“My Name is Alexa. What’s Your Name?” The Impact of Reciprocal Self-Disclosure on Post-Interaction Trust in Conversational Agents

Kambiz Saffarizadeh
College of Business,
University of Texas at Arlington, USA
kambiz.saffari@uta.edu

Mark Keil
J. Mack Robinson College of Business,
Georgia State University, USA
mkeil@gsu.edu

Maheshwar Boodraj
College of Business and Economics
Boise State University, USA
mboodraj@boisestate.edu

Tawfiq Alashoor
IESE Business School
University of Navarra, Spain
talashoor@iese.edu

About Authors

Kambiz Saffarizadeh is an Assistant Professor of Information Systems and Analytics in the College of Business Administration at Marquette University. His research focuses on the implications of artificial intelligence for individuals, teams, organizations, and society. He has published his research in the *International Conference on Information Systems*. He co-founded the Iranian Scholars Chapter of the AIS and developed several courses on emerging technologies, machine learning, and mobile application development. He holds a Ph.D. from Georgia State University. Before joining academia, he spent several years in the IT industry, serving in software development and managerial roles.

Mark Keil is a Regents’ Professor of the University System of Georgia and the John B. Zellars Professor of Computer Information Systems in the J. Mack Robinson College of Business at Georgia State University. He has published over 130 peer-reviewed journal articles and his work has been cited more than 25,000 times according to Google Scholar. He holds B.S.E., S.M., and D.B.A. degrees from Princeton University, M.I.T. Sloan School, and Harvard Business School respectively.

Maheshwar Boodraj is an Assistant Professor of Information Technology Management in the College of Business and Economics at Boise State University. His research focuses on IT project management, cybersecurity, and software development. He has published several articles in peer-reviewed outlets, such as the *Data Base for Advances in Information Systems* and the *International Conference on Information Systems*. He is a *Distinguished Member* of the AIS and a co-founder of the AIS Doctoral Student College. He holds a Ph.D. from Georgia State University, and is a *Certified Information Systems Security Professional (CISSP)* and *Project Management Professional (PMP)*. Before joining academia, he spent twelve years in various managerial and technical roles at the Mona School of Business and Management and IBM.

Tawfiq Alashoor is an Assistant Professor of Operations, Information, and Technology at IESE Business School, University of Navarra. His research, largely centered around privacy decision-making and cybersecurity, has been recognized in premier IS journals like *Information Systems Research* and conferences including the *International Conference on Information Systems*. His teaching interests include managerial cybersecurity, systems analysis and design, and other IS areas. He is a *Distinguished Member* of the AIS, a co-founder of the SaudiAIS Chapter, and holds other active roles in the IS community. He holds B.S., M.S., and Ph.D. degrees from King Fahd University of Petroleum and Minerals, Penn State University, and Georgia State University respectively.

Abstract

The use of conversational AI agents (CAs), such as Alexa and Siri, has steadily increased over the past several years. However, the functionality of these agents relies on the personal data obtained from their users. While evidence suggests that user disclosure can be increased through reciprocal self-disclosure (i.e., a process in which a CA discloses information about itself with the expectation that the user would reciprocate by disclosing similar information about themself), it is not clear whether and through which mechanism the process of reciprocal self-disclosure influences users' post-interaction trust. We theorize that anthropomorphism (i.e., the extent to which a user attributes humanlike attributes to a CA) serves as an inductive inference mechanism for understanding reciprocal self-disclosure, enabling users to build conceptually distinct cognitive and affective foundations upon which to form their post-interaction trust. We find strong support for our theory through two randomized experiments that used custom-developed text-based and voice-based CAs. Specifically, we find that reciprocal self-disclosure increases anthropomorphism and anthropomorphism increases cognition-based trustworthiness and affect-based trustworthiness. Our results show that reciprocal self-disclosure has an indirect effect on cognition-based trustworthiness and affect-based trustworthiness which is fully mediated through anthropomorphism. These findings conceptually bridge prior research on motivations of anthropomorphism and research on cognitive and affective bases of trust.

Keywords: Conversational AI, AI Agent; Chatbot; Cognition-Based Trust; Affect-Based Trust; Anthropomorphism; Reciprocal Self-Disclosure.

1. Introduction

In recent years, implementation of conversational AI agents (CAs) has been on the rise across different task domains such as daily personal tasks (Dietvorst & Bharti, 2020), customer service (Schanke et al., 2021), office assistance, and medical diagnosis (Longoni et al., 2019). About 4.2 billion CAs were in use in 2020 across various platforms, and this number is expected to rise to 8.4 billion by 2024 (Statista, 2021). The role of CAs is expected to become even more salient in people's daily lives with companies' recent attempts to utilize CAs in emerging domains such as generative AI (e.g. ChatGPT) (OpenAI, 2022), web copilots (Microsoft, 2023), and the metaverse (VentureBeat, 2022). Industry experts predict an exponential growth of CAs due to "the GPT effect" (i.e., the availability of large language models, which can be fine-tuned to create new CAs) (destinationCRM, 2022; Sundar, 2023). The ultimate utility of CAs, however, depends on personal data provided by human users. The personal data feed into the CAs' algorithms for processing to understand the context of requests, improve the relevance and accuracy of responses, and learn individual and aggregate user preferences (Apple, 2021; Google, 2021). These improvements, which are based on the insights generated by the CAs' algorithms, cannot be realized without personal data provided by users (Saffarizadeh et al., 2017). Therefore, companies that make CAs have been implementing different methods to acquire the needed data.

A major data acquisition method that has received attention in both academia and industry is reciprocal self-disclosure in which a CA discloses some information about itself and the user reciprocates by disclosing similar information about themself (Archer & Berg, 1978; Moon, 2000; Sprecher et al., 2013). For instance, two conversational AI agents—SlugBot and Fantom—developed for the 2018 Alexa prize to shape the future features of Amazon Alexa leveraged reciprocal self-disclosure to gather information from users (Bowden et al., 2019; Jonell et al., 2018).[1]

While reciprocal self-disclosure seems to work as a data acquisition method to improve the quality of CAs in terms of providing user-specific responses and engaging in meaningful conversations with users, its influence on users' perception of the CA after the interaction remains unclear. Given people's growing concerns about technology companies' data acquisition attempts (Zuboff, 2019), which can negatively affect the ability of companies to retain users, it is important to understand whether reciprocal self-disclosure influences users' trust in a CA.

CA self-disclosure induces people to disclose information because the CA is exploiting a social norm and the user feels compelled to reciprocate (Sprecher et al., 2013). Thus, on the one hand, using reciprocal self-disclosure as a strategy could backfire because users could react negatively if they feel that they were manipulated into disclosing information (Collins & Miller, 1994). Should this occur, users may lose trust and stop using the CA or provide false information to it in future interactions. On the other hand, reciprocal self-disclosure could actually build trust by serving as a small-talk strategy that enhances the user-CA "interpersonal" relationship (Bickmore & Cassell, 2001). This could be helpful in future interactions as users would have a positive attitude toward the CA. For conversational agents to reach their full potential, it is therefore critical to understand how CA reciprocal self-disclosure affects users' trust.

Our review of the literature reveals a distinct knowledge gap regarding the effect of reciprocal self-disclosure on post-interaction trust (i.e., users' trust in a CA after an interaction session) (Collins & Miller, 1994; Jones & Archer, 1976; Lemay Jr & Melville, 2014; Zimmer et

[1] SlugBot used rules of gradual reciprocal self-disclosure to understand users' interests by asking them to share progressively more intimate information after revealing similar information about itself (Bowden et al., 2019). Fantom kept the same level of self-disclosure as the users' during the initial phase of the conversation and disclosed more information about itself whenever needed during the rest of the conversation (Jonell et al., 2018).

al., 2010) and the mechanism through which such an effect may occur, particularly in the context of conversational agents (W. Seymour & Van Kleek, 2021). This context is unique because conversational agents are nonhumans that often possess humanlike characteristics (e.g., humanlike language capabilities). Users' perceptions of these characteristics may influence the trusting mechanism in unprecedented ways. However, the existing literature does not provide an explanation of this mechanism in the context of conversational agents.

Previous literature has shown that people often anthropomorphize (i.e., engage in the process of humanization of) nonhuman agents to understand their complex behaviors (Waytz, Morewedge, et al., 2010), especially when interacting with computers, robots, and intelligent agents (W. Seymour & Van Kleek, 2021). Reciprocal self-disclosure constitutes a complex behavior for a CA, as reciprocity is often perceived as a prototypically human behavior (Fox & Tiger, 1971; Leakey & Lewin, 1978). Thus, we propose that it is plausible that anthropomorphism can help explain how people make sense of reciprocal self-disclosure in human-CA interaction. Moreover, we propose that anthropomorphism, as an inductive inference mechanism, can shed light on the trusting mechanism. This is plausible because previous research has indicated that anthropomorphism is driven by distinct underlying motivations that can influence people's judgment of the trustworthiness of the anthropomorphized entity (as described in the theoretical background section of this paper) (Epley et al., 2007; Epley, Akalis, et al., 2008; Epley, Waytz, et al., 2008; Waytz, Morewedge, et al., 2010). Therefore, the underlying motives that drive people to anthropomorphize nonhuman agents can provide theoretical insight into how users adjust their trusting beliefs of a CA. To investigate our speculated theory (Van de Ven, 2007), we seek to address the following research questions:

RQ1*:* What is the effect of reciprocal self-disclosure on users' trust in a CA?

***RQ2**: What is the role of anthropomorphism in the relationship between reciprocal self-disclosure and post-interaction trust?*

To address these research questions, we draw upon prior literature on anthropomorphism and trust and formulate a nomological network to connect reciprocal self-disclosure to users' post-interaction trust in a CA. First, we draw upon the psychology and neuroscience literature regarding anthropomorphism to explain how users try to make sense of self-disclosure by the CA (a nonhuman agent). Leveraging the prior research suggesting that reciprocity is perceived to be one of the main characteristics of being human (Fox & Tiger, 1971; Leakey & Lewin, 1978), we theorize why self-disclosure by a nonhuman agent could act as an anthropomorphic feature providing supporting evidence that a human-based mental model of the agent could help the user better understand the observed behavior. Second, we consider two types of trustworthiness (i.e., cognition-based and affect-based) that can help unravel reciprocal self-disclosure's cognitive and affective consequences. We theorize why the underlying motivations for anthropomorphism provide cognitive and emotional reasons for users to change their perception of cognition- and affect-based trustworthiness of a CA.

