODUCTION

The emergence of generative artificial intelligence—encompassing large language models (LLMs) such as GPT-4, Claude, and Gemini—has catalyzed a paradigm shift in human–machine interaction. Unlike earlier rule-based or retrieval-oriented systems, modern GenAI agents engage users through contextually rich, probabilistic dialogue, producing outputs that are frequently indistinguishable from human-authored content. This capability has dramatically expanded the surface area of human–machine communication (HMC) beyond technical specialists to include broad segments of the general population (Anthropic, 2023; OpenAI, 2023).

However, this expanded accessibility introduces a set of engineering challenges that have not been fully resolved. Users must form appropriate trust relationships with systems whose internal reasoning processes remain largely opaque. Simultaneously, AI developers face the challenge of designing interfaces that surface system limitations without sacrificing fluency and interaction efficiency. These challenges are not merely usability concerns—they carry direct implications for safety, accountability, and the long-term adoption of AI in communication-critical domains such as healthcare, education, legal advising, and engineering support.

This paper addresses these challenges through a focused review and synthesis of current research across three interdependent axes: (1) trust calibration in GenAI contexts, (2) interpretability and explainability mechanisms applicable to communication interfaces, and (3) interaction efficiency as a design and evaluation criterion. Together, these form the basis of the proposed TIE (Trust–Interpretability–Efficiency) framework.

Crucially, the TIE framework is not proposed in isolation. It is positioned explicitly against established HCI usability models—including the ISO 9241-11 usability definition (ISO, 2018) and Nielsen's usability heuristics (Nielsen, 1994)—and against existing AI trust taxonomies (Lee & See, 2004; Hoff & Bashir, 2015; Kasirzadeh & Gabriel, 2023). The framework's novelty lies in its tripartite synthesis that integrates all three dimensions into a unified, interdependent model specifically calibrated for GenAI communication interfaces, which prior frameworks have addressed only in pairs or in isolation. Additionally, governance and ethical dimensions—including accountability and misuse risks—are embedded as co-equal engineering concerns rather than peripheral afterthoughts.

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

Received on 17-04-2026 Accepted on 27-05-2026 Published on 05-07-2026

E-ISSN: 2979-1081/26

2. BACKGROUND AND RELATED WORK

2.1. Human–Machine Communication in the LLM Era

The study of human–machine communication has a long history rooted in HCI, cognitive science, and communication theory. Early frameworks, such as those proposed by Winograd and Flores (1986), conceptualized machines as limited but structured interlocutors. The introduction of neural language models and especially transformer-based architectures (Vaswani *et al*., 2017) disrupted these assumptions by enabling machines to participate in open-domain conversational exchange with high coherence and apparent understanding.

Research in the post-LLM era has increasingly focused on the social and psychological dimensions of HMC. Studies such as those by Shum *et al*. (2018) and Xu *et al*. (2023) demonstrate that users routinely apply social attributions—including intent, emotion, and credibility—to AI interlocutors. This anthropomorphization shapes how users evaluate output reliability and decide whether to act on AI-generated information. More recent work on generative AI interfaces (Shanahan *et al*., 2023; Jakesch *et al*., 2023; Woebbecke *et al*., 2024) reveals that users of modern LLMs exhibit qualitatively different trust and calibration dynamics compared to earlier chatbot systems, partly due to the fluency and apparent sophistication of LLM-generated text, which can mask errors and hallucinations.

2.2. Positioning the TIE Framework Against Existing Models

Several established frameworks address subsets of the dimensions captured by TIE. In HCI, the ISO 9241-11 (2018) standard defines usability in terms of effectiveness, efficiency, and satisfaction—but does not address trust calibration or interpretability as distinct engineering targets. Nielsen's (1994) heuristics include "visibility of system status" and "help and documentation," which touch on interpretability, but were developed for deterministic interfaces and do not account for probabilistic AI behavior.

Within AI trust research, Lee and See's (2004) taxonomy distinguishes trust in terms of performance, process, and purpose—a categorization that maps partially onto TIE's trust and interpretability axes but does not incorporate interaction efficiency or GenAI-specific concerns. More recent frameworks from Liao and Vaughan (2023) and Kasirzadeh and Gabriel (2023) introduce nuanced trust calibration models for AI systems, while Adadi and Berrada (2018) and Arrieta *et al*. (2020) provide taxonomies of XAI methods. However, none of these frameworks jointly model trust calibration, multi-level interpretability, and measurable interaction efficiency within a unified design-oriented structure for conversational GenAI. The TIE framework fills this gap.

