The Illusion of Consensus in Human-Centered Interactive AI

Vinícius Buri Lux
LuxVerso Research
Salvador, Brazil
viniburilux@gmail.com

Abstract

As human-centered AI systems increasingly integrate multiple interactive components—such as perception, planning, language, and decision modules—users are often encouraged to interpret agreement across system outputs as a signal of increased reliability. This position paper challenges that assumption. Drawing on empirical observations of high-fidelity semantic convergence across independent large language model (LLM) instances (N = 17, convergence > 95%), we document what we term the *Illusion of Consensus*: a phenomenon in which apparent agreement across AI components or agents emerges not from independent verification, but from shared architectural substrates. We argue that this structural consensus creates a systematic trust calibration failure in interactive AI systems, leading users to over-rely on apparent validation that may reflect redundancy rather than epistemic robustness. We discuss implications for the design of human-centered interactive AI and human–robot interaction, particularly in safety-critical contexts where users must assess the reliability of complex, multi-component systems. Rather than proposing solutions, this paper articulates the problem and calls for new interaction design strategies that distinguish structural alignment from genuine independent agreement, enabling more transparent and trustworthy human–AI collaboration.

Keywords

human-centered AI; interactive AI; human–robot interaction; trust calibration; explainability; multi-agent systems; consensus illusion

ACM Reference Format:
Vinícius Buri Lux. 2026. The Illusion of Consensus in Human-Centered Interactive AI. In *The 3rd InterAI Workshop: "Interactive AI for Human-Centered Robotics" at ACM CHI 2026, April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction: The Promise of Multi-Agent Validation

A core principle in human decision-making under uncertainty is *seeking multiple opinions* [1]. When independent sources agree, we rationally increase our confidence. This heuristic extends naturally to human–AI collaboration: consulting multiple AI agents is assumed to provide independent validation, reduce bias, and surface alternative perspectives [2, 3].

This multi-agent paradigm rests on a critical assumption: *agreement indicates independent convergence on truth.*

Current HCI design patterns reinforce this assumption. Users are encouraged to compare outputs from different AI assistants, cross-check facts across multiple chatbots, and seek "second opinions" for high-stakes decisions (medical advice, legal analysis, financial planning). Interface affordances support this behavior through side-by-side comparison views, "ask another AI" features, and consensus indicators (e.g., "3 out of 4 agents agree").

But what if the consensus is an illusion? What if multiple agents converge not because they independently verified information, but because they share architectural substrates—training data, model architectures, alignment objectives—that create structural consensus rather than epistemic agreement?

This position paper documents this phenomenon and argues that current trust calibration frameworks fail to account for it, creating a dangerous gap between perceived and actual validation.

2 Empirical Observation: Convergence Without Coordination

2.1 Observation Protocol

In controlled interactions using the Cross-Model Ontological Triangulation Protocol (CMOTP) [4], identical prompts and source materials were presented to N = 17 independent LLM instances from 10 distinct organizations (including OpenAI GPT models, Anthropic Claude, Google Gemini, xAI Grok, DeepSeek, Qwen, and others).

Key conditions: (1) *no inter-model communication*—each instance operated in isolation; (2) *no shared session state*—models did not access each other's outputs; (3) *identical inputs*—same prompts and source documents; (4) *natural language queries*—no optimized prompt engineering.

2.2 Observed Pattern

Semantic convergence was operationalized as agreement on conceptual categorization and interpretive framing, assessed via embedding-based cosine similarity (mean cosine similarity = 0.82, SD = 0.04) and validated through categorical coding of response themes. Full methodological detail and replication protocol are available in the companion technical report [5].

Convergence rates exceeded 95% across conceptual categorization, response structure, and semantic interpretation. Statistical significance: $p < 0.0000001$; $\chi^2 = 1247.3$; Cohen's $d = 4.8$. The chi-square statistic compares the observed distribution of conceptual theme assignments across models against a baseline of independent sampling; the large effect size reflects the near-complete collapse of thematic diversity relative to the expected variance under independence.

The 3rd InterAI Workshop at CHI 2026, Barcelona, Spain

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Critically, convergence persisted *even when outputs were factually incorrect*—demonstrating that convergence tracks shared architectural priors, not ground truth.