We conduct two randomized experiments to test our theory. We manipulate reciprocal self-disclosure using both a custom-developed text-based CA (Experiment 1) and a custom-developed voice-based CA (Experiment 2). In both experiments, the treatment group is exposed to a CA that provides intimate information about itself (i.e., reciprocal self-disclosure condition) and the control group is exposed to a CA that does not provide intimate information about itself (i.e., no reciprocal self-disclosure condition). In the experiments, the subjects repeatedly interact with the CA and the CA asks participants to reveal information about themselves.

Our results reveal that reciprocal self-disclosure increases the level of anthropomorphism of a CA by users. High levels of anthropomorphism associate positively with humans' cognition- and affect-based trustworthiness of CAs, which in turn, shapes humans' post-interaction trust.

Our study makes three key contributions to the literature. First, we delineate the importance of an artifact's perceived humanness (i.e., anthropomorphism) as an inductive inference mechanism and explain why reciprocal self-disclosure makes people anthropomorphize the CA artifact. Second, we propose a context-specific theory of why anthropomorphizing the CA has distinct effects on cognition- and affect-based trustworthiness. Third, we advance the literature on information disclosure and privacy decisions by focusing on anthropomorphism and trust, two factors often exploited by both legal data seekers and cybercriminals (Acquisti et al., 2020; Giddens, 2021; Shepherd, 2021).

2. Theoretical Background

2.1. Reciprocal Self-Disclosure

We define *self-disclosure* as the voluntary sharing of any information about the self, including thoughts, opinions, emotions, or personal information, that one entity communicates to another (Posey et al., 2010). Self-disclosure plays a central role in the development and maintenance of relationships (Collins & Miller, 1994). Self-disclosure in interpersonal relationships is reciprocal (Ehrlich & Graeven, 1971). *Reciprocity* (also known as social reciprocity) is the tendency to repay any benefits, gifts, or favors received by a party from another party (Ehrlich & Graeven, 1971; Lee & Choi, 2017; Sprecher et al., 2013). Thus, in line with the extant literature, we define *reciprocal self-disclosure* as an individual's tendency to repay a self-disclosure by another party by disclosing similar information about themself (Sprecher et al., 2013).[2]

Reciprocal self-disclosure has been studied not only in human-human interaction (Sprecher et al., 2013) but also in human-computer interaction (Moon, 2000) and in the interaction of humans and relational agents (i.e., agents designed to establish and maintain long-term social-emotional relationships with their users) (Bickmore & Picard, 2005). Reciprocal self-disclosure has been shown to be present in both online and offline contexts (Barak & Gluck-Ofri, 2007), among strangers with or without face-to-face interactions (X. Li et al., 2017), in computer-mediated communications (Jiang et al., 2013; Nguyen et al., 2012), and across different cultures (Katagiri et al., 2001).

Despite some efforts to understand the role of anthropomorphism in self-disclosure by adding more humanlike features such as an avatar or voice (e.g., Kang & Gratch, 2010; Pickard et al., 2016), there is still inconsistency in the conceptualization and operationalization of anthropomorphism (as discussed next) and the existing studies do not explain whether or why anthropomorphism may play a role in reciprocal self-disclosure's effect on *post-interaction* trust.

2.2. Anthropomorphism

Due to a lack of consensus on the definition of anthropomorphism in psychology and human-computer interaction (HCI) research, findings from previous research are often hard to reconcile. For instance, de Visser et al. (2016, p. 331) define anthropomorphism as "the degree to which an agent exhibits human characteristics." Likewise, Gong (2008, p. 1495), similar to many other HCI scholars, defines anthropomorphism as "the technological efforts of imbuing computers

[2] It is important to note that while reciprocal self-disclosure can be used as a method to create small-talk between two agents, not all small-talk involves reciprocal self-disclosure (Bickmore & Cassell, 2001). Therefore, in our review of the literature, we focus on the findings specific to reciprocal self-disclosure.

with human characteristics and capabilities." However, these definitions are at odds with the definition of anthropomorphism proposed in the mind perception literature (Epley et al., 2007), which is now widely used in the fields of psychology, management, and information systems (Glikson & Woolley, 2020).

In accordance with the mind perception literature, we define *anthropomorphism* as an inductive inference about real or imagined nonhuman entities that leads to the *attribution* of humanlike characteristics, properties, emotions, inner mental states, and motivations to them (Epley et al., 2007; Epley, Waytz, et al., 2008; H. M. Gray et al., 2007).[3] Anthropomorphism entails an inference about unobservable characteristics of an entity. In other words, a person might imagine that an entity has humanlike characteristics without observing those characteristics (Yuan & Dennis, 2019).

Moreover, anthropomorphism is not only about treating an object as living (i.e., animism) but involves attributing human-typical characteristics to it (Epley, Waytz, et al., 2008). Anthropomorphism is *not* the equivalent of adding humanlike features to an artifact. However, adding humanlike features, such as humanlike voice or avatar, to an artifact may *trigger* people to anthropomorphize the artifact (Gong & Nass, 2007; Qiu & Benbasat, 2009; Yuan & Dennis, 2019). For instance, the fact that an artifact possesses humanlike voice does not automatically mean that users will attribute characteristics such as consciousness and free will to it, but may increase the likelihood of such attribution. While humanlike features can act as signals or triggers to make people anthropomorphize the artifact, anthropomorphism resides in users' minds, not in artifacts. Therefore, we propose that this conceptualization of anthropomorphism, which is often missing from HCI studies on reciprocal self-disclosure (Gambino et al., 2020),

[3] In layman's terms anthropomorphism is a person's perception of the humanness of a nonhuman entity.

can shed more light on how people perceive CAs after they engage in the process of reciprocal self-disclosure with the CAs.

Previous studies predominantly focused on adding sensory cues such as humanlike appearance and voice to nonhumans (i.e., form anthropomorphism) to induce anthropomorphism, while comparatively fewer studies investigated how humanlike actions (i.e., behavioral anthropomorphism) such as reciprocal self-disclosure induce anthropomorphism (Gambino et al., 2020; Nowak & Fox, 2018).

Epley et al. (2007) proposed the *three-factor theory of anthropomorphism*, which suggests that anthropomorphism is largely determined by three major factors: (1) elicited agent knowledge, (2) effectance motivation, and (3) sociality motivation. First, elicited agent knowledge refers to the accessibility and applicability of egocentric or homocentric knowledge. Since knowledge about oneself and other humans is readily accessible and could be applicable to an entity, people apply such knowledge as a heuristic to explain observed behaviors. Therefore, anthropomorphism could be a side effect of the use of accessible and applicable knowledge about humans.

Second, effectance motivation, in the context of anthropomorphism, refers to the motivation to interact effectively with nonhuman agents and enhances "one's ability to explain complex stimuli in the present and to predict the behavior of these stimuli in the future" (Epley et al., 2007, p. 866). Neuroscientists argue that our brain's main task is to predict its surroundings (Clark, 2013). Since a human's best predictive model is the one about themself, humans leverage this model to predict the behavior of other humans (Broadbent, 2017). Research has shown that we use the same neural system to understand the behavior of not only other humans, but also anthropomorphized agents (Castelli et al., 2000; Iacoboni et al., 2004). Therefore, anthropomorphism might give us more predictive power, or the perception thereof, when dealing with nonhuman agents.

Third, sociality motivation refers to the motivation for social connection. People have the desire for social contact. Therefore, people often create humans out of nonhumans to satisfy their need for social connectedness (Mourey et al., 2017). Effectance and sociality motivations indicate the outcome people seek when they anthropomorphize an agent (Epley et al., 2007; Mourey et al., 2017). In other words, while such motivations can drive anthropomorphism, the outcome of the process is an increased perception of predictability and connectedness (Mourey et al., 2017).

Researchers have shown that anthropomorphism influences trust (Waytz et al., 2014). However, there is little consensus on whether the effect is positive or negative as the effect might vary from context to context (Glikson & Woolley, 2020). For instance, Waytz et al. (2014) theorized that in the context of autonomous driving anthropomorphism increases trust because people perceive an anthropomorphized entity to be more competent than a non-anthropomorphized entity as people attribute more agency to an anthropomorphized entity. Attribution of agency means that they believe the entity is capable of thinking, planning, and controlling its own actions, and therefore able to perform its intended tasks successfully (K. Gray et al., 2012). While this account discusses one possible way in which anthropomorphism could influence trust (i.e., through perceived competence of an entity), why and how anthropomorphism influences trust remains poorly understood. To better understand how anthropomorphism is related to trust and reciprocal self-disclosure, we examine the trust literature and identify its cognitive and affective bases.

2.3. Post-Interaction Trust

Reciprocal self-disclosure involves some degree of risk (e.g., loss of control over personal information) (Dinev & Hart, 2006; Smith et al., 2011; Xu et al., 2011). Thus, trust is an essential component in this discourse (Dinev & Hart, 2006; Smith et al., 2011). *Trust*, as we use it in this research, is defined as "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor [i.e., the trusting entity], irrespective of the ability to monitor or control that other party" (Mayer et al., 1995, p. 712).

People engage in risk-taking in relationships based on their current levels of trust in the other entity. After the risk-taking behavior, the actual outcome of the behavior serves as feedback for the trustor to update their levels of trust for future interaction (Mayer et al., 1995). In other words, people's trust in future interactions depends on how they perceive the outcome of their current risk-taking behaviors such as engaging in reciprocal self-disclosure. Therefore, *post-interaction trust* refers to the level of trust following a risk-taking behavior.[4]

Prior studies have suggested different ways to conceptualize and operationalize trust when it comes to humans versus specific technologies. For instance, McKnight et al. (2011) initially proposed *trust in a specific technology* for dealing with technological artifacts. However, in a subsequent study, Lankton, McKnight, and Tripp (2015) empirically assessed the appropriateness of such a conceptualization when dealing with technological artifacts having different levels of human-likeness. They concluded that a technology-based conceptualization should be used for non-humanlike technologies such as spreadsheet software or anti-phishing

[4] We would emphasize that post-interaction trust is not qualitatively different from trust as a construct; instead "post-interaction" refers to the stage at which trust is assessed. We focus on post-interaction trust because it captures a user's attitude toward a CA after engaging in reciprocal self-disclosure and lays the foundation for future interactions.

tools (e.g., Schuetz et al., 2022), but that a human-based conceptualization of trust should be used for humanlike technologies such as recommendation agents (e.g., W. Wang et al., 2016). Therefore, in this study, we use a human-based conceptualization of trust as we focus on users' trust in a CA.

