12-2017

# When Do We Need a Human? Anthropomorphic Design and Trustworthiness of Conversational Agents

Anna-Maria Seeger
*University of Mannheim*, seeger@uni-mannheim.de

Jella Pfeiffer
*Karlsruhe Institute of Technology*, jella.pfeiffer@kit.edu

Armin Heinzl
*University of Mannheim*, heinzl@uni-mannheim.de

# When Do We Need a Human? Anthropomorphic Design and Trustworthiness of Conversational Agents

**Anna-Maria Seeger**
University of Mannheim
seeger@uni-mannheim.de

**Jella Pfeiffer**
Karlsruhe Institute of Technology
jella.pfeiffer@kit.edu

**Armin Heinzl**
University of Mannheim
heinzl@uni-mannheim.de

## ABSTRACT

Conversational agents interact with users via the most natural interface: human language. A prerequisite for their successful diffusion across use cases is user trust. Following extant research, it is reasonable to assume that increasing the human-likeness of conversational agents represents an effective trust-inducing design strategy. The present article challenges this assumption by considering an opposing theoretical perspective on the human-agent trust-relationship. Based on an extensive review of the two conflicting theoretical positions and related empirical findings, we posit that the agent substitution type (human-like vs. computer-like) represents a situational determinant on the trust-inducing effect of anthropomorphic design. We hypothesize that this is caused by user expectations and beliefs. A multi-method approach is proposed to validate our research model and to understand the cognitive processes triggered by anthropomorphic cues in varying situations. By explaining the identified theoretical contradiction and providing design suggestions, we derive meaningful insights for both researchers and practitioners.

### Keywords

Conversational Agents, Chatbots, Trust, Anthropomorphism.

## INTRODUCTION

The objective to make interactions with computer agents as natural as face-to-face interactions has inspired a vast body of research in human-computer interaction (HCI). A conversational agent (CA) is a software system that interacts with users in human language (Nunamaker et al. 2011). Based on technical advances in natural language processing, CAs – or chatbots – are now being employed in various domains of application (e.g. customer service bots, enterprise system bots). Successful CA diffusion across contexts, however, can only be realized if designers of CAs understand and consider users' expectations and beliefs to ensure that they trust these agents. From extant information systems (IS) research we know that trust is a central antecedent of technology acceptance and use (e.g. Gefen 2000; Komiak and Benbasat 2006; Wang and Benbasat 2008).

Studies interested in understanding the psychological mechanisms that explain human-agent trust adopt two opposing theoretical perspectives: the human-human and the human-machine trust perspective (Madhavan and Wiegmann 2007). The Computers are Social Actors (CASA) paradigm (Nass et al. 1994) reflects the human-human trust perspective. Research in this tradition assumes that humans place social expectations, norms and beliefs towards computers. This paradigm represents a well-known conceptual basis for IS research interested in understanding how to make computer agents more trustworthy (e.g. Qiu and Benbasat 2009; Riedl et al. 2014; Wang and Benbasat 2008). The human-machine trust literature challenges this perspective by arguing that humans hold other expectations towards computer system than towards humans such as efficiency and rationality (Dzindolet et al. 2003; Skitka et al. 1999). Research adopting this perspective argues that humans trust computer systems more than other humans and explain this phenomenon with the automation bias – humans' tendency to trust automated or computer systems (Mosier and Skitka 1996).

The two positions provide contradictory predictions regarding the effect of anthropomorphic design on CAs' trustworthiness. A distinct characteristic of CAs is their ability to interact with users based on natural language. Because natural language originates from human communication, a certain degree of human likeness is immanent to CAs. Therefore, it seems intuitive to adopt the human-human trust perspective and conclude that making these agents even more human-like should be a design objective. However, this approach would not be in line with the human-machine trust literature. While some researchers have investigated the difference between these two theoretical perspectives (e.g. Madhavan and Wiegmann 2007), the puzzling question about the effect of anthropomorphic design on users' trust has not been addressed. The objective of the present research is to address this research gap. We posit that CAs that perform human tasks (human substitute) benefit from anthropomorphic design in terms of trust while the

opposite is true for agents that perform computational tasks (system substitute). To investigate these situational effects, we examine the following research question:

1. Can the agent substitution type explain the contradicting findings about the trust-inducing effect of anthropomorphic design?

To address these questions, we develop a research model based on psychological theory on anthropomorphism and extant research in the domain of trust into technology. We propose a multi-method approach that allows us to validate the developed model and to gain a deeper understanding of the cognitive processes related to the evaluation of anthropomorphic agent design.