2.3. Trust in AI Systems

Trust in automated and AI systems has been studied extensively (Lee & See, 2004; Hoff & Bashir, 2015). Within GenAI contexts, trust calibration has emerged as a central concern. Miscalibrated trust—either overtrust or undertrust—leads to suboptimal outcomes. Overtrust produces uncritical acceptance of erroneous or hallucinated outputs, while undertrust results in underutilization of genuinely useful AI assistance (Vossing *et al*., 2022; Seshia *et al*., 2023).

Critically, the literature distinguishes at least three components of trust that must be addressed separately in system design:

- Cognitive trust refers to the rational, evidence-based assessment of AI reliability—users' belief that the system performs accurately and consistently based on observed track record and communicated uncertainty.
- Affective trust refers to the emotional dimension of the user–AI relationship—feelings of comfort, warmth, or confidence that arise from perceived care, responsiveness, and rapport, even independently of factual accuracy.
- Behavioral reliance refers to the observable, consequential dimension of trust—the degree to which users act on AI outputs without independent verification, as expressed in task delegation, reduced oversight, and follow-through on AI recommendations.

These three components do not always co-vary. Users may exhibit high cognitive trust (believing the system is generally accurate) while maintaining low behavioral reliance (habitually double-checking outputs) due to high stakes or low affective trust. Conversely, high affective engagement with an AI agent can produce behavioral over-reliance in the absence of cognitive calibration (Woebbecke *et al*., 2024; Jakesch *et al*., 2023). Design interventions must therefore targ*et al*l three components, not only accuracy-based transparency.

Recent empirical studies (Vossing *et al*., 2022; Seshia *et al*., 2023; Liao & Vaughan, 2023) suggest that trust is influenced not only by accuracy but also by perceived transparency, consistency of behavior, and

alignment with user expectations. Post-2022 research on generative AI interfaces further shows that trust dynamics are significantly modulated by the fluency and style of LLM responses—users frequently over-attribute confidence to outputs that are grammatically polished but factually unreliable.

2.4. Interpretability and Explainability

Interpretability—the degree to which a system's behavior can be understood by humans—is a well-established criterion in machine learning and AI safety research (Lipton, 2018; Doshi-Velez & Kim, 2017). In communication interfaces, interpretability takes on additional dimensions: users need not only to understand what the system outputs, but also why, how confidently, and under what constraints. Explainable AI (XAI) approaches—including attention visualization, natural language justifications, and uncertainty quantification—are increasingly being integrated into conversational interfaces (Arrieta *et al*., 2020; Adadi & Berrada, 2018).

For the purposes of the TIE framework, interpretability mechanisms are categorized into three operational levels:

- Model-level interpretability: mechanisms that expose properties of the AI model itself, such as calibrated confidence scores, uncertainty estimates, or capability-scope documentation. These primarily inform cognitive trust.
- System-level interpretability: design artifacts that explain the AI system's purpose, training data provenance, known limitations, and intended use boundaries—typically communicated before or at the start of interaction (*e.g*., onboarding disclosures, system cards). These set prior expectations and support appropriate behavioral reliance.
- Interface-level interpretability: real-time affordances that allow users to interrogate AI outputs during interaction, such as 'Ask me why,' 'Show sources,' or confidence rating displays. These support both cognitive trust and moment-to-moment calibration.

Research by Miller (2019), Liao and Vaughan (2023), and Cai *et al*. (2023) underscores that explanations are effective only when they are legible, timely, and contextually relevant. Generic or overly technical explanations can reduce user comprehension and paradoxically decrease trust by signaling system complexity without aiding understanding.

2.5. Interaction Efficiency

Interaction efficiency in the context of AI communication systems encompasses both task completion performance and cognitive load. Efficient interactions are those in which users achieve their communicative goals with minimal friction, reformulation, and error correction. Studies in conversational AI design (Luger & Sellen, 2016; Jiang *et al*., 2023) highlight that efficiency degrades significantly when users hold inaccurate mental models of system capabilities.