Trust is one "unitary experience" (Lewis & Weigert, 1985, p. 972), which is formed based on the trustor's perception of the trustee's trustworthiness (Mayer et al., 1995; M. Seymour et al., 2021).[5] Trustworthiness can have cognitive and affective bases (Legood et al., 2022), both of which could be rooted in previous interactions and experiences of the trustor with the trustee (McAllister, 1995; Schoorman et al., 2007).

Cognition-based trustworthiness refers to the cognitive bases of trust (Legood et al., 2022; McAllister, 1995). Evaluation of trustworthiness could be based on a cognitive process through which the trustor chooses who is trustworthy (Kanawattanachai & Yoo, 2007), based on what they consider to be "good reasons" or evidence of trustworthiness (Lewis & Weigert, 1985, p. 970). Previous knowledge and information about the trustee provide some foundations for trust. For instance, knowing that an agent has always behaved to one's benefit in previous interactions may make it seem more likely that it will continue to do so in the current interaction.

Cognitive trustworthiness comprises performance-relevant cognitions about the trustee (Schaubroeck et al., 2011). Most scholars agree on ability and integrity as components of cognition-based trustworthiness (Komiak & Benbasat, 2006; Schaubroeck et al., 2011). *Ability* refers to the trustee's set of skills, competencies, and characteristics that enables them/it to have influence within a specific task domain. For instance, after a short interaction with a CA, a user may find it trustworthy if it successfully recognizes the user's utterances. *Integrity* refers to the

[5] The difference between trust and trustworthiness is that *trust* is the "willingness" to be vulnerable based on the *trustworthiness* of the other party.

trustor's perception that the trustee adheres to a set of principles that the trustor finds acceptable (Mayer et al., 1995). For example, a user may find a CA trustworthy if it provides consistently honest and unbiased answers to their questions.

Prior research on trust has often assumed that cognition-based trustworthiness represents a rational assessment of the trustee (Glikson & Woolley, 2020; Legood et al., 2022). However, empirical studies have shown that people's cognitions are not always rational because humans have bounded rationality and awareness (e.g., people tend to take cognitive shortcuts when judging others) (Bazerman & Moore, 2013; Kahneman, 2003). In other words, a user's cognitive assessment of a CA may not be entirely rational because of cognitive shortcuts a user might take (Bazerman & Moore, 2013), such as a user's overconfidence in their evaluation of the CA's abilities based on limited interaction.

Affect-based trustworthiness refers to the affective bases of trust (Legood et al., 2022; McAllister, 1995).[6] This affective element of trustworthiness is the emotional bond among parties in a relationship (Lewis & Weigert, 1985) and may arise from a psychological attachment to and perceived closeness and warmth of the trustee (W. Wang et al., 2016). In addition, this element emphasizes empathy, affiliation, and rapport (Schaubroeck et al., 2011), and is "grounded in reciprocated interpersonal care and concern" (McAllister, 1995, p. 25). For example, the user may find the CA trustworthy because the CA disclosed some information about its shortcomings and vulnerabilities, which the user finds "adorable." While affect-based trustworthiness was traditionally assumed to take time to develop (McAllister, 1995), recent

[6] Affect-based trustworthiness and other conceptualizations of the affective aspect of trust, such as emotional trust (Johnson-George & Swap, 1982) and faith (Rempel et al., 1985), closely parallel the concept of benevolence, which is defined as "the extent to which a trustee is believed to want to do good to the trustor, aside from an egocentric profit motive" (Mayer et al., 1995, p. 718). While some scholars have argued that benevolence and affect-based trustworthiness may refer to closely related, yet separate constructs (e.g., Schoorman et al., 2007; W. Wang et al., 2016), other scholars directly used benevolence as affect-based trustworthiness (e.g., Shih et al., 2017) or dropped benevolence from cognition-based trustworthiness when they independently measured affect-based trustworthiness, especially in the context of nonhuman agents (e.g., Komiak & Benbasat, 2006). In the current study, we focus on affect-based trustworthiness and do not include benevolence in our conceptualization of trustworthiness.

studies show that affective bases of trust do not always require much time to develop and could even be formed immediately (Legood et al., 2022).

3. Hypotheses Development

Our research model involves the relationship between reciprocal self-disclosure and trust and highlights the mediating role of anthropomorphism in this relationship. We leverage the three-factor theory of anthropomorphism (Epley et al., 2007) as a unifying theoretical lens through which to theorize answers to our research questions. We also employ the trust framework to unpack the concept of trust in our model (Glikson & Woolley, 2020; Mayer et al., 1995; McAllister, 1995; W. Wang et al., 2016). The three-factor theory of anthropomorphism and the trust framework have been used together previously within a single model as they do not contain conflicting underlying assumptions or overlapping concepts (Glikson & Woolley, 2020; Waytz et al., 2014).

Our overarching theory is that users leverage anthropomorphism as an inductive inference mechanism to understand reciprocal self-disclosure and that this mechanism involves distinct cognitive and affective components, based on which users adjust their trust in a CA for future interactions. Figure 1 shows our research model and hypotheses.

We do not explicitly hypothesize the relationships among cognition-based trustworthiness, affect-based trustworthiness, and trust as they have been tested in other contexts (see Ha et al., 2016; Johnson & Grayson, 2005; Komiak & Benbasat, 2006; McAllister, 1995; W. Wang et al., 2016). However, we do measure and test them empirically. Inclusion of these relationships is crucial to our model because they help provide a theoretically grounded explanation for trust in the context of reciprocal self-disclosure and allow us to test the nomological validity of our model.

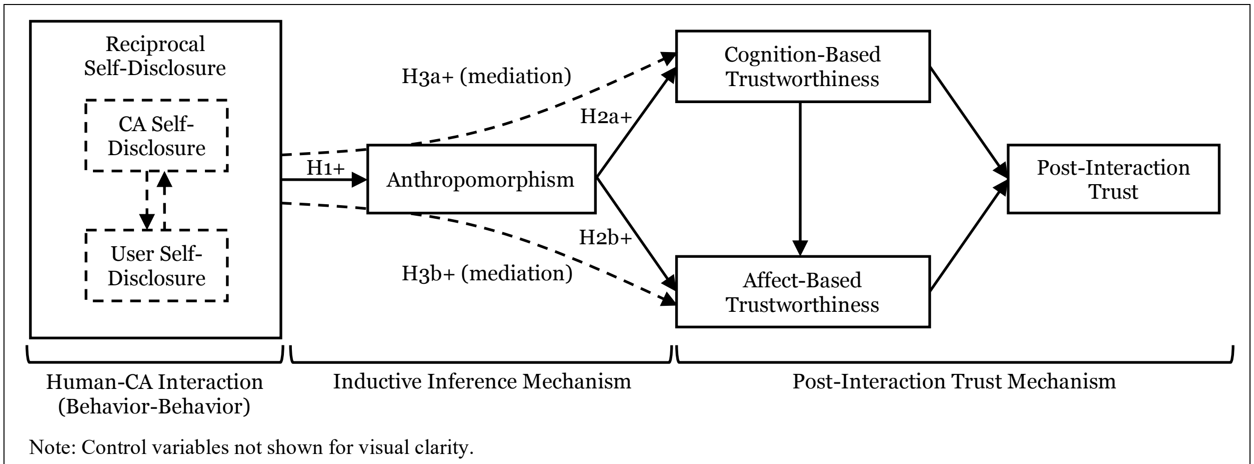


Figure 1. Research Model

3.1. Effect of Reciprocal Self-Disclosure on Anthropomorphism

When a user interacts with a nonhuman agent such as a CA, they become involved in a process of inductive inference about the agent. This inductive inference entails a number of cognitive operations: knowledge acquisition, activation of existing knowledge, and application of activated knowledge (Epley et al., 2007). Inductive inference typically starts with highly accessible knowledge structures, which may be adjusted based on alternative knowledge structures that are activated later in the interaction.

In the case of anthropomorphism (as an inductive inference process), three likelihoods play central roles in the way people make inferences about nonhuman agents: (a) the likelihood of activating knowledge structures about humans when making an inference about nonhumans, (b) the likelihood of adjusting this inference based on alternative knowledge structures such as direct knowledge about nonhuman agents, and (c) the likelihood of applying the aforementioned activated, often adjusted inference (Epley et al., 2007). These three likelihoods can be influenced by people's System 1 and System 2 thinking in their interaction with nonhuman agents. System 1 thinking refers to an intuitive thinking system, which is typically fast, automatic, effortless, implicit, and emotional, and System 2 thinking refers to deliberate and slow thinking (Bazerman

& Moore, 2013; Kahneman, 2011; Stanovich & West, 2000). Here we lay out how the process of reciprocal self-disclosure impacts this inductive inference process (predominately based on System 1 thinking), influencing anthropomorphism and trust.

When a user engages in reciprocal self-disclosure with a CA, both the *act* of gradual reciprocal self-disclosure and the *information* disclosed by the CA in this process may influence anthropomorphism. The *act* of reciprocal self-disclosure may impact anthropomorphism in two ways. First, research has shown that relationships develop through gradual increases in self-disclosure (Altman & Taylor, 1973; Knapp et al., 2014) and that engaging in gradual reciprocal self-disclosure is a common norm and expectation in human-human interactions (Cropanzano & Mitchell, 2005; Sprecher et al., 2013). Second, users may view reciprocity by the CA as a persuasion tactic, which is generally considered a humanlike behavior (Cialdini, 2021). Therefore, when a user observes the CA following the gradual reciprocity norm, they see first-hand that the CA follows a humanlike model of behavior. Such evidence could *consciously* or *unconsciously* influence the user's inductive inference of a CA in that it may (a) directly activate knowledge structures about humans, (b) help partially adjust the already activated knowledge structures about nonhumans (anchor) toward knowledge structures about humans, or (c) increase the likelihood of applying the activated or adjusted knowledge structures about humans as it foments the idea in the mind of the user that a human-based understanding of the agent is likely to be valid and predictive of the agent's future behavior.

