From Unilateral Behavioral Tuning to Human–AI Co-Configuration

Rong-Ching Chang
University of California, Davis
Davis, California, USA
rocchang@ucdavis.edu

Hao-Chuan Wang
University of California, Davis
Davis, California, USA
hciwang@ucdavis.edu

Abstract

The HCI community increasingly treats large language model (LLM)-based agents as collaborating companions, yet people often infer beliefs and intentions from observed AI behaviors that current systems do not actually have. In this position paper, we present an iceberg metaphor as a conceptual framework to separate visible AI behavior from hidden assumptions and mental misattributions. We propose three paradigms for designing human-AI collaboration —free-form interaction, behavioral engineering, and human-AI co-configuration—as a lens to conceptualize this gap. We argue that HCI and AI communities should jointly engage in the third paradigm, human-AI co-configuration, focusing on transparentizing and structuring interactions through co-configurable contractual agreements rather than unilateral behavior tuning alone. We illustrate this emerging paradigm through three design opportunities: long-term calibration with interaction-level indicators, explicit interaction contracts users can adjust, and shared protocols that supports trust calibration.

CCS Concepts

• **Human-centered computing → HCI theory, concepts and models**.

Keywords

Human-AI co-configuration, theory of mind

ACM Reference Format:
Rong-Ching Chang and Hao-Chuan Wang. 2026. From Unilateral Behavioral Tuning to Human–AI Co-Configuration. In *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Grounding

The HCI community increasingly frames LLM-powered agents not as tools but as collaborative partners in joint work [14]. This raises a core question: what happens when a "collaborator" evokes social cues without the internal commitments that make such collaboration dependable?

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI, Barcelona, Spain
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN
https://doi.org/XXXXXXX.XXXXXXX

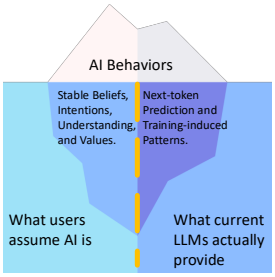

Figure 1: The iceberg framework contrasts visible AI behavior with hidden user assumptions and model processes, creating asymmetric inference where AI can often infer users better than users can infer AI commitments.

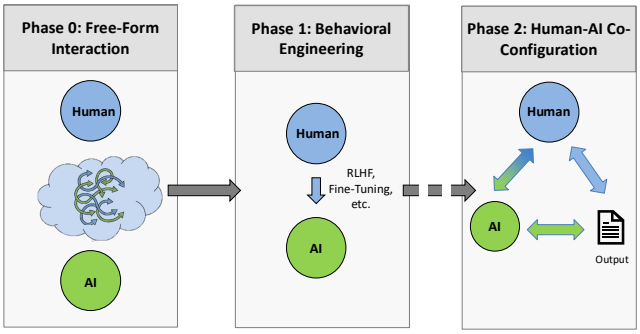

Figure 2: Three interaction design paradigms for human-AI interactions.

Theory of Mind (ToM)—the ability to attribute beliefs, intentions, and desires to others—is central to human-human collaboration. Evaluations of AI's readiness to make ToM attributions of humans showed mixed results: GPT-4 performs at or above human levels on some ToM tasks while struggling on others [12], and LLM-based agents show emergent collaboration (e.g. delegation, leadership) alongside systematic multi-agent failures such as failing to hallucination [3]. Nevertheless, the reverse direction is equally important for the human-AI alignment in interactions: do people routinely attribute ToM *to* AI? For example, it has been noted that users may apply social rules to computers, with or without AI, even without believing machines have feelings [4], and AI-powered agents, such as conversational LLMs, may further intensify this attribution with observable behavioral cues such as fluent language use in dialogues and anthropomorphic features [2]. As Suchman argues,

such interaction is fundamentally asymmetric: systems access human action through a "very small keyhole," while people tend to interpret interaction with their own rich, situated contexts [13].

This reveals a persistent mismatch between what users can infer and what deployed LLM systems can guarantee and deliver. Users read linguistic fluency, persona consistency, and value-signaling responses as evidence of stable internal beliefs, intentions, or values; deployed systems remain next-token predictors that can at best approximate stances of human roles, without stable nor inspectable commitments [8]. This is not merely a problem of AI literacy: automation bias such as favoring autonomy over non-autonomous approaches regardless of efficacy, tends to persist among novices and experts, resisting training-based correction [5], so even informed users can have mismatched expectations beyond what the system warrants. Our observation of this mismatch leads to the discussion of interaction paradigms presented in Section 3.

2 The Iceberg: Conceptualizing Misattribution

We use an *iceberg* metaphor to represent issues of misattribution in human-AI interactions (see Figure 1). Above the waterline is *AI behavior*: what the system says and does in interactions. Below are two layers that users and designers often conflate: users' assumed stable beliefs, intentions, understanding, and values, and what current LLMs *actually* provide—next-token prediction and training-induced patterns that can mimic stances without stable mental states. The gap between these layers clarifies where alignment interventions operate. In deployed systems, most alignment practices (e.g., RLHF, safety fine-tuning) works *primarily* above the waterline by shaping outputs. Adjacent work—including interpretability, mechanistic analysis, and causal evaluation—seeks to connect behavior to underlying mechanisms, but is not yet consistently translated into user-facing affordances for everyday collaboration.

When optimization targets visible behavior, recurring failure modes emerge: sycophancy (agreement over truth) [9], over-refusal (lexical refusal without intent assessment) [7], and a value-action gap (stated commitments that do not reliably regulate downstream behaviors) [10]. Across these cases, systems are tuned for output quality while users can read outputs as signs of internal states. Paradoxically, better above-the-waterline behavior can widen the mismatch. As AI behavior becomes more fluent and responsive, users' "underwater assumptions" of internal states can look increasingly plausible to users despite its actual working [5]. Anthropomorphic attribution can inflate perceived capability and skew moral and trust judgments [6].