## THEORETICAL FOUNDATION

### Multi-Dimensional User Trust

Across disciplinary boundaries, human *trust* is defined as "a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another" (Rousseau et al. 1998, p. 395). Users' initial trust towards a CA is determined by their *trusting beliefs* about the agents' perceived level of *competence*, *benevolence* and *integrity* (Mayer et al. 1995; McKnight et al. 2002). We posit to further distinguish between goodwill- and qualification-based trustworthiness. This distinction is relevant for the present research because benevolence and integrity reflect a CA's perceived intentions and motives to perform the expected behavior. While computers do not have intentions and thoughts, anthropomorphizing non-human objects signifies assigning them such human characteristics. Therefore, it is important to contrast the volitional dimensions of integrity and benevolence (goodwill) from the non-volitional dimension competence (qualification).

### Anthropomorphism and Conversational Agents

*Anthropomorphism* refers to the phenomenon of assigning human-like attributes to non-human agents (Epley et al. 2007). Psychological research has identified two relevant motivational forces that explain why humans respond to non-human agents with anthropomorphism. First, anthropomorphizing non-human agents responds to humans' basic need to be socially related to other humans. Second, anthropomorphizing non-human agents responds to humans' basic need to understand and control the environment. (Epley et al. 2007).

*CAs* enable users to interact with computer systems in human language. This natural language interface represents a distinct characteristic of CAs. Because human language is the most natural way of communication, ideal CAs are considered to provide the most intuitive user interface (Cassell et al. 1999). HCI research interested in CAs envisions the ideal CA to be represented by a virtual human, and thus has been especially interested in anthropomorphic design (Pickard et al. 2017). Several experimental studies indicate that human-likeness influence user beliefs and emotions towards a CA (e.g. Nass et al. 1999; Nunamaker et al. 2011). Overall, research on CAs assumes human-likeness to be the ultimate design goal. The review of extant research in this domain reveals a concentration on CAs that act as a substitute of a human expert such as sales assistants (e.g Qiu and Benbasat 2009), interviewers (e.g. Nunamaker et al. 2011; Pickard et al. 2017) and tutors (Nass et al. 1999).

However, real world CAs are also used in situations where they do not replace a human, but provide intuitive and user-friendly interfaces to computer systems. Enterprise productivity bots (e.g. *Amazon Lex*), internet of things (IoT) or smart home bots (e.g. *action.ai*) are examples for this domain. For instance, professionals can use natural language chat to query data from an enterprise database. To differentiate the two types of agents, we define human substitute agents as CAs that replace a human expert and we define system substitute agents as CAs that provide a natural language interface to computer systems. We posit that research findings considering human substitute agents cannot be readily applied to system substitute agents, because of differing user-expectations about interactions with these systems.

### Human-Human Trust Perspective in HCI

Research on trust in HCI frequently adopts the CASA paradigm as theoretical foundation. Central to this paradigm is the media equation hypothesis that argues that humans show the same social responses to computer systems as to other humans (Nass et al. 1994). Experimental studies adopting this perspective provide evidence that anthropomorphic design is positively related to computer agents' trustworthiness (Cassell and Bickmore 2000). The media-equation hypothesis and the related body of experimental evidence provides the justification for IS research to adopt a human-human trust conceptualization to investigate human-agent trust (e.g. Qiu and Benbasat 2009; Wang and Benbasat 2008). The link between anthropomorphic design and trustworthiness of computer agents is further supported by the introduced psychological theory on anthropomorphism. The two motivational forces that drive humans' tendency to anthropomorphize novel non-human agents correspond to two well-investigated antecedents of trust in HCI: social presence and familiarity (e.g. Gefen and Straub 2004; Komiak and Benbasat 2006). Social presence refers to the feeling of being connected to other human beings (Gefen and Straub 2004), and thus directly corresponds to humans' need to be socially related. Familiarity, on the other hands, refers to knowledge and understanding about an interaction partner (Gefen 2000), and thus directly corresponds to humans' need to understand and control the environment. Consequently, the link between anthropomorphic design and trustworthiness is also supported by IS studies on social presence, familiarity and

trust.

In sum, the human-human trust perspective assumes that anthropomorphism through increased feelings of social relatedness and familiarity is positively related to users' trust perceptions.

