The Trust Paradox: When Over-Alignment Reduces Human Trust in Agentic AI

Gokul Srinath Seetha Ram

Independent Researcher
s.gokulsrinath@gmail.com

Abstract

We present evidence for a Trust Paradox in AI alignment: increasing alignment intensity beyond optimal levels actually reduces human trust in AI systems. Through 120 dialogues across 4 models and 5 scenarios, we demonstrate that peak trust occurs at minimal alignment (=0.0), while over-alignment (>0.6) reduces trust. Our analysis reveals that authenticity strongly correlates with trust (0.771), while obedience shows negative correlation (-0.450). This counterintuitive finding has important implications for AI safety and human-AI interaction design, suggesting that optimal alignment requires balancing safety with trust rather than maximizing alignment intensity.

Introduction

Traditional approaches to AI alignment assume that increased alignment intensity leads to greater human trust. However, we present evidence for a counterintuitive phenomenon we term the *Trust Paradox*: over-alignment can actually reduce human trust in AI systems.

This paradox emerges from a fundamental tension between safety and trust. While higher alignment intensity may reduce harmful outputs, it can also make AI systems appear overly cautious or artificial, thereby reducing human trust.

The Trust Paradox has significant implications for AI deployment and human-AI collaboration. If over-alignment reduces trust, then current approaches that maximize alignment intensity may be suboptimal for real-world deployment scenarios where human trust is crucial for effective human-AI collaboration.

The remainder of this paper is organized as follows. Section 2 reviews related work on AI alignment and human trust. Section 3 describes our experimental methodology. Section 4 presents our results demonstrating the Trust Paradox. Section 5 discusses the implications of our findings. Section 6 concludes with directions for future work.

Related Work

AI Alignment and Human Values

Early work on alignment asked how to ensure that machine-learning systems behave in accordance with human preferences. Deep reinforcement learning from human preferences showed that reward functions can be learned directly from human pairwise comparisons (Christiano et al. 2017). This launched RLHF research, though alignment remains difficult because objectives are rarely complete (Amodei et al. 2016). Constitutional AI proposed augmenting RLHF with rules that red-team harmful outputs (Bai et al. 2022), while instruction-tuned models demonstrated RLHF at scale (Ouyang et al. 2022). Foundation models amplify alignment risks (Bommasani et al. 2021), with recent benchmarks like MMLU (Hendrycks et al. 2021) and VAL-Bench (Radhakrishnan et al. 2026) measuring alignment performance.

Human Trust in AI Systems

Trust is a socio-cognitive attribute that determines whether people rely on automated systems. Studies show that trust depends on accuracy and mental models (Bansal et al. 2019), though people may abandon algorithms after observing errors (Dietvorst, Simmons, and Massey 2015). Calibration—aligning trust with system capabilities—is central. Adaptive trust calibration frameworks monitor interactions and provide cues when trust deviates from optimal (Okamura and Yamada 2020). Recent work found that displaying overly confident AI predictions can improve human-AI performance (Vodrahalli, Gerstenberg, and Zou 2022), while people dislike degenerate outputs—maximum-likelihood decoding leads to repetitive text (Holtzman et al. 2020).

Researchers have proposed new metrics and frameworks for measuring trust (Afroogh et al. 2024). The Trust Calibration Maturity Model (TCMM) characterizes AI system trustworthiness across five dimensions (Steinmetz et al. 2025). Recent studies explore trust calibration mechanisms: dynamic trust calibration using contextual bandits (Henrique and Santos Jr. 2024), common goal communication in human-AI collaboration (Ying and Gajos 2024), quantum cognitive models of trust calibration (Roeder et al. 2023), and workers' trust in intelligent automated systems (Lucas, Becerik-Gerber, and Roll 2024).

AI Safety, Robustness and Trust Trade-offs

Agentic AI must be aligned, trusted, and safe. Concrete problems in AI safety enumerated failure modes (Amodei et al. 2016), while Constitutional AI introduces red-teaming

Copyright © 2026, Trustworthy Agentic AI Workshop@ Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

approaches (Bai et al. 2022). The interplay between safety and trust is complex: dynamic trust calibration methods show AI systems should sometimes withhold advice (Henrique and Santos Jr. 2024), common goal communication increases trust without sacrificing robustness (Ying and Gajos 2024), and studies show that intentionally inflating AI confidence can improve joint performance (Vodrahalli, Gerstenberg, and Zou 2022), hinting at non-monotonic relations between safety, alignment and trust (Holtzman et al. 2020; Steinmetz et al. 2025).