This mismatch also appears longitudinally. Analysis of human-LLM conversations shows that, as conversations unfold, LLMs converge with users on *surface* linguistic features while drifting on psychological and contextual dimensions [1] over time. Interaction can look increasingly aligned above the waterline as the model mirrors a user's style, while below it, users infer deeper understanding even as the model continues producing statistically similar tokens without deeper convergence.

To design AI as collaborating partners, we identify three *interaction design paradigms* of human-AI collaboration in reflection of the current states and future prospect. They differ in how the interaction processes are configured, initiated, and continued over time, and in how they address the recurring misattribution and mismatch as identified.

3 Three Paradigms of Human-AI Collaboration

Paradigm 0: Free-Form Interaction. Collaboration is largely ad hoc: users issue requests, generic models respond, and there's no or limited configuration effort in attempting to coordinate human versus AI behaviors across turns. Free-form interaction can work for lightweight tasks, but it provides weak scaffolding for sustained joint work. Assumptions users made remain implicit, expectations can drift, and there's no explicit design for handling human-AI misalignment.

Paradigm 1: Behavioral Engineering (Current). The current paradigm is *unilateral behavioral tuning*: optimizing outputs to appear helpful, harmless, and honest via RLHF, safety tuning, and model specifications. This has improved usability and safety in many contexts, yet persistent failure patterns remain: sycophancy [9], over-refusal [7], and unresolved tensions among plural values [11]. These are not isolated edge cases. Behavior does not equate belief [10]; LLMs can simulate social roles but role nuances and plural values cannot be collapsed into a single optimization target [11]. In Suchman's term, tuning reproduces the planning-model error: seeing alignment as a specification to encode and replicate, rather than as situated action continuously worked out in practice [13]. The result is a structural mismatch: more polished behavior can invite more attributions and following expectations that aren't currently situated, risking greater misalignment.

Paradigm 2: Human-AI Co-Configuration (Emerging). This paradigm shifts from unilateral tuning to *co-configuration*: transparent, re-structured human-AI interaction through co-configurable contractual agreements. Instead of fixing a collaboration plan at design time, both parties can shape goals, constraints, evidentiary standards, and ambiguity-handling strategies during use or the run-time. System behaviors would remain open to ongoing negotiation and reworking, not frozen by general or historical priors [13]. In this co-configuration paradigm, alignment becomes an interaction property that is maintained, monitored, and revised dynamically over time, with user-verifiable evidence rather than guesswork.

(1) **Long-term calibration via interaction indicators.** Co-configuration should support continuous adaptation across repeated collaboration, using interaction history to reveal misalignment and support co-configuration, and provide indicators when system behaviors could be misperceived.
(2) **Co-configurable interaction contracts.** Co-configuration should make task-level accounts explicit and revisable (goals, evidence thresholds, boundaries, fallback policies), so coordination rests on negotiated commitments rather than inferred intents of both sides.
(3) **Shared protocols for trust re-calibration.** Co-configuration should establish protocols for how both user and system respond when goals, confidence, or context shifts, and clearly signal these changes as interaction unfolds, enabling both sides to re-plan, diagnose breakdowns, and recalibrate trust based on legible traces and situated contexts.

References

[1] Rong-Ching Chang and Hao-Chuan Wang. 2025. Convergence, Reciprocity, and Asymmetry: Communication Accommodation Between Large Language Models and Users Across Cultures. In *Proceedings of the 47th Annual Meeting of the Cognitive Science Society*.

[2] Nicholas Epley, Adam Waytz, and John T. Cacioppo. 2007. On seeing human: A three-factor theory of anthropomorphism. *Psychological Review* 114, 4 (2007), 864–886. doi:10.1037/0033-295X.114.4.864

[3] Huao Li, Yu Quan Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Michael Lewis, and Katia Sycara. 2023. Theory of Mind for Multi-Agent Collaboration via Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 180–192.

[4] Clifford Nass and Youngme Moon. 2000. Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues* 56, 1 (Jan. 2000), 81–103. doi:10.1111/0022-4537.00153

[5] Raja Parasuraman and Dietrich H. Manzey. 2010. Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors* 52, 3 (2010), 381–410. doi:10.1177/0018720810376055

[6] Adriana Placani. 2024. Anthropomorphism in AI: hype and fallacy. *AI and Ethics* 4, 3 (Feb. 2024), 691–698. doi:10.1007/s43681-024-00419-4

[7] Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 5377–5400.

[8] Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature* 623, 7987 (Nov. 2023), 493–498. doi:10.1038/s41586-023-06647-8

[9] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. 2024. Towards understanding sycophancy in language models. In *Proceedings of the 12th International Conference on Learning Representations*.

[10] Hua Shen, Nicholas Clark, and Tanu Mitra. 2025. Mind the Value-Action Gap: Do LLMs Act in Alignment with Their Values?. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 3097–3118. doi:10.18653/v1/2025.emnlp-main.154

[11] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*.

[12] James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour* 8 (July 2024), 1285–1295. doi:10.1038/s41562-024-01882-z

[13] Lucy Suchman. 2007. *Human-Machine Reconfigurations: Plans and Situated Actions* (2nd ed.). Cambridge University Press.

[14] Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. From Human-Human Collaboration to Human-AI Collaboration: Designing AI Systems That Can Work Together with People. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 1–6. doi:10.1145/3334480.3381069

Received 12 February 2026