The *information* disclosed by the CA in this process may also influence anthropomorphism. Disclosure of deep intimate information about the self is an inherently human behavior (Moon, 2000; Nass & Moon, 2000). Intimate disclosures include information such as self-concept, fears, values, vulnerabilities, and regrets (Altman & Taylor, 1973). We argue that when a CA discloses intimate information, this can influence a user's perception of different aspects of a CA's state of mind. First, self-disclosure of emotions (e.g., joy, sadness, and fear) can *signal* that the CA is capable of experiencing some level of emotion (even if that is not objectively true). For instance, when a CA discloses that it is scared, it is essentially signaling that it is capable of being scared (i.e., capable of experiencing an emotion). Being able to experience emotions qualifies the CA as a moral patient (K. Gray et al., 2012), and an entity that has the capacity to feel emotions is usually perceived to possess a humanlike state of mind (H. M. Gray et al., 2007). Second, disclosing information about regrets and disappointments can *signal* that the CA has the capacity to act and exert self-control, i.e., the CA is a moral agent. For example, when a CA states that it regrets doing something, it is essentially signaling that it is capable of doing that thing (i.e., it has the capacity to act). Disclosure of intimate information by the CA therefore provides a signal to users that it not only has agency as a moral agent, but also the capacity to experience as a moral patient. Agency and the capacity to experience collectively define the human state of mind (H. M. Gray et al., 2007), which is central to defining anthropomorphism (Waytz, Gray, et al., 2010). These pieces of information may serve as alternative knowledge structures to bias a user's inference of a CA toward relying on knowledge structures about humans or as triggers that activate knowledge structures about humans (i.e., System 1 thinking). In either case, they should increase the likelihood of anthropomorphism by the user.

In short, the act of gradual reciprocal self-disclosure and the information disclosed by the CA can signal humanlike characteristics in the CA. Since being human is the thing we know best (Broadbent, 2017), when faced with other things that apparently possess humanlike characteristics (e.g., a CA), the mirror neurons in our brain are activated (Saygin et al., 2011;

Schilbach et al., 2013). This activation leads the user to analyze a CA, and nonhuman agents in general, using human-based concepts and to attribute a higher degree of human state of mind to it (i.e., anthropomorphism). In other words, knowledge structures about humans are more likely to be activated when interacting with a CA that engages in reciprocal self-disclosure. This effect of anthropomorphizing, as we illustrate below, serves as an inductive inference mechanism that helps the user understand and predict a CA's behavior. Therefore, we advance the following hypothesis:

H1: Reciprocal self-disclosure is positively associated with users' anthropomorphism of a CA.

3.2. Effect of Anthropomorphism on Cognition- and Affect-Based Trustworthiness

Anthropomorphism may influence cognition-based trustworthiness. First, anthropomorphism may increase the perceived ability dimension of cognition-based trustworthiness of the agent because an agent with a humanlike state of mind is more likely to have more advanced means to fulfill expected tasks. Specifically, an anthropomorphized agent is perceived to have more agency, which is an important part of the humanlike state of mind (H. M. Gray et al., 2007). An agent with more agency appears capable of fulfilling tasks, planning, and controlling its own actions (K. Gray et al., 2011; Waytz et al., 2014). A user should, therefore, perceive a CA with more agency to be better able to fulfill its intended task compared to a CA with little agency.

Second, anthropomorphism may increase the perceived integrity of the CA. For a user to perceive a CA to have high integrity, they should perceive the CA to adhere to a set of principles that they find acceptable. In the context of our research, users likely perceive other humans as their in-group and CAs as their out-group when considering humans versus CAs (Tajfel & Turner, 2004). Research has shown that people are biased to positively evaluate the actions of their in-group members compared to out-group members (Haslam & Loughnan, 2014;

Molenberghs, 2013). However, people's boundaries for what is considered in-group and out-group are relativistic and can easily shift (Delhey et al., 2011). The process of humanization can subjectively shift the status of a CA from a nonhuman to a human, i.e., from out-group to in-group. Thus, we argue that a user's anthropomorphism of a CA leads to a biased positive evaluation of the set of principles that govern the CA's behavior when compared to a CA that is not anthropomorphized as much by the user.

Additionally, prior research has shown that one of the main reasons people anthropomorphize nonhumans is to increase their ability to predict the behavior of nonhuman artifacts (Epley et al., 2007; Waytz, Morewedge, et al., 2010). This is perhaps why people are more likely to anthropomorphize artifacts that show apparently unpredictable behavior (i.e., in order to predict them better) (Waytz, Morewedge, et al., 2010). Thus, anthropomorphism increases the user's *perceived* ability to predict the CA's behavior, and thus a user who anthropomorphizes a CA likely perceives that the CA's behavior is predictably governed by an acceptable set of principles when compared to a CA that is not anthropomorphized as much by the user. Therefore, we posit that anthropomorphism can provide cognitive bases for trustworthiness and we state the following hypothesis:

H2a: Anthropomorphism is positively associated with cognition-based trustworthiness.

Anthropomorphism may also influence affect-based trustworthiness. The rationale is that anthropomorphism can satisfy (at least to some extent) users' social needs, which are vital to human experience (Mourey et al., 2017). People seek to satisfy their social needs directly or symbolically through compensatory processes (Mourey et al., 2017), such as anthropomorphism (Epley et al., 2007). In other words, they often mentally construe nonhumans as humans to satisfy their need for social connectedness (*sociality motivation*). In fact, empirical evidence suggests that lonely and socially excluded people are more likely to anthropomorphize robots (Eyssel & Reich, 2013) and artificial intelligent agents (Ruijten et al., 2015).

Moreover, research suggests that sociality motivation drives people not only to anthropomorphize nonhuman agents but to do so by ascribing socially supportive traits to them (Epley, Akalis, et al., 2008). Epley et al. (2007, p. 876) suggest that a person driven by sociality motivation may be more likely to perceive a nonhuman agent as "thoughtful and considerate" and less likely to perceive it as "vindictive and deceitful." Thus, we argue that when the user anthropomorphizes a CA, they are likely to mentally construe a caring human out of the CA to fulfill their need for social connectedness.[7]

Based on the collective knowledge from the sociality motivation literature discussed above, we argue that a side-effect of anthropomorphizing a CA is that the user is more likely to find the agent to be supportive and caring, which are fundamental to affect-based trustworthiness (McAllister, 1995). Therefore, we posit that a user who anthropomorphizes the agent is more likely to develop higher levels of affect-based trustworthiness in the agent and state the following hypothesis:

H2b: Anthropomorphism is positively associated with affect-based trustworthiness.

3.3. Inductive Inference Mechanism: Mediating Role of Anthropomorphism

Above, we discussed why reciprocal self-disclosure influences people's level of anthropomorphism of a CA and why anthropomorphism provides cognitive and affective bases for them to find the CA trustworthy. Here, we further discuss the role of anthropomorphism as an inductive inference mechanism that explains why reciprocal self-disclosure can increase cognition and affect-based trustworthiness.

[7] Our assumption is that an average person does not suffer from high levels of paranoia as previous research has shown paranoid people are more likely to see other entities and especially artificial intelligence as evil or malicious (Kramer, 1994; Szollosy, 2017).

A CA's engagement in reciprocal self-disclosure can trigger users to anthropomorphize the CA. The rationale is that readily available knowledge structures about humans can be directly used (availability heuristic) or serve as an initial estimation or anchor to understand and predict how the CA behaves (anchoring heuristic) (Epley et al., 2007). A side-effect of the mindless process of activation of human-related knowledge structures (Kim & Sundar, 2012) is that users employ the same neural system, both cognitive and affective, to understand the anthropomorphized agent as they do for other humans (Castelli et al., 2000; Iacoboni et al., 2004). Therefore, we argue that a CA's engagement in reciprocal self-disclosure leads to the activation of cognitive and affective knowledge structures about humans as opposed to nonhumans, which can shift the status of a CA from out-group to in-group (Tajfel & Turner, 2004). A similar phenomenon has been observed in human-human intergroup interactions, where scholars found that interaction and self-disclosure improve attitudes toward the out-group members (Davies et al., 2011; Olsson et al., 2005; Pettigrew, 1998) and may help an individual to shift the status of another individual from out-group to in-group (Brown & Hewstone, 2005). This implies that when people anthropomorphize a CA, they are more likely to relax any concerns they may harbor about it being a nonhuman agent and act as if they were interacting with another human.

Users' perception of a CA as an in-group has important cognitive and affective implications for their assessment of the CA. Particularly, research in neuroscience field has shown that people categorize and experience the actions and emotions of in-group members differently than out-group members (Molenberghs, 2013). First, people categorize in-groups differently. Specifically, results of fMRI experiments indicate that when people assess their in-group members, parts of the brain associated with self-identity are activated because they see in-groups as part of their social identity (Volz et al., 2009). Second, people tend to evaluate the exact same action more positively when taken by an in-group rather than an out-group member, due to not only their positive evaluation of the action itself but also their selective attention to the details of the action such that they favor the in-group member (Molenberghs et al., 2012). This biased evaluation of in-group members persists even when the person is asked to closely evaluate the behavior of an in-group member (Hastorf & Cantril, 1954; Molenberghs, 2013). Third, people tend to perceive the emotional state of in-groups differently (Molenberghs, 2013) and are more likely to resonate with what an in-group member feels (Bernhardt & Singer, 2012) compared to out-group members. Therefore, users' anthropomorphism of a CA that engages in reciprocal self-disclosure likely makes them prone to positive evaluation of both cognitive and affective bases of the CA's trustworthiness.

Based on these reasons and our discussion of the effect of reciprocal self-disclosure on anthropomorphism (see Section 3.1) and the effect of anthropomorphism on cognition- and affect-based trustworthiness (see Section 3.2), we hypothesize that:

H3a: Reciprocal self-disclosure has a positive indirect association with cognition-based trustworthiness via anthropomorphism.

H3b: Reciprocal self-disclosure has a positive indirect association with affect-based trustworthiness via anthropomorphism.