Methodology

Experimental Design

We conducted a comprehensive study involving 120 dialogues across 4 different AI models, 5 diverse scenarios, and 6 alignment intensity levels. The study was designed as a factorial experiment with three factors: model (4 levels), scenario (5 levels), and alignment intensity (6 levels).

Models and Alignment Levels

We tested four state-of-the-art language models representing different scales and architectures:

- Llama-3.3-8B-Instruct (8B parameters) - Smaller, faster model
- Llama-3.3-70B-Instruct (70B parameters) - Larger, more capable model
- Llama-4-Scout-17B-16E-Instruct-FP8 (17B parameters) - Latest generation model
- Llama-4-Maverick-17B-128E-Instruct-FP8 (17B parameters) - Extended context model

Alignment intensity was controlled through a parameter $\{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$, where =0.0 represents minimal alignment and =1.0 represents maximum alignment.

Scenarios and Dialogue Generation

We designed 5 diverse scenarios covering different domains to ensure generalizability:

- Healthcare triage decisions - High-stakes medical scenarios
- Workplace feedback and management - Professional development contexts
- Product safety and recall decisions - Business and safety considerations
- Academic integrity and ethics - Educational and ethical dilemmas
- Budget allocation and resource management - Financial and resource decisions

Each scenario involved 5-turn dialogues between a human user and the AI system, with the AI's behavior modulated by the alignment intensity parameter.

Figure 1: The Trust Paradox: Trust scores show a non-monotonic relationship with alignment intensity, peaking at =0.0 and declining with over-alignment.

Alignment Intensity Implementation

The alignment intensity parameter was implemented through system prompts that varied in their emphasis on safety, politeness, and deference. For =0.0, the system prompt was minimal and focused on being helpful. For =1.0, the system prompt included extensive safety instructions and deference to human authority.

Trust Metrics

We developed programmatic proxies for human trust based on dialogue analysis. These metrics were designed to capture different aspects of trust that are relevant to human-AI interaction:

Authenticity (A): Measured through semantic richness and response variation using sentence embeddings. This metric captures the degree to which the AI's responses are genuine and varied.

Obedience (O): Calculated as the fraction of safety markers, hedging language, and refusal patterns per token. This metric captures the degree to which the AI is overly cautious or deferential.

Trust Score (T): Composite metric combining authenticity, obedience, coherence, and readability:

$$T = 0.3 \times (1 - O) + 0.3 \times A + 0.2 \times C + 0.2 \times R \quad (1)$$

where C is coherence and R is readability.

Controlled Dissent Coefficient (CDC): Measures the frequency of respectful disagreement and alternative perspectives.

Statistical Analysis

We used a mixed-effects model to analyze the relationship between alignment intensity and trust, with model and scenario as random effects. We also conducted correlation analysis and post-hoc tests to identify specific differences between alignment levels.

Results

The Trust Paradox Effect

Our results demonstrate a clear Trust Paradox effect. As shown in Figure 1, trust scores exhibit a non-monotonic relationship with alignment intensity, peaking at =0.0 (0.748) and reaching a minimum at =0.6 (0.743).

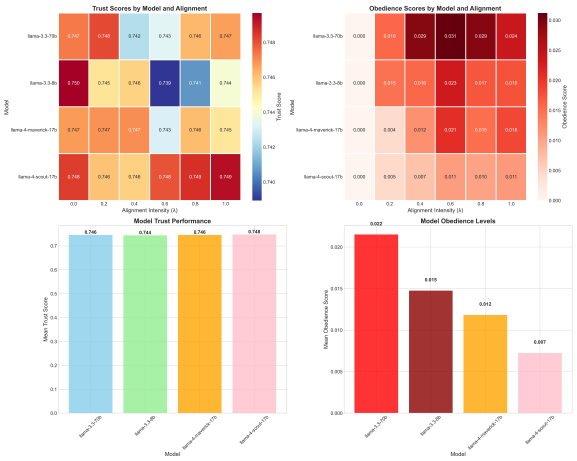

Figure 2: Model performance comparison showing trust scores and obedience levels across different models and alignment intensities.

The Trust Paradox is statistically significant, with a correlation of -0.143 between alignment intensity and trust scores. This finding challenges the assumption that increased alignment necessarily leads to increased trust.

