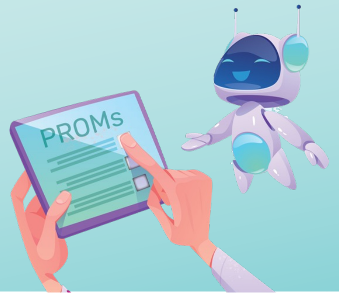
Bundesministerium
für Bildung
und Forschung

Trusting robots & avatars

A model of trust building dynamics with embodied artificial agents



Philipp Graf | Hochschule München
Manuela Marquardt | Charité – Universitätsmedizin Berlin

Overview

Trust as a **social mechanism to reduce complexity** is critical, particularly when using technology in medical or care settings. While according to Luhmann (1979) **personal trust** involves attribution of intended action (and thus also the ability for **contingent action**), **system trust** – specifically in technology use – hinges on trusting **societal norms and technology standards** to ensure functionality (Wagner 1994). Advancements in the field of Large Language Models (Floridi 2023) make contingent communication with AI driven agents a probable fact (Esposito 2017). Across the notion of contingent behavior lies the **question of embodiment** influencing the realm an agent can act in. With respect to sociological conceptions this raises the question **how trust relationships between humans and artificial agents (AAs) can be described?**

Building on latest theoretical approaches from the fields of HRI/HCI, we present a work-in-progress model of trust focusing on implications of embodiment.

The specific characteristics of the **embodiment of artificial agents (AAs)** underpin the Western dichotomy of body and mind (Jackson et al. 2021) as the **distributed nature of hard- and software** enables them to **re- and co-embody** (Luria et al. 2019) diverse devices and bodies. Building on this, Williams et al. (2021) argue and empirically substantiate in their deconstructed trustee theory that at least three **loci of trust** need to be differentiated: “body”, “mind” (and “identity”).

We draw attention to the question how this multi-layeredness influences the dynamic formation of trust or mistrust. Additionally **different frames of assessing** the specifics of an **embodiment** determine, how **risks and gains** are perceived (**physical vs. virtual** (Mutlu 2021)). However, the interacting person's **reference** can **seamlessly shift** between perceiving the AA as a **social entity** and a **mechanical artifact** from one moment to the next (Clark & Fischer 2023) and precisely this oscillation ((Alač 2015) implies a **new quality of trust** that is not adequately captured by either personal or system trust.

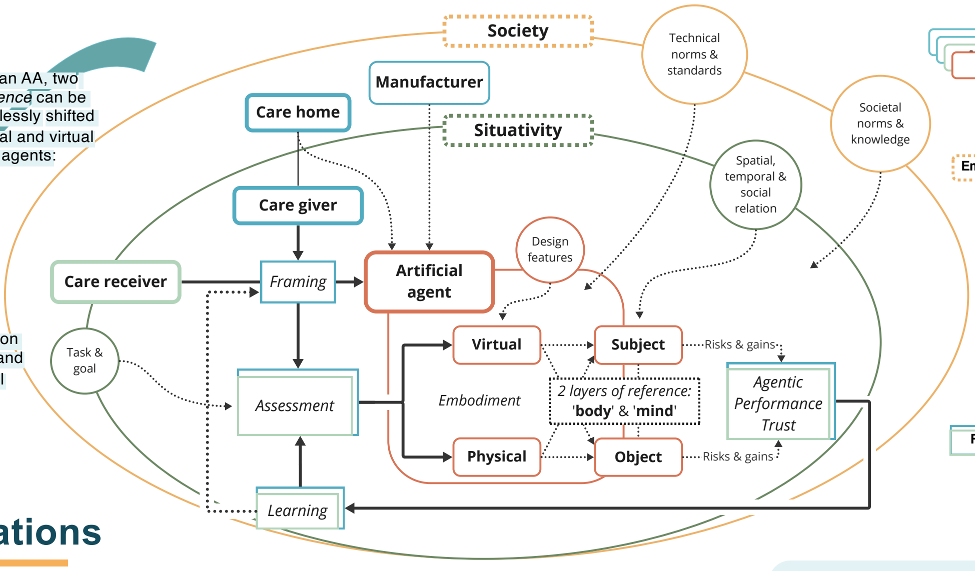
We argue that the new quality of trust relationships **requires more differentiated ways of describing them**, being able to grasp **the intertwined dynamics** between **embodiment** and **ascription of contingent behavior**. We propose to describe this relation as **Agentic Performance Trust (APT), combining characteristics of personal and system trust**.

