Could More Human-Like AI Undermine Trust?

By Phanish Puranam , INSEAD, and Bart Vanneste , University College London

Designing modern AI in a way that makes it appear to have more agency could backfire by making people trust the technology less.

From virtual assistants to self-driving cars to medical imaging analysis, the use of modern AI technologies based on deep-learning architectures has increased dramatically in recent years. But what makes us put our faith in such technologies to, say, get us safely to our destination without running off the road? And why do we perceive some forms of AI to be more trustworthy than others?

It may depend on how much agency we believe the technology possesses. Modern AI systems are often perceived as agentic (i.e. displaying the capacity to think, plan and act) to varying degrees. In our research, recently published in ***Academy of Management Review***, we explore how different levels of perceived agency can affect human trust in AI.

Modern AI appears agentic

The perception of agency is crucial for trust. In human relationships, trust is built on the belief that the individual you place your trust in can act independently and **make their own choices**. For instance, if someone is likely to perform an action because they have no choice in the matter, there is no need to trust them. Trust is only relevant if the trustee has some level of agency and can therefore choose between acting in a way that may benefit or harm the trustor.

Modern AI systems can be perceived as possessing greater agency than older, rule-based AI technologies because they rely on a connectionist approach (which involves neural networks that learn from data or interactions with an environment) instead of a symbolic approach (which is based on strict, fixed and pre-defined rules created by human designers).

Given their ability to make decisions more independently of their original programming compared to older AI technologies, modern AI systems exhibit less predictable behaviour, leading to higher perceptions of intentionality and free will. When users don't understand how the AI reaches its conclusions, they are more likely to assume it has agency.

AI agency and trust

Our research adds two new ideas about agency perceptions and trust.

First, when the AI technology appears to have significant levels of agency, trust is mostly about the AI itself. The more animated the puppet, the less noticeable its strings are. But if the AI is perceived to have low agency, trust relies more on the trustworthiness of its designer. This shift occurs because agency perception influences who is seen as responsible for the AI's actions.

Second, humans' natural aversion to betrayal hinders trust. As this is heightened when dealing with entities perceived to possess high agency, the anticipated psychological cost of the AI technology violating trust increases with how agentic it seems to be. If AI with high perceived agency fails or acts against the user's interests, the feeling of betrayal is more significant compared to technology that appears to have less agency.

Implications for AI developers

AI developers draw on several **strategies** to increase human trust in AI. These include enhancing the AI's autonomy (i.e. giving it a greater ability to decide without consulting a human), improving its reliability (i.e. increasing

performance), providing greater transparency (i.e. explaining the logic for its decisions and actions) and anthropomorphising it (i.e. making it more human-like in appearance and capabilities).

However, as outlined in our model, interventions meant to stimulate trust in AI can have the opposite effect. As an example, designers often strive to make AI appear more human-like to increase trust in the technology. They do so by endowing it with characteristics that suggest higher agency, such as giving a name, gender and voice to autonomous vehicles and virtual assistants. Since humans are normally credited with agency, users of these human-like AI systems see them as also having agency. Although this could make the AI appear capable and benevolent, resulting in greater trust, this could be offset or even outweighed by heightened concerns about betrayal.

It is therefore important for AI developers to balance the degree of perceived agency with the intended trust outcomes. An overemphasis on making AI seem human-like can backfire if this is not carefully managed. Communicating the system's capabilities and limitations transparently can help manage user expectations and foster appropriate levels of trust.

It comes down to trust

Human trust in AI is multifaceted and significantly influenced by how agentic the AI is perceived to be. By highlighting the nuanced ways in which agency perceptions shape trust dynamics, we offer a framework that designers and policymakers can use to better understand and predict trust in AI.

Whether or not human trust in AI should be enhanced in the first place, though, is highly context-specific. When there is good reason to believe that scepticism is preventing people from using AI systems that can benefit them, our analysis offers interventions to help overcome this by enhancing or suppressing the perceived agency of the AI to reduce betrayal aversion and promote trust.

In contrast, there are situations where there is a risk of overreliance on AI. This has been documented in **certain medical applications** and is a particular concern given the rapid diffusion of large language models like **ChatGPT** and Gemini. In such cases, designers could use our findings to consciously alter agency perceptions to lower the trust that humans are likely to place in the AI system.

*To find out more about the insights outlined in this article, we invite you to read a **critique** of the research and the **response** by the co-authors.*

Find article at

https://knowledge.insead.edu/strategy/could-more-human-ai-undermine-trust

About the author(s)

Phanish Puranam is a Professor of Strategy and the Roland Berger Chaired Professor of Strategy and Organisation Design at INSEAD. He also directs the **Transforming Your Business with AI** programme.

Bart Vanneste is an Associate Professor in the Strategy & Entrepreneurship group of the UCL School of Management.

About the research

"**Artificial Intelligence, Trust, and Perceptions of Agency**" is published in *Academy of Management Review*.

About the series

AI: Disruption and Adaptation

Delve deeper into developments in artificial intelligence, especially the disruptions across value chains. This series examines AI's impact on a range of sectors, including business consulting, education and the media. It also sizes up the regulatory and ethical questions tied to this game-changing technology.