DO YOU TRUST ME? – DEVELOPMENT OF A FRAMEWORK TO ANALYZE THE ROLE AND MEANING OF TRUSTWORTHINESS OF AVATARS

TREO Paper

Julia Bräker, University of Hamburg, Hamburg, Germany, julia.braeker@uni-hamburg.de

Sofia Schöbel, University of Osnabrück, Osnabrück, Germany, sofia.schoebel@uni-osnabrueck.de

Martin Semmann, University of Hamburg, Hamburg, Germany, martin.semmann@uni-hamburg.de

1 Motivation

Virtual worlds are becoming more and more present in our private and working lives. At the same time, the importance and relevance of artificial intelligence is constantly increasing. As soon as users are in a virtual space - be it in a Teams meeting or on a gaming platform - they are represented as a virtual image (Mohler et al. 2010). If users take part in a Teams meeting, for example, they can be seen by other participants, or they can be represented as an avatar because they are not feeling well on the day of the meeting (Panda et al. 2022). With the growing presence of generative artificial intelligence (AI), users can represent themselves online in various ways. Users can represent themselves as a completely different person or even appear as fictional characters in a virtual world (Batinovic et al. 2024). In addition, we are seeing more and more avatars appear online that are completely designed by artificial intelligence and are not human (Batinovic et al. 2024). These avatars often act without human intervention. Most recently, for example, the German Chancellor Olaf Scholz was created as a digital image. This image was visually and acoustically able to completely replace the real person.

These developments raise questions about trust in virtual presences - otherwise known as avatars. Research in the field of trust in AI-generated avatars is still in its early stages (Weisman and Pena 2021). In particular, the depth of trust in different types of avatars has not yet been explored in more detail. In general, the question arises in which situations and with which type of appearance users trust avatars. With generative artificial intelligence, for example, it is even possible to imitate the voices and complete appearance of other people so that the other person can no longer be sure whether and with whom they are interacting (Qui et al. 2005). In summary, there is a lack of design implications in research as to how avatars can be used with confidence and under what conditions. The project presented below, therefore, addresses the following research question:

RQ: What is the role and meaning of different kinds of avatars in digital, virtual environments?

Once our work is completed, we contribute to theory and practice in different ways. First, we provide more details on the role and meaning of avatars in digital virtual worlds and can describe the circumstances under which they are experienced as trustworthy. Second, we can provide insights on how users experience different types of avatar shapes. We provide practical implications to organizations that want to use avatars, whether it is for the design of hybrid meetings or the representation of their employees in virtual worlds.

2 Dimensions

When considering trust regarding virtual avatars, we can take different perspectives. In the following, we will focus on the user's perspective, specifically how a user perceives the trustworthiness of other avatars. Avatars can be classified according to two dimensions, and their trustworthiness can be assessed based on these characteristics.

Avatar Representation:

The representation of an avatar in virtual worlds can be (1) exclusively *human* nature, (2) *hybrid*, or (3) *artificial*. This applies to both their outer appearance and their voice. A solely human avatar is defined as the representation of the user in its natural form with their real body and voice. A hybrid representation includes any kind of virtually modified or extended appearance of the user. An artificially generated avatar is rendered completely digital and replaces the user entirely. The avatar representation relates to the technical rendering, i.e., what I see is the reality or it is computer-generated.

Realistic Appearance:

Avatars can have varying degrees of realism, being (a) *realistic/natural*, (b) *partially realistic/mixed*, or (c) *unrealistic/different shape*. The degree of realism also applies to the outer appearance and voice of an avatar. A realistic and natural-looking avatar looks like a real person and has a natural voice (I look and sound like I do in reality). A 3D scan of the body for virtual worlds that creates a realistic representation would be a realistic avatar in this case. An avatar with a mixed degree of realism is modified or extended in an unrealistic way, but the natural appearance is still visible (I look and sound like I do in reality, but parts of the "real" me are changed, e.g., by visual filters). An unrealistic avatar has a completely different shape or voice and looks or sounds unrealistic (I do not appear like a natural person/human being). The avatar can no longer be traced back to a natural person.

Degree of Trustworthiness:

The perception of trustworthiness is highly subjective and individual. Trustworthiness is not a binary measure (I trust or I do not trust) but instead varies along a continuum. From the user's point of view, trustworthiness can be examined from two perspectives (user roles). On the one hand, the question arises as to whether the user trusts other avatars regarding their identity, i.e., is the person behind the avatar really who they claim to be. The other perspective looks at the extent to which other people trust the user regarding their identity, i.e., do other people trust my avatar and me.

3 Exemplary Cases to Evaluate the Role of Trust in Avatars

We present two exemplary cases that can be explained by our framework.