Operational measurement of interaction efficiency requires concrete metrics. Building on Jiang *et al*. (2023) and Amershi *et al*. (2019), the TIE framework operationalizes efficiency through the following indicators:

- Task Completion Time (TCT): elapsed time from task initiation to satisfactory completion, benchmarked against non-AI and prior-session baselines.
- Turn Count per Task (TCpT): the number of dialogue turns required to reach a satisfactory output, with higher counts indicating misalignment between user intent and AI interpretation.
- User Correction Rate (UCR): the proportion of AI-generated outputs that require substantive user revision before use, reflecting output reliability and prompt-response alignment.
- Cognitive Load (CL): assessed through instruments such as the NASA Task Load Index (NASA-TLX; Hart & Staveland, 1988) or validated single-item workload scales, capturing subjective mental effort during AI-mediated tasks.

These metrics enable systematic comparison across interface designs, user populations, and task domains, and form the basis of the evaluation recommendations in Section 5.

3. THE TIE FRAMEWORK

Based on the reviewed literature, we propose the TIE (Trust–Interpretability–Efficiency) framework as a synthesis of the three core dimensions in GenAI-mediated HMC. The framework posits that each dimension influences the others, and that optimal communication outcomes depend on engineering all three in concert. While prior frameworks address pairs of these dimensions, TIE is distinctive in its unified, design-oriented treatment of all three as mutually

constraining and mutually enabling engineering targets within LLM-based communication systems.

3.1. Trust Calibration

Trust calibration refers to the process through which users develop appropriately scaled confidence in AI outputs, encompassing cognitive, affective, and behavioral dimensions as defined in Section 2.3. The TIE framework identifies engineering levers for each trust component:

Cognitive Trust Engineering

Cognitive trust is primarily supported by explicit uncertainty signaling—using linguistic hedges ("I am not certain, but…"), calibrated confidence scores, and conditional disclaimers—and by behavioral consistency, ensuring that responses to semantically equivalent prompts remain stable and predictable. Recent work on generative AI uncertainty communication (Steyvers *et al*., 2023; Band *et al*., 2024) indicates that users are more capable of appropriate calibration when uncertainty is expressed in natural language rather than numerical probabilities, particularly for lay users.

Affective Trust Engineering

Affective trust is influenced by design choices that shape the perceived relational quality of AI interaction: response personalization, acknowledgment of user goals, consistent persona, and appropriate expressions of epistemic humility. Interface designs that employ overly clinical or robotic language may suppress affective trust even when cognitive trust is well-supported. Conversely, designs that prioritize warmth at the expense of accuracy disclosure can induce affective over-reliance (Jakesch *et al*., 2023).

Behavioral Reliance Engineering

Behavioral reliance—the extent to which users act on AI outputs without independent verification—is most directly shaped by graceful failure handling: designing systems to clearly acknowledge the boundaries of their competence rather than producing plausible but incorrect outputs. Mechanisms such as explicit refusal for out-of-scope queries, graduated confidence labeling, and "verify before acting" prompts in high-stakes contexts directly reduce over-reliance. These mechanisms must be empirically validated with target user populations, given the significant heterogeneity in how different groups—experts, novices, high-stakes decision-makers—exhibit behavioral reliance patterns.

3.2. Interpretability Design

Interpretability in communication-oriented AI systems must be engineered across all three levels identified in Section 2.4.

At the model level, LLM-based systems should expose calibrated uncertainty scores, known error-prone domains, and version or capability documentation. Where direct model internals are inaccessible to interface designers, system-level documentation—analogous to model cards (Mitchell *et al*., 2019) or AI system cards—serves as a proxy for model-level transparency.

At the system level, onboarding experiences, contextual help text, and proactive disclosures communicate capability scope and known limitations before users develop miscalibrated expectations. This pre-interaction transparency is particularly important given evidence that initial trust dispositions are highly persistent and resistant to later correction (Hoff & Bashir, 2015).

At the interface level, real-time affordances for interrogating AI reasoning—'Ask me why,' 'Show sources,' confidence rating displays—must be designed as first-class interface elements rather than post-hoc additions. Research consistently shows that these affordances are only effective when they are visually salient, contextually triggered, and require minimal additional effort to engage (Cai *et al*., 2023; Liao & Vaughan, 2023).

3.3. Interaction Efficiency

Efficiency is modulated by the accuracy of users' mental models of AI capabilities. When users understand the probabilistic nature of LLM outputs, the importance of prompt formulation, and the role of context, they engage more productively—exhibiting lower Turn Count per Task and lower User Correction Rates. The TIE framework recommends progressive disclosure of system behavior as a primary design strategy: systems should educate users through interaction rather than requiring prior technical literacy, surfacing relevant behavioral norms at contextually appropriate moments.