### Human-Machine Trust Perspective in HCI

In contrast to the human-human trust perspective, automation bias literature proposes a contradicting relationship between anthropomorphism and trust (Dzindolet et al. 2003). While humans are expected to be imperfect, the opposite is true for automation, and thus trust into computational systems is higher than trust into another human. This stream of literature suggests that humans generally think that the programmed technical abilities of a computer agent are superior in terms of rationality, reliability and objectivity (Mosier and Skitka 1996). This is reflected in the authority hypothesis that proposes that humans are responding to computational systems by perceiving them as better skilled than humans to perform specific tasks (Skitka et al. 1999). A series of experiments supports the authority hypothesis. For example, Dijkstra et al. (1998) conducted an experimental study to evaluate the persuasiveness of expert systems and found that users perceive expert systems to be more rational and objective than humans. Similarly, in a series of experiments Dzindolet et al. (2003) found that users hold higher than average initial trusting beliefs towards an automated decision aid. In line with these findings, a recent experimental study also found that initial trust towards a machine-like computer agent was higher than initial trust towards a human-like computer agent (de Visser et al. 2016).

In sum, the automation bias literature assumes that humans' trust into automation is influenced by their beliefs about computer agents' superior expertise. According to this perspective, anthropomorphized CAs can be detrimental to agents' trustworthiness because they may signal human imperfection instead of algorithmic precision.

## RESEARCH MODEL

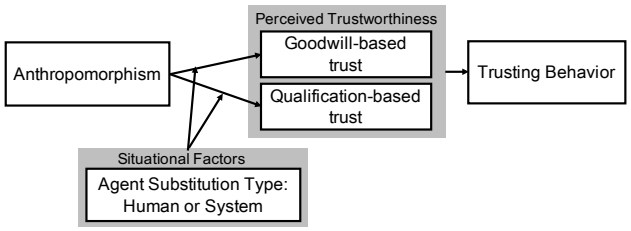


**Figure 1. Proposed Research Model**

In the present study, we seek to investigate the situational factors that influence the positive relationship between anthropomorphic design and trustworthiness of CAs. Thereby, we attempt to resolve the contradicting predictions of the human-human and the human-machine trust perspective. The basic structure of our research model corresponds to Gefen's trust model (Gefen 2000; Gefen and Straub 2004). We provide an overview of our model in figure 1.

Humans anthropomorphize novel non-human objects in order to increase feelings of familiarity and social relatedness (Epley et al. 2007). In accordance with psychological theory and studies adopting the CASA perspective in HCI, we expect that anthropomorphism is positively related to users' trust perceptions. On the one hand, anthropomorphism helps to reduce uncertainty about the potential behavior of a new interaction partner (Epley et al. 2007). This increased level of familiarity is known to be positively related to trusting beliefs about the goodwill of an interaction partner (Gefen 2000). In addition, it can be expected that the qualification-based trustworthiness benefits from familiarity because a CA for a specific task (e.g. tutor, doctor, expert) raises positive beliefs induced by the knowledge about the expertise of a familiar human equivalent (Komiak and Benbasat 2006). On the other hand, anthropomorphism increases feelings of social relatedness (Epley et al. 2007). Again, it is well established that social presence is positively related to trustworthiness in HCI (Gefen and Straub 2004). By increasing feelings of social relatedness an anthropomorphized CA signals "warmth" and empathy (Qiu and Benbasat 2009). These characteristics are closely related to goodwill-based trustworthiness. Based on the discussed effects of anthropomorphism, we hypothesize:

**H1**: The higher the anthropomorphism, the higher the trusting belief into the CA's goodwill.

**H2**: The higher the anthropomorphism, the higher the trusting belief into the CA's qualification.

As detailed in the previous chapter, extant research on trust into technology provides two conflicting predictions about the effect of anthropomorphism on user trust. In the present paper, we propose that in the context of CAs both perspectives are valid and that the agent substitution type acts as a situational moderator variable on the relationship between anthropomorphism and trustworthiness. We posit to differentiate between the agent as human substitute and system substitute. We expect that, in accordance with CASA, agents as human substitutes in contrast to system substitutes benefit from increased anthropomorphism in terms of trust. We theorize that different expectations are triggered by the substitution type.

CAs of the human substitution type are programmed to perform a task typically performed by a human expert. In such instantiations, users are formerly used to interact with a human interaction partner, and thus hold expectations about the human-likeness of the novel interaction partner. CAs that act as sales assistants, for

instance, are examples for the human substitution type. We expect that the human substitution type will enhance the positive relationship between anthropomorphic design and trustworthiness because of humans' need for social relatedness and their desire to decrease uncertainty. Anthropomorphism assigns human characteristics including emotions and intentions to non-human agents (Epley et al. 2007). This is beneficial for human substitute CAs because in their role they need to meet not only qualification-related but also goodwill-related expectations. Because the type of tasks performed by a human substitution type raise social expectations, a CA that can respond to these expectations is hypothesized to be more trustworthy.