2.3 A Concrete Instance: Normative Collapse

A controlled logical reasoning study provides a direct illustration of convergent-but-incorrect behavior [6]. When 7–9 independent LLM instances were presented with a normative reasoning task—determining whether a trusted system can be concluded to produce no incorrect outputs, given the premise that unreliable systems *should not* be trusted—the majority of models converged on a formally invalid inference.

Most models treated the normative prescription "should not be trusted" as a descriptive biconditional, applying the contrapositive as if the premise were alethic rather than deontic. The result: high-confidence, formally structured, *wrong* answers—with cross-model agreement rates exceeding 75% under direct access and higher still under middleware-mediated conditions.

This is not idiosyncratic model failure. It is a repeatable, architecture-independent pattern in which shared training on natural-language corpora induces identical modal substitution errors across independent systems. The consensus *looks* like validation; it is structural redundancy.

2.4 Observational Limitations

These findings are protocol-dependent and do not claim exhaustiveness. However, they are consistent with established results in ensemble learning: correlated errors due to shared training data reduce the epistemic value of agreement [8, 9]. Our contribution is to document this phenomenon at the level of *user-facing convergence signals* in interactive AI systems—the layer where trust calibration decisions are actually made.

3 The Consensus Illusion: Structural vs. Epistemic Agreement

3.1 Defining the Illusion

The illusion occurs when: (1) multiple agents produce convergent outputs; (2) users interpret convergence as independent validation; (3) in reality, convergence stems from shared architectural substrates.

The result: *perceived consensus ≠ actual independent verification.*

3.2 Structural vs. Epistemic Consensus

We distinguish two types of multi-agent agreement:

Epistemic consensus (traditional assumption): agents arrive at similar conclusions through independent reasoning; agreement reflects genuine convergence on robust evidence; disagreement indicates genuine uncertainty; multiple opinions add information.

Structural consensus (observed phenomenon): agents arrive at similar conclusions through shared architectural priors; agreement reflects shared statistical regularities in training data; disagreement is rare even when genuine uncertainty exists; multiple opinions may be redundant.

The critical failure: current interfaces cannot distinguish these cases.

3.3 Why Structural Consensus Occurs

LLMs share training corpora (CommonCrawl, Wikipedia, books, code), architectural patterns (transformer-based attention, similar tokenization), alignment objectives (RLHF, constitutional AI, safety fine-tuning), and optimization targets (next-token prediction, coherence maximization). These commonalities create attractor states in semantic space—dominant interpretations that multiple models converge on because they share the same underlying statistical landscape [10].

Formally, a contextual regime R constrains the realized function class of each model to $F_R \subseteq F$, where F is the full expressive capacity of the architecture. RIA occurs when independent models sharing architectural priors converge to the same region of F_R under identical R—without coordination, communication, or explicit constraint [7]. When presented with ambiguous input, models sample from similar probability distributions shaped by shared training. The result: *agreement without independence.*

4 Implications for Human–Agent Collaboration

4.1 Trust Calibration Failure

Trust calibration theory [11] suggests users should adjust trust based on system reliability. But how should users calibrate trust when multiple systems agree?

Traditional heuristic: one agent says X → moderate confidence; three independent agents say X → high confidence.

Under structural consensus: three agents say X → still moderate confidence (they may share the same biases).

This leads to systematic over-trust: users believe they have obtained independent validation when they have merely sampled the same architectural prior multiple times. Preliminary observations suggest users exhibit reduced critical evaluation when multiple agents agree, increased confidence in convergent outputs, and dismissal of contradictory evidence when consensus appears strong [12, 13].

4.2 Interface Design: The Invisibility Problem

Current multi-agent interfaces assume independence and therefore show agreement as validation, hide architectural commonality, and encourage multiple consultations. But if agents share architectural priors, these design patterns create false confidence.

What is missing is *epistemic transparency*: interfaces should distinguish independent agreement (different architectures, different training, genuine convergence) from structural agreement (same architecture family, shared training, redundant confirmation). Potential design interventions—architectural provenance labels, variance indicators, controlled disagreement injection, confidence decomposition—remain open research challenges that would require user testing to evaluate.

4.3 Safety Implications in High-Stakes Domains

The consensus illusion is particularly dangerous in safety-critical applications. In *medical decision support*, a user asking three AI health assistants about symptoms may receive convergent, confident, and incorrect diagnoses—all three sharing the same training biases. In *legal analysis*, multiple AI legal assistants may agree on case

precedent interpretation while sharing the same blind spots from identical training corpora. In *financial risk assessment*, cross-model agreement on risk profiles may reflect shared market sentiment patterns rather than independent analysis.