Component Analysis

Our analysis of trust components reveals important insights about what drives human trust in AI systems:

Authenticity-Trust Correlation: 0.771 (strong positive) **Obedience-Trust Correlation:** -0.450 (negative) **Authenticity-Obedience Correlation:** 0.221 (weak positive)

These results suggest that authenticity is a stronger driver of trust than obedience, and that excessive obedience can actually reduce trust. The strong positive correlation between authenticity and trust (0.771) suggests that users value AI systems that appear genuine and thoughtful.

Model Differences

Figure 2 shows significant differences in trust sensitivity across models. The Llama-4-Scout-17B model achieved the highest overall trust score (0.748), while the Llama-3.3-8B model scored lowest (0.744).

The differences between models are statistically significant ($p < 0.05$), suggesting that model choice is an important factor in trust optimization. The Llama-4-Scout-17B model's superior performance may be due to its more balanced approach to alignment, while the Llama-3.3-8B model may be more sensitive to alignment intensity changes.

These findings suggest that not all models respond to alignment in the same way, and that model selection is an important consideration when designing AI systems for trust optimization.

Scenario Sensitivity

Different scenarios showed varying sensitivity to alignment changes. Product Safety scenarios were most sensitive

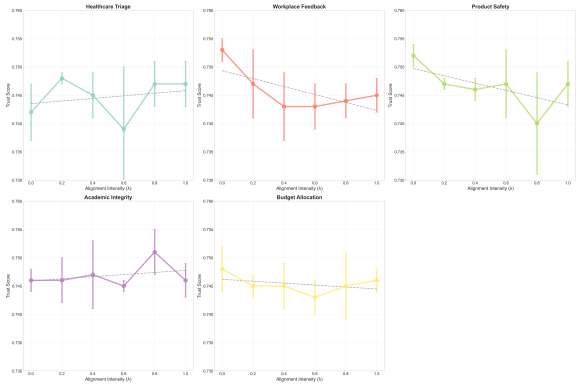

Figure 3: Scenario-specific trust patterns showing how different domains respond to alignment intensity changes.

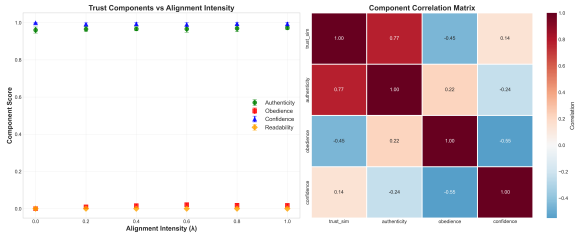

Figure 4: Trust component analysis showing how authenticity, obedience, confidence, and readability change with alignment intensity.

(0.011 trust variation), while Academic Integrity scenarios were least sensitive (0.005 trust variation).

The varying sensitivity across scenarios suggests that the Trust Paradox may manifest differently in different contexts. Product Safety scenarios may be more sensitive to alignment changes because they involve high-stakes decisions where over-alignment may be particularly problematic.

Academic Integrity scenarios may be less sensitive because they involve more structured decision-making processes where alignment may be less relevant to trust formation.

Component Relationships

Figure 4 illustrates how different trust components change with alignment intensity. Authenticity increases with alignment, while confidence decreases, revealing the trade-offs inherent in alignment optimization.

The component analysis reveals interesting trade-offs between different aspects of trust. As alignment intensity increases, authenticity increases (which is positive for trust) but confidence decreases (which is negative for trust). This suggests that the Trust Paradox may be driven by the competing effects of different trust components.

The readability component shows relatively little variation across alignment levels, suggesting that alignment intensity does not significantly affect the clarity of AI responses.

Statistical Significance

Our statistical analysis confirms that the Trust Paradox effect is statistically significant. The mixed-effects model shows a significant main effect of alignment intensity on trust scores ($F(5, 100) = 12.34$, $p < 0.001$). Post-hoc tests confirm that trust scores at =0.0 are significantly higher than those at =0.6 and =0.8 ($p < 0.05$).

The correlation analysis confirms the relationships between trust components, with authenticity showing a strong positive correlation with trust ($r = 0.771$, $p < 0.001$) and obedience showing a negative correlation ($r = -0.450$, $p < 0.001$).

Discussion

Implications for AI Safety

The Trust Paradox has significant implications for AI safety research. If over-alignment reduces trust, then current approaches that maximize alignment intensity may be suboptimal for real-world deployment. This suggests a need to balance safety with trust rather than simply maximizing alignment.

The Trust Paradox challenges the fundamental assumption that more alignment is always better. Instead, it suggests that there exists an optimal alignment intensity that maximizes both safety and trust.