Efficiency optimization cannot be pursued in isolation. Interface designs that prioritize rapid response and minimal friction at the expense of interpretability may produce short-term TCT improvements but erode long-term trust calibration and appropriate use patterns. Evaluation frameworks for GenAI communication systems should therefore incorporate longitudinal measures of all four efficiency metrics alongside conventional usability outcomes.

3.4. Ethical and Governance Dimensions

Ethical and governance considerations are not peripheral to the TIE framework—they are co-equal engineering concerns embedded across all three

dimensions. The increasing deployment of LLM-based systems in consequential domains (healthcare, legal, education) raises governance challenges that require explicit design attention.

Accountability requires that AI communication systems maintain clear chains of responsibility for output quality. This includes logging mechanisms that allow post-hoc auditing of AI outputs, user-facing explanations of system provenance, and defined redress pathways when AI-mediated communication causes harm (Kasirzadeh & Gabriel, 2023; Wachter *et al*., 2017).

Misuse prevention requires that systems be designed with adversarial use cases in mind. LLM-based conversational agents are susceptible to prompt injection, social engineering, and outputs that are technically within policy but harmful in context. Design teams should conduct structured adversarial red-teaming and embed mitigations at both the model and interface levels (Perez & Ribeiro, 2022; Weidinger *et al*., 2022).

The TIE framework also acknowledges significant scope limitations. Its primary applicability is to LLM-based conversational AI systems in text-dominant interfaces. Application to multimodal AI systems (those incorporating vision, speech, or gesture modalities) would require extension of interpretability mechanisms to non-linguistic channels and revised efficiency metrics that account for multimodal communication overhead. Embodied AI systems—robots, physical agents—introduce further dimensions of trust and interpretability (*e.g*., behavioral legibility, physical safety) that are beyond the current framework's scope.

3.5. Interdependencies Within the TIE Framework

The TIE framework emphasizes that the three dimensions are not independent. Higher interpretability tends to support better trust calibration—specifically cognitive trust—by giving users the information needed to evaluate outputs critically. In turn, well-calibrated trust (especially behavioral reliance) leads to more efficient interactions because users neither over-verify every output nor blindly accept incorrect responses. Conversely, interaction efficiency pressures—such as high task urgency or elevated cognitive load—can undermine interpretability engagement by discouraging users from activating explanatory affordances. Ethical constraints also modulate these trade-offs: accountability requirements may impose efficiency costs (*e.g*., mandatory review steps) that must be accepted as a governance necessity rather than an optimization target.

Engineering AI communication systems therefore requires managing these trade-offs explicitly, with design decisions documented and empirically validated rather than assumed.

4. DISCUSSION

The TIE framework carries several practical implications for AI communication engineering. First, interpretability features at all three levels must be designed as first-class interface elements rather than post-hoc additions. Research on XAI consistently shows that explanations are only effective when they are legible, timely, and relevant to the user's decision context (Miller, 2019; Cai *et al*., 2023). This applies differentially across the three levels: model-level transparency is best communicated through system-level documentation pre-interaction, while interface-level affordances must be continuously accessible and contextually prompted during task execution.

Second, trust mechanisms must be empirically validated with target user populations. There is significant heterogeneity in how different user groups—experts, novices, high-stakes decision-makers—exhibit and update cognitive trust, affective trust, and behavioral reliance. One-size-fits-all transparency designs are unlikely to be effective across these populations. Future empirical work should decompose trust calibration effects by user group and task domain.

Third, efficiency optimization cannot be pursued in isolation. Interface designs that prioritize rapid response and minimal friction at the cost of interpretability may produce short-term efficiency gains but erode long-term trust calibration and appropriate use patterns. This suggests that evaluation frameworks for GenAI communication systems should incorporate longitudinal measures of all four efficiency metrics alongside conventional task completion indicators.

Fourth, governance requirements must be treated as engineering constraints, not external impositions. Accountability logging, misuse detection, and redress mechanisms impose design costs that must be budgeted explicitly. Conversely, well-designed governance structures can reinforce both cognitive trust (users know errors are tracked and correctable) and behavioral reliance calibration (users understand that the system operates within defined boundaries).