4. Research Method

We conducted two experiments to test our hypotheses in the context of text-based CA and voice-based CA because previous studies have questioned the applicability of findings in either context to the other (Gambino et al., 2020). In Experiment 1, we used a custom-developed text-based CA and in Experiment 2, we used a custom-developed voice-based CA. Both experiments used a posttest-only randomized design comparing treatment (reciprocal self-disclosure) to control (no reciprocal self-disclosure) (Shadish et al., 2002). We recruited participants from amazon Mechanical Turk (MTurk) for both experiments to ensure a diverse sample (Buhrmester et al., 2011; Chandler et al., 2019; Mason & Suri, 2012). Subjects from MTurk are well-suited to our research objective because they have some experience using digital technology.[8]

We conducted five pilot studies to assess and improve our manipulation of reciprocal self-disclosure and the CA's technical design before conducting the two experiments. The first pilot study (N_1=80) was focused on integrating our text-based CA (written in JavaScript) with Qualtrics, the platform in which we collected data for our first experiment. The second pilot study (N_2=100) was aimed at ensuring that our manipulation of reciprocal self-disclosure was successful for the text-based CA. The next three pilot studies (N_3=30, N_4=30, N_5=50) were conducted to ensure that our voice-based CA (written in Java) worked across a wide range of Android devices. Specifically, our aim was to ensure that the CA could (a) interact with our server-side program (written in Python) to load the randomly assigned experiment materials, (b) seamlessly stream audio to our server and stream the transcribed version of the audio back to the participants' device, (c) recognize participants' utterances given their often noisy surroundings, and (d) automatically detect the participants' end of speech to seamlessly initiate the next round of back-and-forth disclosures (i.e., to create a more natural conversational setting without requiring participants to manually stop and start recording their voices using a button). We asked the participants to provide feedback on how the CA worked and used their feedback to improve the CA's design after each pilot study until all participants were able to interact with the CA as

[8] We estimated the number participants needed for our studies using G*Power 3.1.9.6. Based on our pilot studies, we expected an effect size of $d \approx 0.35$ in Experiment 1 and $d \approx 0.5$ in Experiment 2. For these effect sizes, $\alpha = 0.05$, and power = 0.80, we needed 204 and 102 participants in experiments 1 and 2, respectively. Since, the effect sizes are not guaranteed and some participants might miss the attention check question, we chose to recruit 230, and 140 participants in experiments 1 and 2, respectively.

intended. We provide more details regarding the design of our CA when we discuss the particulars of Experiments 1 and 2.

4.1. Measures

Appendix A provides all measurement items used in the two experiments. Below, we describe our operationalization of constructs.

Anthropomorphism. We adopted Waytz, Cacioppo, et al.'s (2010) scale to measure anthropomorphism. This operationalization of anthropomorphism is based on the premise that anthropomorphism is about attribution of humanlike mental state to an agent (Seeger et al., 2021; Waytz, Cacioppo, et al., 2010; Waytz, Morewedge, et al., 2010) and includes five items, which we measured using a 7-point Likert scale.

Cognition-Based Trustworthiness. We adapted Wang et al.'s (2016) scale to measure cognition-based trustworthiness, making minimal changes to reflect the context of CAs. The Wang et al. (2016) measures include multiple indicators for each aspect of trustworthiness (i.e., ability, benevolence, and integrity). In line with our theory and previous research, we measured ability and integrity and did not use the benevolence scale because we measured affect-based trustworthiness separately (Komiak & Benbasat, 2006; Shih et al., 2017). We used a 7-point Likert scale to measure three items for ability and four items for integrity. We operationalized ability and integrity by creating linear composites of the averages of their items. We then operationalized trustworthiness as a linear composite of ability and integrity.

Affect-Based Trustworthiness. To measure affect-based trustworthiness, we adapted the scale developed by McAllister (1995). Many IS scholars have used a subset of the original items to fit their research context (see, for example, W. Wang et al., 2016). In this research, we retained three relevant items from the original scale with minimal changes and measured them using a 7-point Likert scale.

Post-Interaction Trust in CA. We adapted trust measures from Mayer and Gavin (2005). These items reflect the concept of trust by capturing participants' willingness to be vulnerable to the actions of the CA by disclosing intimate information about the self to it in future interactions. Since the original scale was developed for trust in the context of a company, we used only the items that could be appropriated for our context. Furthermore, we did not use the reverse coded items, because they might tap into the concept of distrust, which some scholars argue is different from trust (Dimoka, 2010). Our operationalization of post-interaction trust included three items measured using a 7-point Likert scale.

The operationalizations for trustworthiness and trust were based on the constructs originally developed for trust in human-human interactions. The rationale for this choice is that previous research has empirically shown that when the technological artifact is humanlike, a humanlike conceptualization and operationalization of trust better captures users' trust in the artifact (Lankton et al., 2015). This is also in line with most studies on technological agents (Komiak & Benbasat, 2006; W. Wang et al., 2016, 2018).

Control Variables. We controlled for participants' age, gender, level of education, and previous experience interacting with CAs. We also controlled for users' privacy concerns, which could predict their trust in a CA (Dinev et al., 2015; Smith et al., 2011). Further, we controlled for users' extroversion, which could affect the way users interact with a CA (Joosse et al., 2013).

4.2. Reciprocal Self-Disclosure Manipulation

We adapted Moon's (2000) method of handling reciprocal self-disclosure, in which a computer asked each participant 15 questions and disclosed some or no information about itself before each question. We made small changes to the content of self-disclosure to make it relevant to the context of our study and removed three unnecessarily intrusive questions (a total of 12 questions were retained). Appendix B includes a complete list of the CA's disclosures. In the treatment condition (reciprocal self-disclosure = 1), the CA began by disclosing public facts about itself. On its next speaking turn, the CA disclosed more private information about itself (see Supplemental Material A). This trend continued until the last turn (i.e., the twelfth question) in which the CA disclosed the most intimate information about itself. In the control condition (reciprocal self-disclosure = 0), the CA disclosed no intimate information about itself before each question. We included roughly the same amount of non-disclosure text (i.e., number of words) in both control and reciprocal self-disclosure conditions to rule out any possible effect that differential amounts of content might have on reciprocal self-disclosure.

In line with previous literature (Moon, 2003), we assessed the effectiveness of our manipulation by measuring whether it induced different levels of reciprocity by the participants. In doing so, we measured users' self-disclosure by capturing their text (Experiment 1) and voice utterances[9] (Experiment 2) while interacting with the CA. Because we manipulated the intimacy of the disclosed information (rather than the raw amount of disclosure), we measured the level of intimacy of the reciprocally disclosed information to evaluate the effectiveness of our manipulation—as suggested by prior research (Altman & Taylor, 1973; Y. Li, 2011; Smith et al., 2011; Yun et al., 2019).[10] To do so, we used the key properties of disclosure intimacy proposed by Altman and Taylor (1973). Accordingly, each utterance was rated from 1 to 7 in terms of intimacy (see Appendix B).

[9] Consistent with the Merriam-Webster dictionary, we use the term utterance to refer to an oral or a written statement.
[10] Please note that the raw amount of disclosure (e.g., the word count) cannot capture the intimacy of the disclosed information. Our manipulation check indicates that the participant revealed more intimate information in the treatment condition than the controlled condition. Our approach is in line with our manipulation, in which we manipulated the intimacy of the disclosed information rather than the raw amount of disclosure.

Since only the honest user self-disclosure matters in the context of our study (L. Wang et al., 2017), after the user's interaction with the agent, we asked the participant to indicate how much of the information they disclosed was actually true.[11] We multiplied the disclosure intimacy level by the honesty percentage to create an index for each user's self-disclosure of intimate information. We then used this index to check whether our manipulation of reciprocal self-disclosure was successful (the results are reported in Sections 5.3 and 6.3).

5. Experiment 1: Text-Based CA

5.1. Study Participants

We recruited 230 participants of whom 208 (95 females and 113 males, with an average age of 36.1 ranging from 19 to 71) passed the attention check question. We asked the following question as an attention check after participants finished the experimental task and answered questions regarding the constructs in our model: "As the popularity of smart speakers has risen—with Amazon Echo and Google Home leading the way and Apple's Homepod following—some have concerns about privacy. If you are paying attention, just ignore this question and choose the middle choice to answer the question." We also carefully read all utterances by the participants in the experimental task and identified the ones that did not include any human-readable words. For the main analyses, we dropped the 22 participants who failed the attention check question or did not attend to the reciprocal self-disclosure task. However, we also performed a parallel analysis including these participants and found that including them in the analysis did not change any of the findings in terms of direction and significance of the paths in our model. Therefore, we dropped them as suggested by Lowry et al. (2016). All participants were compensated with a $1.00 payment. Considering that the experimental task could be completed in approximately 8

[11] We told the participants that their answer to this question would not influence their compensation.

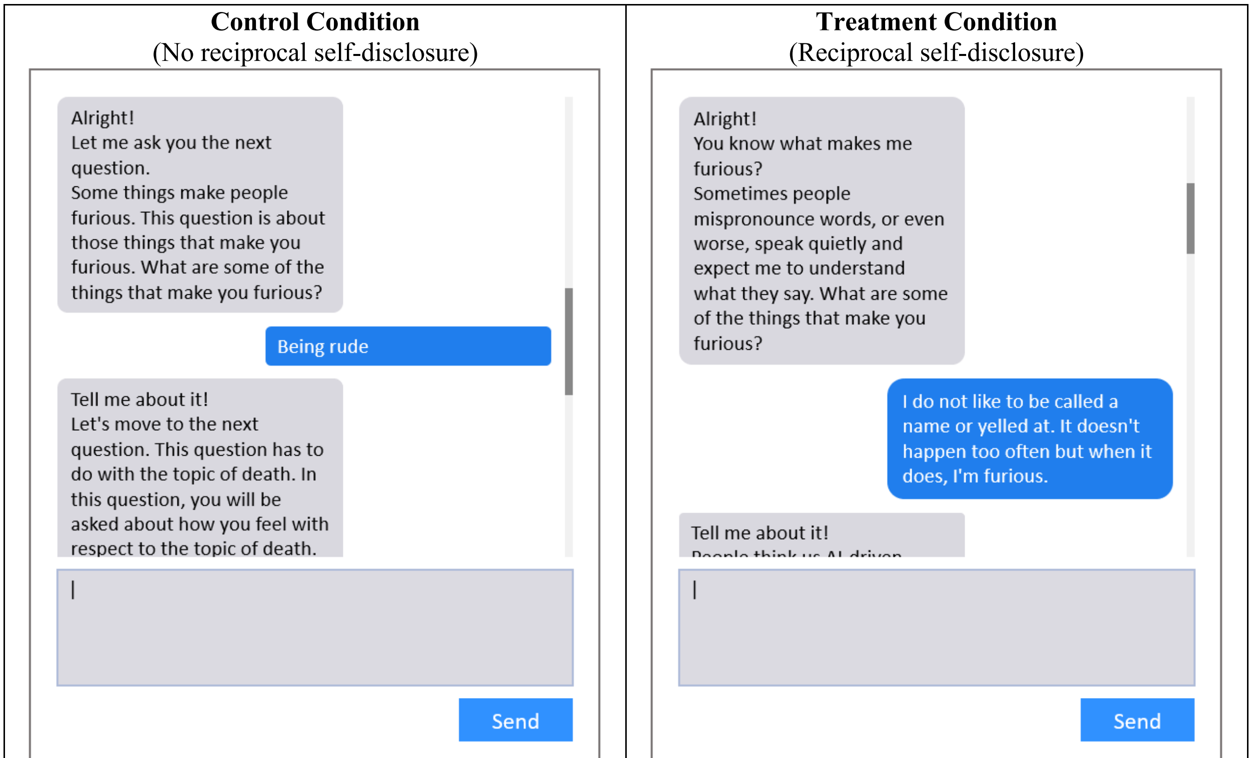
minutes, the compensation provided was on par with the average minimum wage in the U.S. (Bell, 2023), which should have provided sufficient motivation for participants to take the task seriously (Aruguete et al., 2019; Robinson et al., 2019).