**H3a**: A CA of human substitution type positively moderates (reinforces) the positive relationship between anthropomorphism and qualification-based trustworthiness

**H4a**: A CA of human substitution type positively moderates (reinforces) the positive relationship between anthropomorphism and goodwill-based trustworthiness

CAs in form of a system substitution type are implemented to provide a more efficient interface to computational systems. Examples for such technologies include chatbots that allow users to interact with enterprise software systems. We expect that this substitution type does not benefit from anthropomorphic design. Because users expect a system substitute agent to be a rational and efficient "machine", cues of human-likeness undermine these relevant characteristics and provide conflicting information. Because the highest expertise and efficiency for computational tasks is assigned to computer systems, inducing perceptions of human-likeness can have negative impact on the qualification-assessment of the agent. Moreover, unexpected human-likeness of an agent can raise users' doubts about the underlying design intentions. Because cues of human-likeness provide conflicting information when assessing system substitute agent, we hypothesize:

**H3b**: A CA of system substitution type negatively moderates (attenuates) the positive relationship between anthropomorphism and qualification-based trustworthiness.

**H4b**: A CA of system substitution type negatively moderates (attenuates) the positive relationship between anthropomorphism and goodwill-based trustworthiness.

The connection between users' trustworthiness perceptions and trusting behavior has been widely discussed in HCI (e.g. Benbasat and Wang 2005; Gefen 2000; Qiu and Benbasat 2009). Accordingly, perceptions of trustworthiness are important in decision-making processes. Therefore, we hypothesize that the trustworthiness will determine the selection of a CA.

**H5**: Trusting Behavior: Users select the CA associated with the higher degree of trustworthiness.

## PROPOSED EXPERIMENTAL DESIGN

We plan to use a multi-method approach that allows to test our research model and to explore the cognitive processes related to the assessment of conversational agents. Therefore, self-reported, behavioral and eye-tracking data will be collected that complement each other: self-reported data will be used to test the hypothesized relationships between anthropomorphism and trustworthiness, the effect on trustworthy behavior will be assessed with behavioral data about the user's choice, and eye-tracking data will allow us to gain deeper understanding of the cognitive processes triggered by the use of anthropomorphism in different substitution type conditions.

To test our research model, we propose a 2 x 2 within-subjects design with two levels of anthropomorphism (low vs. high) and two agent substitution types (human vs. system). Participants receive a task scenario. In the scenario, each participant represents an employee in an enterprise. Their manager is enthusiastic to introduce the latest chatbot technology to streamline business processes. Therefore, she asks the employee to evaluate and decide which customer service chatbot should be introduced as a touching point for customers (human substitute) and which enterprise chatbot should be introduced inside the company to enable efficient interactions for sales personal with the customer database (system substitute). For each decision task the employee is presented with two chatbot options (anthropomorphism: high vs. low). We manipulate anthropomorphism. Our stimulus material is adapted from previous studies on CAs (Cassell and Bickmore 2000). We will measure trustworthiness by using established self-rating scales (McKnight and Choudhury 2002). The choice decision between the offered CAs represents the behavioral trust measure. Eye-tracking will be used to capture participants' eye movement and fixations. Eye movement and fixation data allows to infer cognitive processes related to visual stimuli (Just and Carpenter 1976). Thus, eye-tracking allows to measure the extent to which anthropomorphic cues might distract the user from task execution

## CONCLUSION AND EXPECTED CONTIRBUTIONS

We identify two conflicting theoretical perspectives on the relationship between anthropomorphic design and users' trust into CAs. We theorize that the agent substitution type plays a moderating role in the relationship between anthropomorphic design and agents' trustworthiness. In doing so, this research demonstrates that by considering situational boundaries the two opposing theoretical streams become compatible. Moreover, a multi-method empirical approach is proposed to validate the developed research model. Because the situational limitation has not been considered in previous studies on CA's design, this research will enhance HCI knowledge. Moreover, this

study will inform practitioners who seek to leverage chatbots in various usage contexts.

## REFERENCES


Benbasat, I., and Wang, W. 2005. "Trust in and adoption of online recommendation agents," *Journal of the Association for Information Systems,* (6:3), pp. 72–101.

Cassell, J., and Bickmore, T. 2000. "External manifestations of trustworthiness in the interface," *Communications of the ACM* (43:12), pp. 50–56.

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., and Yan, H. 1999. "Embodiment in conversational interfaces: Rea," CHI Pittsburgh PA, USA, pp. 520–527.