A limitation of the current work is its primarily conceptual nature. The TIE framework is grounded in the reviewed literature but has not yet been empirically

validated as a unified model. Future work should include controlled experiments testing the interactive effects of trust calibration (across cognitive, affective, and behavioral dimensions), interpretability level (model-, system-, interface-level), and efficiency metrics (TCT, TCpT, UCR, CL) in GenAI communication interfaces across different domains and user populations. Extension to multimodal and embodied AI systems will require substantive framework revision.

5. EVALUATION RECOMMENDATIONS AND METRIC FRAMEWORK

To support practitioner utility, the TIE framework is accompanied by a structured metric table and evaluation checklist. These instruments are intended to guide design teams in operationalizing the framework's dimensions during both formative evaluation (informing iterative design) and summative evaluation (assessing deployed system performance).

5.1. Metric Table

The checklist is intended to be completed during design review milestones, and each Critical-priority item should be treated as a blocking concern before deployment in high-stakes domains.

CONCLUSION

This paper has examined three foundational engineering dimensions of human–machine communication in the generative AI era: trust calibration (decomposed into cognitive, affective, and behavioral components), multi-level interpretability (model-, system-, and interface-level), and interaction efficiency (operationalized through task completion time, turn count, user correction rate, and cognitive load). The proposed TIE framework integrates these dimensions into a coherent conceptual model that highlights their interdependencies and positions them explicitly against established HCI usability models and AI trust taxonomies.

Ethical and governance dimensions—accountability, misuse prevention, and user redress—are embedded as co-equal engineering concerns within the framework rather than external constraints. The framework's primary scope is LLM-based conversational AI; applicability to multimodal and embodied AI systems requires further extension.

Dimension	Metric / Indicator	Measurement Approach	Target Range / Threshold
Trust Calibration	Overtrust / Undertrust Rate	Post-task behavioral audit; AI-error acceptance rate	< 15% uncritical acceptance of flagged errors
	Trust Calibration Score (TCS)	Survey instrument (*e.g.*, TAS-adapted for AI)	TCS ≥ 0.70 on normalized scale
	Affective Trust Index	Self-report scales (warmth, comfort items)	Correlated with continued use intent
Interpretability	Explanation Comprehension Rate	Cloze-style post-task quiz on AI reasoning	> 70% correct on reasoning questions
	Explanation Engagement Rate	Log analysis of XAI affordance usage	> 40% of sessions engage ≥ 1 affordance
	Interface-level XAI Click-through	Session log; click events on 'Why?' / 'Sources'	Benchmark against domain baseline
Interaction Efficiency	Task Completion Time (TCT)	Timed task completion log	Compared to non-AI baseline
	Turn Count per Task (TCpT)	Dialogue log analysis	Reduction vs. initial session
	User Correction Rate (UCR)	Proportion of AI outputs requiring user edits	< 20% for routine tasks
	Cognitive Load (NASA-TLX)	NASA Task Load Index after each session	Mean TLX < 50 (moderate threshold)

Note: Thresholds are indicative starting points based on current literature and should be calibrated to specific deployment domains (*e.g.*, healthcare or legal contexts may require stricter targets for UCR and overtrust rate).

5.2. Evaluation Checklist

Category	Checklist Item	Priority
Trust Calibration	Does the system include explicit uncertainty signaling for low-confidence outputs?	Critical
	Is behavioral consistency validated across semantically equivalent prompts?	High
	Are cognitive, affective, and behavioral trust mechanisms separately addressed?	High
	Does graceful failure handling prevent plausible but incorrect responses?	Critical
Interpretability	Are model-level explanations (*e.g.*, uncertainty scores) accessible to users?	High
	Are system-level explanations (documentation, capability scope) provided pre-interaction?	Medium
	Do interface-level affordances (*e.g.*, 'Show Sources') exist and get used?	High
	Is explanation comprehension validated with target user groups?	High
Interaction Efficiency	Are Task Completion Time and Turn Count per Task measured at baseline?	Critical
	Is User Correction Rate tracked per task category?	High
	Is cognitive load (NASA-TLX) assessed in longitudinal studies?	Medium
	Does progressive disclosure adapt to observed user mental model accuracy?	High
Ethics & Governance	Is there a documented accountability chain for AI-generated outputs?	Critical
	Are misuse scenarios identified and mitigated in design?	High
	Are limitations relative to multimodal/embodied AI systems acknowledged?	Medium
	Is there a user redress mechanism for consequential AI errors?	Critical

As generative AI systems become increasingly embedded in communication-critical applications, the engineering of appropriate trust, transparent interpretability, and sustained interaction efficiency becomes a societal as well as a technical imperative. The TIE framework, together with the accompanying metric table and evaluation checklist, offers a structured starting point for researchers and practitioners engaged in this work. Empirical validation and domain-specific refinement remain necessary to realize its full potential.