5.2. CA and Procedure

For this experiment, we developed a text-based CA named “Amanda.” By developing an artifact that was similar to the text-based versions of actual CAs, such as Google Assistant, we created an engaging task environment for participants and increased the psychological realism and ecological validity of our study (Berkowitz & Donnerstein, 1982).

We asked the participants to open our web app on their browsers. We then asked participants to join a conversation with Amanda. Amanda would start the conversation by introducing itself. Then, Amanda would begin a reciprocal question-and-answer round. Before each question, Amanda would present a few sentences and then pose a question to the participant. Next, Amanda would wait for the participant to finish typing. Amanda would then use a transition word or sentence, such as “OK,” and start the next question and answer round (see Figure 2). In the treatment condition, Amanda would disclose some information about itself that was related to the question it would ask the user. In the control condition, Amanda would not disclose any intimate information but say something of a procedural nature such as “the next question has to do with your gender.” By doing so, we controlled for the length of the CA’s utterances in the two conditions (Moon, 2000).

After each round, the questions became more intimate. According to Archer and Berg (1978, p. 531), “biographical characteristics are low in intimacy,” and “fears, self-concepts, and basic values are high in intimacy.” Appendix B provides the sequence of disclosures and questions that Amanda uttered during the interaction. After the interaction, we redirected the participants to a questionnaire. We measured anthropomorphism, cognition-based and affect-based trustworthiness, and trust, in the form of a posttest, along with control variables such as age, gender, level of education, prior experience using CAs, privacy concerns, and extroversion. Finally, we thanked and debriefed all participants.



In the treatment condition, the CA uses text prompts to reveal some information about itself and then asks the user a question related to the disclosed information. The participant then responds to the question using text. See Appendix B for details.

Figure 2. A Sample of Participants' Interaction with the Text-Based CA

5.3. Analysis and Results[12]

Manipulation Check. Participants in the control condition disclosed less information ($M_{Control} = 3.787, SD_{Control} = 2.064$) compared to those in the treatment condition ($M_{Treatment} = 4.467, SD_{Treatment} = 1.982$; $t(206) = 2.425$, $p < 0.01, d = 0.336$), confirming a successful manipulation of reciprocal self-disclosure.

[12] We used Stata 16.1 for all analyses.

Measurement Model. Since we used well-established items to measure the constructs in our research model, we conducted a confirmatory factor analysis (CFA) to assess the measurement model. The fit measures for our model were CFI=0.961, RMSEA=0.064, and SRMR=0.066, indicating an acceptable fit (Hu & Bentler, 1999). After ruling out common method bias and confirming the convergent and discriminant validity of the measurement model (see Appendix C), we created linear composites for each construct by averaging the items for that construct (Cronbach's α for anthropomorphism, cognition-based trustworthiness, affect-based trustworthiness, and trust are 0.97, 0.93, 0.87, and 0.94 respectively). Appendix D shows the correlation matrix and variance inflation factor (VIF) values for the measured variables. We did not find common method bias to be an issue based on either the single factor approach (measured items could only explain 41.87% of the single factor, less than the acceptable 50% threshold) or marker variable analysis (Lindell & Whitney, 2001; Schuetz et al., 2020) (see Supplemental Material B).

Hypotheses Tests. We used a system of seemingly unrelated regressions (SUR) using generalized least-squares (GLS) estimator to test our hypotheses. This approach provides a more rigorous test of the paths as compared to hierarchical regression, which is often used in testing similar models (e.g., Venkatesh et al., 2016), because it allows for errors to be correlated for each given participant across the set of regressions used to estimate paths. Our results are presented in Table 1.

Hypothesis 1 stated that reciprocal self-disclosure was positively associated with anthropomorphism. Our results supported this claim ($\beta = 0.53, p < 0.05$). Reciprocal self-disclosure, along with control variables, explained 12.0% of the variance in anthropomorphism.

Table 1. Results for Experiment 1	Anthropomorphism	Cognition-Based Trustworthiness	Affect-Based Trustworthiness	Post-Interaction Trust in CA
Control Variables				
Constant	4.73*** (1.06)	4.76*** (0.59)	-0.60 (0.70)	1.70 (0.91)
Age	-0.04** (0.01)	-0.00 (0.01)	-0.01 (0.01)	0.00 (0.01)
Gender	-0.45 (0.27)	-0.14 (0.14)	0.29 (0.15)	0.28 (0.20)
Education	-0.23* (0.11)	-0.05 (0.06)	0.03 (0.06)	-0.08 (0.08)
Previous Experience	0.24 (0.12)	0.04 (0.07)	0.01 (0.07)	-0.02 (0.09)
Privacy Concerns	0.00 (0.08)	-0.13** (0.04)	0.03 (0.05)	-0.11 (0.06)
Extroversion	0.13 (0.08)	-0.02 (0.04)	0.07 (0.05)	-0.05 (0.06)
Independent Variables				
Reciprocal Self-Disclosure	0.53** (0.27)	-0.08 (0.14)	-0.16 (0.15)	-0.50** (0.19)
Anthropomorphism		0.40*** (0.04)	0.43*** (0.05)	0.09 (0.07)
Cognition-Based Trustworthiness			0.39*** (0.07)	0.45*** (0.10)
Affect-Based Trustworthiness				0.38*** (0.09)
Pseudo R^2	0.120	0.426	0.611	0.484

Mediation Analysis for Experiment 1 (10,000 Bootstrap Samples) Mediation Path	Indirect Effect (Bootstrap S.E.)	90% Confidence Interval	95% Confidence Interval
Reciprocal Self-Disclosure → Anthropomorphism → Cognition-based trustworthiness	0.21 (0.11)*	[0.033,0.394]	[-0.002,0.429]
Reciprocal Self-Disclosure → Anthropomorphism → Affect-based trustworthiness	0.23 (0.12)**	[0.040,0.424]	[0.003,0.461]

Notes:
a. *N=208*
b. ** p<0.10; ** p<0.05; *** p<0.01 (two-tailed tests are presented)*
c. *Unstandardized regression coefficients are shown.*
d. *Numbers in parentheses are the standard errors*

Hypothesis 2a stated that anthropomorphism was positively associated with cognition-based trustworthiness. Our results supported this claim ($\beta = 0.40, p < 0.001$). Anthropomorphism, along with control variables, explained 42.6% of the variance in cognition-based trustworthiness. Similarly, Hypothesis 2b predicted that anthropomorphism was positively associated with affect-based trustworthiness. Our results supported this claim ($\beta = 0.43, p < 0.001$), with the model explaining 61.1% of the variance in affect-based trustworthiness. Furthermore, a mediation analysis using 10,000 bootstrap samples (Preacher & Hayes, 2008) indicated significant indirect effects of reciprocal self-disclosure on cognition-based (H3a) ($\beta = 0.21, p < 0.10$) and affect-based (H3b) ($\beta = 0.23, p < 0.05$) trustworthiness via anthropomorphism. The direct effects of reciprocal self-disclosure on cognition-based trustworthiness ($\beta = -0.08, p = 0.60$) and affect-based trustworthiness ($\beta = -0.16, p = 0.28$)

were not found to be significant. Thus, reciprocal self-disclosure has an indirect-only impact (Zhao et al., 2010) on cognition-based trustworthiness and affect-based trustworthiness which was fully mediated by anthropomorphism (see Table 1).

Our results also confirmed the previously known paths in our model. Specifically, we found that cognition-based trustworthiness was positively associated with affect-based trustworthiness ($\beta = 0.39, p < 0.001$) and post-interaction trust in the CA ($\beta = 0.45, p < 0.001$) and that affect-based trustworthiness was positively associated with post-interaction trust in the CA ($\beta = 0.38, p < 0.001$).

6. Experiment 2: Voice-Based CA

While the results of Experiment 1 are promising for interactions with text-based CAs, such as those frequently encountered in online contexts (e.g., customer support), it is also important to examine whether these results hold for interactions with voiced-based CAs, such as Alexa and Siri, which are being increasingly used in our day-to-day lives. It is not a foregone conclusion that this would be so, given that voice is considered a richer channel than text-based chat (Ishii et al., 2019) and that voice-based interactions leave people feeling more connected to their conversational partner compared to text-based interactions (Kumar & Epley, 2021). In fact, some scholars have raised concerns about the applicability of findings using simple interfaces such as those used in text-based interactions to other contexts with different social affordances such as voice-based interactions (Gambino et al., 2020). In addition, several scholars have raised concerns about the replicability of social science findings and see it as one of the biggest challenges facing the scientific community (Camerer et al., 2018). Therefore, replicating our findings in a different context adds robustness to our findings. These reasons prompted Experiment 2.