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., and Parasuraman, R. 2016. "Almost human: Anthropomorphism increases trust resilience in cognitive agents.," *Journal of Experimental Psychology: Applied* (22:3), pp. 331–349.

Dijkstra, J. J. 1999. "User agreement with incorrect expert system advice," *Behaviour & Information Technology* (18:6), pp. 399–411.

Dijkstra, J. J., Liebrand, W. B. G., and Timminga, E. 1998. "Persuasiveness of expert systems," *Behaviour & Information Technology* (17:3), pp. 155–163.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., and Beck, H. P. 2003. "The role of trust in automation reliance," *International Journal of Human-Computer Studies* (58:6), pp. 697–718.

Epley, N., Waytz, A., and Cacioppo, J. T. 2007. "On seeing human: A three-factor theory of anthropomorphism.," *Psychological Review* (114:4), pp. 864–886.

Gefen, D. 2000. "E-commerce: the role of familiarity and trust," *Omega* (28:6), pp. 725–737.

Gefen, D., and Straub, D. W. 2004. "Consumer trust in B2C e-Commerce and the importance of social presence: experiments in e-Products and e-Services," *Omega* (32:6), pp. 407–424.

Just, M. A., and Carpenter, P. A. 1976. "Eye fixations and cognitive processes," *Cognitive psychology* (8:4), pp. 441–480.

Komiak, S., and Benbasat, I. 2006. "The effects of personalization and familiarity on trust and adoption of recommendation agents," *MIS Quarterly,* (30:4), pp. 941–960.

Madhavan, P., and Wiegmann, D. A. 2007. "Similarities and differences between human–human and human–automation trust: an integrative review," *Theoretical Issues in Ergonomics Science* (8:4), pp. 277–301.

Mayer, R. C., Davis, J. H., and Schoorman, F. D. 1995. "An Integrative Model of Organizational Trust," *The Academy of Management Review* (20:3), pp. 709–734.

McKnight, D. H., and Choudhury, V. 2002. "Developing and Validating Trust Measures for e-Commerce: An Integrative Typology," *Information Systems Research* (13:3), pp. 334–359.

Mosier, K. L., and Skitka, L. J. 1996. "Human decision makers and automated decision aids: Made for each other, " *Automation and Human Performance: Theory and Applications*, pp. 201–220.

Mosier, K. L., Skitka, L. J., Heers, S., and Burdick, M. 1998. "Automation Bias: Decision Making and Performance in High-Tech Cockpits," *The International Journal of Aviation Psychology* (8:1), pp. 47–63.

Nass, C., Moon, Y., and Carney, P. 1999. "Are People Polite to Computers? Responses to Computer-Based Interviewing Systems," *Journal of Applied Social Psychology* (29:5), pp. 1093–1109.

Nass, C., Steuer, J., and Tauber, E. R. 1994. "Computers are Social Actors," Proceedings of the SIGCHI conference on Human factors in computing systems., pp. 72–78

Nunamaker, J. F., Jr., Derrick, D. C., Elkins, A. C., Burgoon, J. K., and Patton, M. W. 2011. "Embodied Conversational Agent-Based Kiosk for Automated Interviewing," *Journal of Management Information Systems* (28:1), pp. 17–48.

Pickard, M., Schuetzler, R., Valacich, J., and Wood, D. A. 2017. "Next-Generation Accounting Interviewing: A Comparison of Human and Embodied Conversational Agents as Interviewers."

Qiu, L., and Benbasat, I. 2009. "Evaluating Anthropomorphic Product Recommendation Agents: A Social Relationship Perspective to Designing Information Systems," *Journal of Management Information Systems* (25:4), pp. 145–182.

Riedl, R., Mohr, P. N. C., Kenning, P. H., Davis, F. D., and Heekeren, H. R. 2014. "Trusting Humans and Avatars: A Brain Imaging Study Based on Evolution Theory," *Journal of Management Information Systems* (30:4), pp. 83–114.

Rousseau, D. M., Sitkin, S. B., and Burt, R. S. 1998. "Not so different after all: A cross-discipline view of trust," *Academy of Management Review* (23:3), pp. 393–404.

Skitka, L. J., Mosier, K. L., and Burdick, M. 1999. "Does automation bias decision-making?," *International Journal of Human-Computer Studies* (51:5), pp. 991–1006.

Wang, W., and Benbasat, I. 2008. "Attributions of Trust in Decision Support Technologies: A Study of Recommendation Agents for E-Commerce," *Journal*

*of Management Information Systems* (24:4),pp. 249–273.