Case 1 – Deep fake video of Olaf Scholz: the deep fake video of Olaf Scholz is an initiative by a Berlin-based activist group called the Centre for Political Beauty (ZPS). The video shows Scholz pretending to announce a ban on the far-right Alternative for Germany (AfD) party. The video was created using artificial intelligence to simulate Scholz's face and voice. This case shows that we can use artificial intelligence to create realistic representations of individuals, even imitating someone else's voice. This case demonstrates an AI-generated avatar ((3) artificial avatar representation) with a realistic and natural degree of realism ((a) realistic/natural appearance), raising the question of how trust constitutes from different perspectives.

Case 2 – Faking multiple persons in video calls: after falling victim to a deepfake scam, a Hong Kong-based finance employee has been swindled out of $25 million. Following a video call from his "chief financial officer and other staff", the employee - who works for a multinational company - was persuaded to part with his funds. After receiving an email requesting a secret transaction, the employee was suspicious. However, after being convinced of the authenticity of the group video call, he ended up sending the scammers a total of 200 million Hong Kong dollars. In the multiperson videoconference, everyone he saw turned out to be fake. This case demonstrates that even with a natural and realistic AI

representation of more than one person, individuals can end up trusting them so much that they are willing to pay money to others.

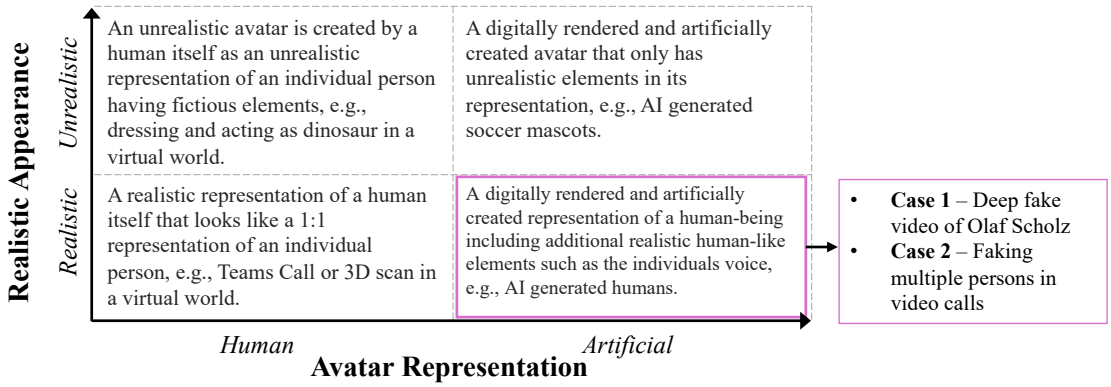

Figure 1. Classification of the cases based on avatar representation and realistic appearance dimensions.

4 Next Steps and Expected Contributions

With our framework, we want to explore different cases in more detail to better understand how trust is constituted in virtual worlds and in relation to different types and kinds of avatars. Once our research is completed, we will categorize different types of cases into clusters by using our framework. To better explore the nature of trust, we plan to run think-aloud studies. In our think aloud studies, users are going to interact with different types of avatars and are requested to report their feelings towards the shown avatars. This approach will be used to evaluate our framework. We are also going to consider an extension of our framework for situations where the represented avatar is not human at all and is completely generated and controlled by an AI. These types of avatars are better known as virtual influencers (Batinovic et al. 2024). Our completed research project contributes to theory and practice. From a theoretical perspective, we can better explain the nature of trust towards different types of avatars. Like the uncanny valley, we think that with different cases involving different types of avatars, we are better able to demonstrate how trust constitutes. This also provides more information on how to design avatars in different virtual environments. From a practical perspective, we can assist companies in using avatars in different types of virtual environments, such as hybrid meetings, social media, or even in virtual worlds.

References

Batinovic, H., Tingelhoff, F., Hammerschmidt, M., & Schöbel, S. (2024). Come Closer, but Not Too Close: How Virtual Influencers Can Facilitate or Restrict Brand Experiences in the Metaverse.

Mohler, B. J., Creem-Regehr, S. H., Thompson, W. B., & Bülthoff, H. H. (2010). The effect of viewing a self-avatar on distance judgments in an HMD-based virtual environment. Presence, 19(3), 230-242.

Panda, P., Nicholas, M. J., Gonzalez-Franco, M., Inkpen, K., Ofek, E., Cutler, R., ... & Lanier, J. (2022, June). AllTogether: Effect of Avatars in Mixed-Modality Conferencing Environments. In 2022 Symposium on Human-Computer Interaction for Work(pp. 1-10).

Qiu, Lingyun, and Izak Benbasat. "Online consumer trust and live help interfaces: The effects of text-to-speech voice and three-dimensional avatars." International journal of human-computer interaction 19.1 (2005): 75-94.

Weisman, W. D., & Peña, J. F. (2021). Face the uncanny: the effects of doppelganger talking head avatars on affect-based trust toward artificial intelligence technology are mediated by uncanny valley perceptions. Cyberpsychology, behavior, and social networking, 24(3), 182-187.