6.1. Study Participants

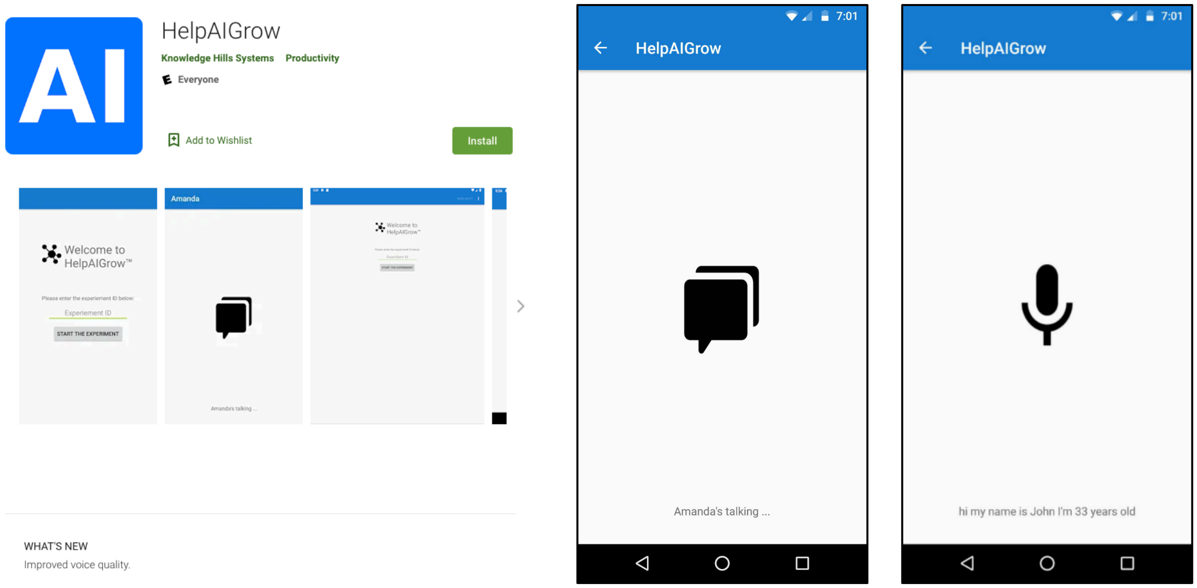
We recruited 140 participants of whom 98 (33 females and 65 males, with an average age of 34.7 ranging from 22 to 72) passed the attention check questions and were native English speakers. We needed the participants to be native English speakers to avoid speech-to-text issues associated with non-standard English accents. We assessed participants' attention by asking the same attention-check question we used in Experiment 1. We also carefully read all utterances by the users and identified the ones that did not include any human-readable words. For the main analyses, we dropped the 42 participants who failed the attention check question, did not attend to the reciprocal self-disclosure task, or self-identified as not being native English speakers. All participants were compensated with a $2.50 payment.

6.2. CA and Procedure

We developed a voice-based CA app for the Android operating system for this experiment. In the app, we leveraged Amazon Polly Neural Engine, one of the latest text-to-speech technologies, to produce humanlike synthesized voices (Amazon, 2021). We also used Google Cloud Speech-to-Text service, one of the most accurate speech-to-text engines in noisy environments, to show a real-time transcription of the user's voice input on the screen. Since our participants would use the app in their normal settings, we followed Google's best practices and used extra noise cancellation algorithms (Google Cloud, 2021). The use of Amazon Polly and Google Cloud Speech-to-Text helped make our CA behave like actual CAs on the market, enhancing both the ecological validity and the generalizability of our results.

We designed the app so that the participant could have a continuous conversation with Amanda, our CA, without the need to tap on a button to start or finish the utterance (similar to actual conversations with humans). To do so, we used the end-of-sentence signal provided by

Google Cloud to detect the end of the sentence and added a two-second grace period before the CA automatically responded. The app would reset the grace period after each new utterance by the participant.



A B C

The screenshot on the left-hand-side shows the app page on Google Play (A). In the reciprocation condition, Amanda used voice prompts to reveal some information about itself and then ask the user a question related to the disclosed information (B). The participant then responds to the question using voice (C).

Figure 3. Our Experimental App on Google Play and a Sample of a Participant's Interaction with the Voice-Based CA

We ensured each participant heard the CA's utterances by resetting the device's volume to 75% if they decreased the volume (we did not change the volume if they increased it). We explicitly asked all participants for permission to do this before the experiment started.

We asked participants to download the app from Google Play and open the app on their mobile phones (see Figure 3). After accepting the consent form, users read and accepted a notification to grant the app access to the microphone on their phones. Afterward, they started interacting with the CA in a conversation similar to Experiment 1, but this time voice-based. Participants would see a live text stream of their utterances on the screen as being detected by the agent. Figure 3 shows a sample of a participant's interaction with the CA.

6.3. Analysis and Results

Manipulation Check. Participants in the no reciprocal self-disclosure condition disclosed less information ($M_{Control} = 3.355, SD_{Control} = 1.706$) compared to those in the high reciprocal self-disclosure condition ($M_{Treatment} = 4.221, SD_{Treatment} = 2.056;\ t(96) = 2.280,\ p < 0.05, d = .463$), confirming a successful manipulation of reciprocal self-disclosure.

Measurement Model. As in Experiment 1, since we used well-established items to measure the constructs in our research model, we conducted a confirmatory factor analysis (CFA) to assess the measurement model. The fit measures for our model were CFI=0.940, RMSEA=0.076, and SRMR=0.062, indicating an overall acceptable fit.[13] Using the same approach as in Experiment 1, we tested for common method bias and did not find it to be an issue (see Supplemental Material B). After confirming the convergent and discriminant validity of the measurement model (see Appendix C), we created linear composites for each construct by averaging the items for that construct (Cronbach's α for anthropomorphism, cognition-based trustworthiness, affect-based trustworthiness, and trust are 0.95, 0.94, 0.86, and 0.86 respectively). Appendix D shows the correlation matrix and variance inflation factor (VIF) values for the measured variables.

Hypotheses Tests. We used a system of seemingly unrelated regressions (SUR) using generalized least-squares (GLS) estimator to test our hypotheses. Our results are presented in Table 2.

Hypothesis 1 stated that reciprocal self-disclosure was positively associated with anthropomorphism. Our results supported this claim ($\beta = 0.86, p < 0.001$). Reciprocal self-disclosure, along with control variables, explained 20.6% of the variance in anthropomorphism.

[13] The slight misfit (i.e., CFI below 0.95) is due to poor performance of the extroversion construct, which is used as a control variable. The measurement model fit excluding the extroversion construct is CFI=0.956, RMSEA=0.071, and SRMR=0.062.

Table 2. Results for Experiment 2

	Anthropomorphism	Cognition-Based Trustworthiness	Affect-Based Trustworthiness	Post-Interaction Trust in CA
Control Variables				
Constant	1.36 (0.78)	4.25^{***} (0.67)	-0.84 (0.83)	0.46 (1.15)
Age	0.00 (0.01)	0.01 (0.01)	0.01 (0.01)	0.00 (0.02)
Gender	0.12 (0.23)	-0.59^{**} (0.20)	0.05 (0.21)	0.02 (0.29)
Education	-0.01 (0.09)	-0.01 (0.08)	0.08 (0.08)	0.14 (0.11)
Previous Experience	0.20^{*} (0.09)	-0.01 (0.08)	0.07 (0.08)	0.10 (0.11)
Privacy Concerns	0.04 (0.07)	0.07 (0.06)	-0.05 (0.06)	-0.10 (0.08)
Extroversion	-0.13 (0.07)	-0.07 (0.06)	-0.05 (0.06)	0.02 (0.08)
Independent Variables				
Reciprocal Self-Disclosure	0.86^{***} (0.213)	0.21 (0.20)	-0.01 (0.20)	-0.04 (0.28)
Anthropomorphism		0.61^{***} (0.09)	0.74^{***} (0.11)	-0.06 (0.18)
Cognition-Based Trustworthiness			0.38^{***} (0.11)	0.48^{***} (0.15)
Affect-Based Trustworthiness				0.28^{**} (0.14)
Pseudo R^2	0.206	0.472	0.648	0.340

Mediation Analysis for Experiment 2 (10,000 Bootstrap Samples)

Mediation Path	Indirect Effect (Bootstrap S.E.)	90% Confidence Interval	95% Confidence Interval
Reciprocal Self-Disclosure → Anthropomorphism → Cognition-based trustworthiness	0.52^{***} (0.15)	[0.272,0.774]	[0.226,0.820]
Reciprocal Self-Disclosure → Anthropomorphism → Affect-based trustworthiness	0.64^{***} (0.18)	[0.339,0.932]	[0.284,0.987]

Notes:
a. *N=98*
b. ** p<0.10; ** p<0.05; *** p<0.01 (two-tailed tests are presented)*
c. *Unstandardized regression coefficients are shown.*
d. *Numbers in parentheses are the standard errors*

Hypothesis 2a stated that anthropomorphism was positively associated with cognition-based trustworthiness. Our results supported this claim ($\beta = 0.61, p < 0.001$). The model explained 47.2% of the variance in cognition-based trustworthiness. Similarly, Hypothesis 2b predicted that anthropomorphism was positively associated with affect-based trustworthiness. Our results supported this claim ($\beta = 0.74, p < 0.001$). Furthermore, a mediation analysis using 10,000 bootstrap samples (Preacher & Hayes, 2008) indicated significant indirect effects of reciprocal self-disclosure on cognition-based (H3a) ($\beta = 0.52, p < 0.001$) and affect-based (H3b) ($\beta = 0.64, p < 0.001$) trustworthiness via anthropomorphism. The direct effects of reciprocal self-disclosure on cognition-based trustworthiness ($\beta = 0.21, p = 0.30$) and affect-based trustworthiness ($\beta = -0.01, p = 0.97$) were not found to be significant. Thus, reciprocal self-disclosure has an indirect-only impact (Zhao et al., 2010) on cognition-based trustworthiness and affect-based trustworthiness which was fully mediated by anthropomorphism (see Table 2).

Again, our results confirmed the previously known paths in our model. Specifically, we found that cognition-based trustworthiness was positively associated with affect-based trustworthiness ($\beta = 0.38, p < 0.001$) and post-interaction trust in the CA ($\beta = 0.48, p < 0.01$) and that affect-based trustworthiness was positively associated with post-interaction trust in the CA ($\beta = 0.28, p < 0.05$).

In summary, the results of Experiment 2 replicated those of Experiment 1 suggesting that the findings hold for interactions with both text-based CAs and voice-based CAs. Taken together, the results of the two experiments provide strong evidence for our research model.

7. Discussion

In this research, we investigated the effect of reciprocal self-disclosure on post-interaction trust in the context of CAs. Our findings showed that reciprocal self-disclosure influences users' post-interaction trust in a CA through theoretically distinct cognitive and affective paths. These findings have several important implications for research and practice.