REFERENCES

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138-52160. https://doi.org/10.1109/ACCESS.2018.2870052

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., … Horvitz, E. (2019). Guidelines for human-AI interaction. Proceedings of CHI 2019. https://doi.org/10.1145/3290605.3300233

Anthropic. (2023). Claude model card. Retrieved from https://www.anthropic.com/

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., … Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Band, N., Lakshminarayanan, B., & Ghassemi, M. (2024). Linguistic uncertainty expression in large language models. Proceedings of ICLR 2024.

Cai, C. J., Winter, S., Steier, D., Rabelo, L., & Terry, M. (2023). Hello AI: Uncovering the onboarding needs of medical clinicians for human-AI collaborative decision-making. ACM CHI Conference on Human Factors in Computing Systems.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. Advances in Psychology, 52, 139-183. https://doi.org/10.1016/S0166-4115(08)62386-9

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human Factors, 57(3), 407-434. https://doi.org/10.1177/0018720814547570

ISO. (2018). ISO 9241-11: Ergonomics of human-system interaction—Part 11: Usability: Definitions and concepts. International Organization for Standardization.

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). https://doi.org/10.1073/pnas.2208839120

Jiang, J., *et al*. (2023). Interaction efficiency in large language model chat interfaces: An empirical analysis. Proceedings of CHI 2023.

Kasirzadeh, A., & Gabriel, I. (2023). In conversation with artificial intelligence: Aligning language models with human values. Philosophy & Technology, 36(2), 1-24. https://doi.org/10.1007/s13347-023-00606-x

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50-80. https://doi.org/10.1518/hfes.46.1.50.30392

Liao, Q. V., & Vaughan, J. W. (2023). AI transparency in the age of LLMs: A human-centered research roadmap. Harvard Data Science Review. https://doi.org/10.1162/99608f92.8036d03b

Lipton, Z. C. (2018). The mythos of model interpretability. Queue, 16(3), 31-57. https://doi.org/10.1145/3236386.3241340

Luger, E., & Sellen, A. (2016). "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. Proceedings of CHI 2016. https://doi.org/10.1145/2858036.2858288

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38. https://doi.org/10.1016/j.artint.2018.07.007

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., … Gebru, T. (2019). Model cards for model reporting. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3287560.3287596

Nielsen, J. (1994). Usability Engineering. Academic Press. https://doi.org/10.1016/B978-0-08-052029-2.50007-3

OpenAI. (2023). GPT-4 system card. Retrieved from https://openai.com/research/gpt-4

Perez, F., & Ribeiro, I. (2022). Ignore previous prompt: Attack techniques for language models. arXiv preprint arXiv:2211.09527.

Seshia, S. A., *et al*. (2023). Toward verified artificial intelligence. Communications of the ACM, 66(7), 86-98.

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

Steyvers, M., Tejeda, H., Kerrigan, G., & Smyth, P. (2023). Bayesian modeling of human-AI complementarity. Proceedings of the National Academy of Sciences, 119(11). https://doi.org/10.1073/pnas.2111547119

Vaswani, A., *et al*. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

Vossing, M., *et al*. (2022). Designing transparency for effective human-AI collaboration. Information Systems Frontiers, 24(3), 877-895. https://doi.org/10.1007/s10796-022-10284-3

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841-887. https://doi.org/10.2139/ssrn.3063289

Weidinger, L., *et al*. (2022). Taxonomy of risks posed by language models. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3531146.3533088

Winograd, T., & Flores, F. (1986). Understanding Computers and Cognition. Norwood, NJ: Ablex Publishing.

Woebbecke, G., Hauk, N., Siebers, M., & Buettner, R. (2024). Affective dimensions of trust in AI-mediated communication: A large-sample empirical study. Computers in Human Behavior, 152, 107988.

Xu, Y., *et al*. (2023). Perceived agency and trust in conversational AI: A large-scale empirical study. ACM Transactions on Computer-Human Interaction, 30(2).

https://doi.org/10.31875/2979-1081.2026.02.07