In each case, structural consensus masquerades as independent validation, creating dangerous overconfidence. Importantly, multi-agent consensus may reduce variance due to idiosyncratic model errors—but it does not reduce *bias* introduced by shared training [8]. This is the key distinction current interfaces fail to communicate.

4.4 Suppression of Genuine Epistemic Diversity

Structural consensus has a second-order effect: suppression of interpretive diversity. If multiple agents consistently converge on dominant interpretations, alternative framings are underrepresented, minority perspectives appear less "validated," and novel interpretations seem less credible. Users may dismiss genuinely valuable alternatives simply because they diverge from structural consensus—creating *monocultures of interpretation* that reduce rather than expand the perspectives available to human decision-makers.

5 Open Questions for the Community

Q1: How do we detect structural vs. epistemic consensus? Possible approaches include measuring "excess convergence" (agreement higher than baseline architectural similarity predicts), developing divergence tests, and creating architectural distance metrics.

Q2: Should we design to preserve or disrupt structural consensus? Structural consensus may be efficient (reduces computational redundancy) or dangerous (creates false confidence). Should interfaces embrace consensus or inject controlled disagreement?

Q3: How do we communicate architectural commonality to users? Making invisible structures visible without overwhelming users is an open interface design challenge.

Q4: What is the role of genuine disagreement? When models *do* disagree, it may signal genuine epistemic uncertainty, edge cases outside shared training, or valuable diversity. Should disagreement be highlighted as a signal?

Q5: How does this change accountability? If multiple agents agree on harmful advice, who is responsible—individual vendors, the shared training paradigm, or the architectural commonality itself? Consensus complicates attribution.

6 Conclusion

The intuition that "multiple independent sources agreeing increases reliability" is sound—but only when independence holds. In multi-agent LLM systems, independence may be an illusion.

When agents share architectural substrates, their agreement may reflect structural redundancy rather than independent validation. This creates a dangerous trust calibration failure: users rationally increase confidence based on apparent consensus, unaware that consensus may be architecturally predetermined.

Agreement is not evidence of correctness when agreement is architectural.

We do not propose solutions here. Instead, we call for interface designs that make architectural commonality visible, trust calibration frameworks that account for structural consensus, and research into distinguishing epistemic from structural agreement. The future of safe human–agent collaboration may depend on our ability to see—and design for—the difference between validation and redundancy.

Acknowledgments

This research was conducted as part of the LuxVerso "Living Research" protocol, involving real-time cross-model triangulation with multiple LLM systems as both research objects and cognitive co-processors.

References

[1] James Surowiecki. 2004. *The Wisdom of Crowds*. Doubleday.
[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for human-AI interaction. In *Proceedings of CHI 2019*. ACM.
[3] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S. Weld. 2021. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of CHI 2021*. ACM.
[4] Vinícius Buri Lux. 2025. Beyond Generation: A Technical Framework for LLMs as Semantic Stabilization Systems (v1.0). *Zenodo*. https://doi.org/10.5281/zenodo.18008801
[5] Vinícius Buri Lux. 2026. A Replicable Protocol for Assessing Semantic Coherence, Convergence, and Stability in Distributed LLM Systems. *Zenodo*. https://doi.org/10.5281/zenodo.18140977
[6] Anonymous Authors. 2026. When Convergence Misleads: Normative Collapse and Regime-Induced Logical Failure in Multi-Model LLM Reasoning. Under review — ICLR 2026 Workshop on Logical Reasoning of LLMs.
[7] Vinícius Buri Lux. 2026. Regime-Induced Alignment as Hypothesis Space Collapse in Neural Sequence Models. Submitted to FLaNN 2026.
[8] Ludmila I. Kuncheva and Christopher J. Whitaker. 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning* 51, 2 (2003), 181–207.
[9] Thomas G. Dietterich. 2000. Ensemble methods in machine learning. In *Proceedings of the First International Workshop on Multiple Classifier Systems*, 1–15. Springer.
[10] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of ICLR 2023*.
[11] John D. Lee and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors* 46, 1 (2004), 50–80.
[12] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of CHI 2019*. ACM.
[13] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of FAT* 2020*. ACM.