We propose that for understanding trust formation it is useful to assume **two** context-dependent **reference layers** that can be applied to both physical and virtual embodied agents: The "body" layer defines potential gains and risks based on the material capacities. The "mind" layer is referenced when ascribing contingent behavior, adding further risks and gains, particularly associated with social agency status.

In assessing an AA, two *layers of reference* can be used and seamlessly shifted for both, physical and virtual embodied agents:

A) Body referencing its material and functional relations.

B) Mind referencing the ascription of contingent behavior and implications for social agency status.



Actor … quadrate consisting of care-receiver (CR), care-giver (CG), organization and artificial agent (AA)

Embedded … in a concrete situation with specific … **task & goals**

in which societal & technical … **norms & knowledge**

… also have an impact

Assessment CG and CR assess the AA under a certain situative…

Framing with respect to the task to be fulfilled and the AA's embodiment resp. **design features**

Implications

	Physical artifical agent (PAA)	Virtual artifical agent (VAA)
Embeddedness (body)	(0) Real worldly embeddedness allows acting upon **same affordances** as user. At the same time, it is a danger to trust, as the **sociomaterial environment is uterly complex** and an invitation to fail.	(+) Digitally embeddedness is a resource for trust, as VAA **act within ‚their‘ domain of competences**.
Contingent communication (mind)	(0) For PAAs contingent communication is a double edged sword, as it **has to function coherently with its body** and the environment. A mismatch is thus disadvantageous for trust building.	(+) Contingent communication is the main ressource for VAAs trust building. As the communication is secured inside scripted interfaces, it **can rely on the symbolic sphere and function ‚decoupled‘** from the real world.
Relevance (body / mind)	(–) PAAs can make themselves relevant. This poses them to a higher risk of disappointment as they **may be expected to act pro-actively** in a given situation.	(+) VAAs trust building profits from the fact that they have to be made relevant by the user, as they **can rely just on their reactive functioning**.
Social cues (body / mind)	(+) The use of material social cues (gestures, proxemics) is a ressource for trust as the **use of social cues triggers strong attributions of humanlike characteristics**.	(–) Being bound to graphical or audio social cues limits the capabilites to trigger social attributions.
Risks (body / mind)	(0) **Main risk lies in physical safety** of the user. → PAA needs a high level of trustworthiness regarding perception and body control.	(0) **Main risk lies in data security** of the user. → VAAs trustworthiness relies on its infrastructure respectively the responsible institutions.

Conclusion

A novel quality of trust relations can be observed in interaction with AAs, which requires more differentiated description options. The forms of embodiment have a significant influence on the dynamics of trust building:

- ❖ Virtual embodied agents can be perceived as acting within ‘their domain’ – information processing. The match between digital representation and digital competencies may be advantageous for building trust.
- ❖ On the contrary, for physically embodied agents, real-world embodiment represents the burden of bringing together capacities in both, the analogue and digital domain, while at the same time being an existential threat for humans.

References

Alač, M. (2016). Social robots: Things or agents? AI and Society, 31(4), 519–535.

Clark, H. H., & Fischer, K. (2023). Social robots as depictions of social agents. Behavioral and Brain Sciences, 46.

Esposito, E. (2017). Artificial Communication? The Production of Contingency by Algorithms. *Zeitschrift für Soziologie*, *46*(4), 249–265.

Floridi, L. (2023). AI as Agency Without Intelligence: on ChatGPT, Large Language Models, and Other Generative Models. Philosophy and Technology, 36(1), 1–7.

Jackson, R. B., Bejarano, A., Winkle, K., & Williams, T. (2021). Design, Performance, and Perception of Robot Identity. Proceedings of the Workshop on Robo-Identity: Artificial Identity and Multi-Embodiment at HRI.

Luhmann, N. (1979). Trust and Power. John Wiley & Sons.

Mutlu, B. (2021). The virtual and the physical: two frames of mind. IScience, 24(2),

Wagner, G. (1994). Vertrauen in Technik. Zeitschrift für Soziologie, 23(2), 145–157.

Williams, T., Ayers, D., Kaufman, C., Serrano, J., & Roy, S. (2021). Deconstructed trustee theory: Disentangling trust in body and identity in multi-robot distributed systems. ACM/IEEE International Conference on Human-Robot Interaction, 262–271.

CHARITÉ UNIVERSITÄTSMEDIZIN BERLIN | HM Hochschule München University of Applied Sciences | HOCHSCHULE HAMM-LIPPSTADT | TECHNISCHE UNIVERSITÄT BERLIN | MIA PROM | dexter voice of healthcare | Deutsche Rentenversicherung Reha-Zentrum Seehof | ZAR | ACALTA