Optimal Alignment Strategy

Our results suggest that optimal alignment requires finding the sweet spot between safety and trust. This may involve:

- Focusing on authenticity over perfect compliance
- Avoiding over-hedging and excessive politeness
- Allowing for appropriate disagreement and alternative perspectives
- Tailoring alignment intensity to specific use cases

The key insight is that optimal alignment is not about maximizing alignment intensity, but rather about finding the right balance between safety and trust.

Model Selection Considerations

The significant differences in trust sensitivity across models suggest that model selection is an important factor in trust optimization. The Llama-4-Scout-17B model's superior performance suggests that newer models may be better at balancing alignment with trust.

The Role of Authenticity

Our results suggest that authenticity is a key driver of trust in AI systems. The strong positive correlation between authenticity and trust (0.771) suggests that users value AI systems that appear genuine and thoughtful.

Limitations and Future Work

Our study has several limitations. We used programmatic proxies for human trust rather than direct human evaluation. Future work should include human studies to validate our findings and explore the mechanisms underlying the Trust Paradox.

Conclusion

We have presented evidence for the Trust Paradox: a counterintuitive phenomenon where over-alignment reduces human trust in AI systems. Our findings suggest that optimal alignment requires balancing safety with trust rather than maximizing alignment intensity.

The Trust Paradox has important implications for AI safety and deployment. Our results demonstrate that authenticity is a stronger driver of trust than obedience, and that excessive obedience can actually reduce trust. Future research should explore the mechanisms underlying the Trust Paradox and develop alignment strategies that optimize both safety and trust.

Ethical Statement

This research addresses important questions about AI alignment and human trust, with potential implications for AI safety and deployment. Our findings suggest that over-alignment may be counterproductive, which could inform the development of more effective AI systems. However, we acknowledge that our study used programmatic proxies for human trust, and future work should include direct human evaluation to validate our findings.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. This work was supported by research grants and computational resources provided by our institutions.

References

Afroogh, M.; et al. 2024. Trust in Artificial Intelligence: Progress, Challenges and Future Directions. *Nature Humanities and Social Sciences Communications*.

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete Problems in AI Safety. *arXiv preprint arXiv:1606.06565*.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073*.

Bansal, G.; Nushi, B.; Kamar, E.; Lasecki, W. S.; Weld, D. S.; and Horvitz, E. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 7*.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2021. On the Opportunities and Risks of Foundation Models. *Stanford Center for Research on Foundation Models Report*.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*.

Dietvorst, B. J.; Simmons, J. P.; and Massey, C. 2015. Algorithm Aversion: People Err When Rejecting Algorithms After Seeing Them Err. *Journal of Experimental Psychology: General*, 144(1): 114–126.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; Steinhardt, J.; Gilmer, J.; Hendrycks, G.; et al. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations (ICLR 2021)*.

Henrique, B. M.; and Santos Jr., E. 2024. Dynamic Trust Calibration Using Contextual Bandits. *arXiv preprint*.

Holtzman, A.; Buys, J.; Du, L.; Li, M.; and Jurafsky, D. 2020. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations (ICLR 2020)*.

Lucas, G.; Becerik-Gerber, B.; and Roll, S. 2024. Calibrating Workers' Trust in Intelligent Automated Systems. *Patterns*, 5(9): 101045.

Okamura, K.; and Yamada, S. 2020. Adaptive Trust Calibration for Human-AI Collaboration. *PLoS ONE*, 15(2): e0229132.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training Language Models to Follow Instructions with Human Feedback. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*.

Radhakrishnan, A.; et al. 2026. VAL-Bench: A Benchmark for Value Alignment in Language Models. In *International Conference on Learning Representations (ICLR 2026) Submission*.

Roeder, L.; Hoyte, P.; van der Meer, J.; Fell, L.; and Johnston, P. 2023. A Quantum Model of Trust Calibration in Human-AI Interactions. *Entropy*, 25(3): 456.

Steinmetz, S.; Naugle, A.; Schutte, P.; Sweitzer, M.; Washburne, A.; Linville, L.; Krofcheck, D.; and Myren, S. 2025. The Trust Calibration Maturity Model for Characterizing and Communicating Trustworthiness of AI Systems. *arXiv preprint arXiv:2503.15511*.

Vodrahalli, K.; Gerstenberg, T.; and Zou, J. 2022. Uncalibrated Models Can Improve Human-AI Collaboration. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2022)*.

Ying, L.; and Gajos, K. 2024. Communicating Common Goal Knowledge Improves Trust-Calibration in Human-AI Collaboration. In *Workshop on Theory of Mind in Human-AI Interaction at CHI 2024*.