## AI trust paradox

The **AI trust paradox** (also known as the **verisimilitude paradox**) is the phenomenon where advanced artificial intelligence models become so proficient at mimicking human-like language and behavior that users increasingly struggle to determine if the information generated is accurate or simply plausible.[1]

Unlike earlier concerns such as Moravec's paradox, which highlighted the surprising difficulty in replicating *simple* human functions in AI, and the automation paradox, which deals with balancing automation and human control, the AI trust paradox specifically addresses the issue of *verisimilitude*—the appearance of truth that leads to misplaced trust.[2][3][*page needed*] The newer challenge arises from the inherent difficulty for users in distinguishing between genuine and misleading content produced by large language models (LLMs) as they become more adept at generating natural and contextually appropriate responses.[4]

### History

In the paper, *The AI Trust Paradox: Navigating Verisimilitude in Advanced Language Models* by Christopher Foster-McBride, published by Digital Human Assistants, the evolution of large language models (LLMs) was explored through a comparative analysis of early models and their more advanced successors.[5][*unreliable source?*] Foster-McBride demonstrated that newer LLMs, with improved architecture and training on extensive datasets, showed significant advancements across key performance metrics, including fluency and contextual understanding.[5] However, this increased sophistication made it increasingly difficult for users to detect inaccuracies, also known as hallucinations.[5]

Foster-McBride highlighted that the newer models not only provided more coherent and contextually appropriate responses but also masked incorrect information more convincingly.[5] This aspect of AI evolution posed a unique challenge: while the responses appeared more reliable, the underlying verisimilitude increased the potential for misinformation going unnoticed by human evaluators.[5]

The study concluded that as models became more capable, their fluency led to a rising trust among users, which paradoxically made discerning false information harder.[5] This finding has led to subsequent discussions and research focusing on the impact of model sophistication and fluency on user trust and behavior, as researchers investigate the implications of AI-generated content that can confidently produce misleading or incorrect information.[5]

### Relation to other paradoxes

The AI trust paradox can be understood alongside other well-known paradoxes, such as the automation paradox, which addresses the complexity of balancing automation with human oversight. Similar concerns arise in Goodhart's law, where an AI's optimization of specified objectives can lead to unintended, often negative, outcomes.[6][7][*page needed*]

These paradoxes highlight that trust in AI is not only technical but behavioral and organizational. Several implementation-stage strategies can help resolve them, including early user involvement, clear accountability structures, and explainable interfaces.[8]

### Current research and mitigation strategies

Addressing the AI trust paradox requires methods such as reinforcement learning with human feedback (RLHF), which trains AI models to better align their responses with expected norms and user intentions.[9][10][11]

Efforts in trustworthy AI focus on making AI systems transparent, robust, and accountable to mitigate the risks posed by the AI trust paradox. Current research in AI safety aims to minimize the occurrence of hallucinations and ensure that AI outputs are both reliable and ethically sound.[12][13][*page needed*]

### See also

- AI effect
- AI alignment
- Polanyi's paradox

### References

1. Trisha Ray,*"The paradox of innovation and trust in Artificial Intelligence". orfonline.org. 22 February 2024. Retrieved 1 October 2024.*

2. Roger Vergauwen & Rodrigo González,*"On the verisimilitude of artificial intelligence". Retrieved 1 October 2024.*

3. Russell, Stuart; Norvig, Peter (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson. ISBN 978-0-13-750513-5.

4. *"The LLM Paradox: High Expectations Coupled With Lack of Trust". theinformation.com. 14 August 2024. Retrieved 1 October 2024.*

5. Christopher Foster-McBride (25 April 2024). "The AI Trust Paradox: Navigating Verisimilitude in Advanced Language Models". Digital Human Assistants. Retrieved 11 September 2024.

6. Al Bowman,*"Humans vs AI: The Trust Paradox". mindfoundry.ai. 29 July 2023. Retrieved 1 October 2024.*

7. Moravec, Hans (1988). *Mind Children: The Future of Robot and Human Intelligence*. Harvard University Press. ISBN 978-0-674-57618-6.

8. *Bakonyi, Zoltán (2024-01-01). "How can companies handle paradoxes to enhance trust in artificial intelligence solutions? A qualitative research". Journal of Organizational Change Management. **37** (7): 1405–1426. doi:10.1108/JOCM-01-2023-0026. ISSN 0953-4814.*

9. Dennis Hillemann,*"The Trust Paradox: Will AI in the Public Sector Trust Humans, and Should We Trust AI?". dhillemann.medium.com. 30 June 2023. Retrieved 1 October 2024.*

10. Ng, Andrew (November 2016). *What Artificial Intelligence Can and Can't Do Right Now*. Harvard Business Review.

11. Unkelbach, Christian; Bayer, Myriam; Alves, Hans; Koch, Alex; Stahl, Christoph (2011). *Fluency and positivity as possible causes of the truth effect*. Consciousness and Cognition. 20 (3): 594–602. doi:10.1016/j.concog.2010.09.015. PMID 21111638.

12. Sarah Kreps,Julie George ,Paul Lushenko,Adi Rao,*Kreps, Sarah; George, Julie; Lushenko, Paul; Rao, Adi (18 January 2023). "Exploring the artificial intelligence "Trust paradox": Evidence from a survey experiment in the United States". PLOS ONE. **18** (7) e0288109. journals.plos.org. Bibcode:2023PLoSO..1888109K. doi:10.1371/journal.pone.0288109. PMC 10353804. PMID 37463148.*

13. Bostrom, Nick (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press. ISBN 978-0-19-967811-2.