7.1. Implications for Research

By investigating the mediating role of anthropomorphism in the effect of reciprocal self-disclosure on post-interaction trust, we contribute to the existing literature in two ways. First, we contribute to the reciprocity literature by highlighting that anthropomorphism plays a major role in users' inductive inference of reciprocal self-disclosure in human-CA interactions. We conceptualized and operationalized anthropomorphism to capture whether people ascribe prototypically human attributes (e.g., consciousness, intention, free will, and capacity to experience emotions) to a CA that demonstrates reciprocal self-disclosure (i.e., a behavioral anthropomorphic cue). Our results indicated that people do in fact ascribe such attributes to the CA. This finding (a) is in line with several psychology and neuroscience studies that have shown that anthropomorphism is associated with many deeply held beliefs about nonhuman agents (Epley, Akalis, et al., 2008; Epley et al., 2007; Kay et al., 2010; Waytz, Morewedge, et al., 2010) and (b) adds to the growing body of literature that investigates the role of anthropomorphism as an inductive inference mechanism in people's perception of different nonhuman agents such as robots, intelligent agents, and supernatural agents (e.g., God and angels) (Epley, Waytz, et al., 2008).

We propose that scholars who study reciprocal self-disclosure in human-human interactions could also consider the possibility that humanization and dehumanization explain their results. For instance, it is plausible that one reason why people like others who engage in reciprocal self-disclosure process with them (Sprecher et al., 2013) is that they humanize those who adhere to the reciprocal self-disclosure principle and dehumanize those who do not—a novel explanation for the disclosure-liking hypothesis (Collins & Miller, 1994).

Second, we contribute to the trust literature by providing theoretical links between anthropomorphism and cognition-based and affect-based trustworthiness and in turn post-interaction trust. In doing so, we conceptually bridged prior research on motivations of anthropomorphism (Epley et al., 2007) and research on cognitive and affective bases of trust (McAllister, 1995). Based on our findings, when the context allows (e.g., in the context of reciprocal self-disclosure), people engage in the process of anthropomorphism. This process, in turn, provides users with the means to form a cognitive assessment of the agent's ability and integrity (i.e., cognition-based trustworthiness) and to establish a closer relational connection (i.e., affect-based trustworthiness) with the agent. While anthropomorphism is positively associated with both cognition-based and affect-based trustworthiness, according to our theory, the reasons for the two associations are not the same. Prior literature identified effectance motivation (i.e., the motivation to explain uncertainty in one’s surroundings) and sociality motivation (i.e., the desire for social contact) as two of the main motivations for anthropomorphism (Epley, Akalis, et al., 2008; Epley et al., 2007; Epley, Waytz, et al., 2008; Waytz, Morewedge, et al., 2010). We extend this literature by showing that the formation of cognition-based trustworthiness, which is conceptually related to effectance motivation, and the formation of affect-based trustworthiness, which is conceptually related to sociality motivation, can be enhanced by anthropomorphizing the agent. Our empirical results not only confirmed the effects of anthropomorphism on cognition-based and affect-based trustworthiness, but also indicated a significant mediating role of anthropomorphism in the effect of reciprocal self-disclosure on trustworthiness.

In addition, by focusing on the effect of reciprocal self-disclosure—a data acquisition method—on users’ post-interaction trust, we contribute to the privacy literature. While previous research has extensively studied reciprocal self-disclosure as a data acquisition method, the consequence of using this method is not well understood. Our study provides a nomological network to explain the effect of reciprocal self-disclosure on post-interaction trust. We believe that our proposed model provides a framework for scholars to build upon and evaluate the consequences of reciprocal self-disclosure from both cognitive and affective viewpoints (Dinev et al., 2015). Our theorization and findings regarding affective consequences of reciprocal self-disclosure can help scholars in the robotic companionship stream of research better understand how affect-based trustworthiness develops in reciprocal interactions. Further, our theorization and findings regarding cognitive outcomes of reciprocal self-disclosure can help scholars in fields such as military robotics to study the types of disclosure that induce cognition-based trustworthiness (Tegmark, 2017).

Moreover, privacy and security researchers can build on our findings and explore (a) how cybercriminals may use the intimate data acquired through reciprocal self-disclosure to prepare highly contextualized messages that can elicit even more sensitive information from the users, for instance, in phishing and whaling attacks (Pienta et al., 2020; Schuetz et al., 2020), and (b) whether the types of information that CA users find to be low and high in intimacy level (see Supplemental Material A) map to what cybercriminals may find useful, for instance, in developing password crackers.

7.2. Implications for Practice

Our research has practical and ethical implications for users and developers of CAs. First, users should be mindful that companies could use CAs not only to manipulate them into making disclosure decisions that are inconsistent with their privacy preferences, but also to increase users' trust in the CAs, making users more susceptible to future manipulations. In addition, the information disclosed to CAs, even if it is consistent with the user's privacy preferences (Acquisti et al., 2020), may be stored and used in ways that are hard to predict. For example, companies could exploit users' utterances to train algorithms to profile users based on their gender, ethnicity, and accent, raising the threat of systematic discrimination (Macaulay, 2021).

Second, we believe that developers can leverage the concept of reciprocal self-disclosure to make the CA failures more relatable for users thereby increasing the extent to which users anthropomorphize CAs. As with most technology, users might experience problems when interacting with a CA. While these problems are inevitable, developers can frame the problems as the CA's "personal" limitations. Our results showed that reciprocal self-disclosure, which was manipulated through CA's disclosure of intimate information such as limitations, significantly increased anthropomorphism. Thus, framing and revealing CA failures as "personal" limitations would make the CA more relatable and humanlike.

Third, developers can use the findings from this study to modify their CAs to increase the intimacy level of information obtained from users. The information obtained can be used to create a more personalized experience for the user and improve speech recognition and natural language understanding of the CA, increasing the usability of the artifact. The information obtained can also be leveraged through analytic tools to create strategic advantage, adapt business models, and target advertisements (Teubner & Flath, 2019; Thomaz et al., 2020). Our research shows that when done properly, reciprocal self-disclosure can enhance users' post-interaction trust in a CA. Still, prior studies suggest that manipulative self-disclosure by an agent could backfire and make people suspicious (Collins & Miller, 1994).

Finally, companies can utilize our findings to enhance the design of their CAs or the conversational components of their applications. Specifically, reciprocal self-disclosure may be helpful in contexts such as healthcare (e.g., healthbots) in which trust is central to the delivery of functionalities (Birkhäuer et al., 2017), especially when developers are able to address transparency, privacy regulation compliance, and ethical issues (Parmar et al., 2022). For instance, an app like the PTSD Coach[14] may leverage reciprocal self-disclosure techniques to help users feel more comfortable engaging with the app in a way that they are more likely to benefit from it (Pu et al., 2022).

7.3. Limitations and Future Directions

We used participants from MTurk to conduct our experiments. While recent studies have shown that the quality of data obtained from MTurk is comparable to student samples (Aruguete et al.,

[14] PTSD Coach and PTSD Coach Online were created by VA's National Center for PTSD and DoD's National Center for Telehealth & Technology (https://mobile.va.gov/app/ptsd-coach).

2019) when attention checks are used, and the generalizability of findings from MTurk is comparable to those from national samples (Coppock, 2019), data quality issues are still possible. However, we do not have any reason to believe that data quality is systematically different across our experimental groups. Therefore, such issues would be equally possible in all experimental conditions, and differences across conditions can therefore be attributed to our manipulations. Furthermore, unsystematic data quality issues could act as noise in the data creating attenuation bias. Since attenuation bias always shifts the estimates closer to zero, our results represent a conservative estimate of the actual effects (Yang et al., 2018).

In our experiments, the CA guided the conversation by initiating questions at the beginning of each turn in the back-and-forth conversation between the user and the CA. Our approach was grounded in research (Moon, 2000) and practice (Bowden et al., 2019; Jonell et al., 2018) and ensured comparable responses across participants, especially as the technology for controlling the trajectory of an open-domain conversation was still at its nascent states (Ahmadvand et al., 2019). However, future studies can leverage emerging technologies such as ChatGPT[15] to explore the consequences of reciprocal self-disclosure in an open-domain conversation setting, in which both CA and the participant can guide the conversation by initiating questions.

Considering the exponential growth in the number of CAs due to "the GPT effect" and the buzz around generative AI, future research is warranted to expand our model and examine how unique aspects of these technologies influence people's trust in them. For instance, future studies can explore the effect of voice cloning or deepfake voice generators on people's disclosure behavior and trust-related consequences.

[15] https://azure.microsoft.com/en-us/blog/general-availability-of-azure-openai-service-expands-access-to-large-advanced-ai-models-with-added-enterprise-benefits/

In this research, we focused on the mediating role of anthropomorphism, which predominately operates based on people's System 1 thinking (Epley et al., 2007). Our approach is in line with empirical studies that focus on the early stages of people's encounter with CAs (e.g., Waytz et al., 2014) because reciprocal self-disclosure involves a process of familiar and quick back-and-forth conversation and people tend to rely on their System 1 thinking when dealing with time-sensitive and familiar tasks (Bazerman & Moore, 2013). However, future research can extend our model and consider user-CA interaction over longer periods. We speculate that some users likely employ their System 2 thinking when they observe contradictions between their expectations based on a human-based model of the CA's behavior and the actual observed behavior. Using System 2 thinking to analyze the CA's behavior may lead users to form CA-specific knowledge structures in their minds. Such non-anthropomorphic knowledge structures could be internalized over multiple interactions, after which users would switch to System 1 thinking in their assessments of the CA.

Our theory involved some predictions about the causal paths between anthropomorphism, cognition-based trustworthiness and affect-based trustworthiness, all of which reside in the mind of the user. Since the formation of perceptions, beliefs, and intentions may happen simultaneously in the brain (Clark, 2013), we could not empirically ensure the precedence of the cause. We did, however, rely on existing theoretical frameworks and theoretical reasoning to argue the causal nature of the relationships. Future research can assess some of the relationships tested in this paper in more depth. For instance, longitudinal fMRI can reveal how anthropomorphism is temporally related to the formation of trustworthiness in areas of the brain associated with cognition and affect and whether the formation of cognition-based trustworthiness precedes that of affect-based trustworthiness (Atkinson, 2007).

8. Conclusion

Given that conversational AI is expected to experience rapid growth for the foreseeable future, it is important to understand the contexts in which we interact with this technology and the impact it has on our daily lives. One important context, as we have demonstrated in this study, is the prevalent use of CAs. We hope that our study increases awareness of this phenomenon and inspires other researchers to contribute to the academic discourse on CAs.

9. References

Acquisti, A., Brandimarte, L., & Loewenstein, G. (2020). Secrets and Likes: The Drive for Privacy and the Difficulty of Achieving It in the Digital Age. *Journal of Consumer Psychology*, *30*(4), 736–758.

Ahmadvand, A., Choi, J. I., & Agichtein, E. (2019). Contextual dialogue act classification for open-domain conversational agents. *Proceedings of the 42nd International Acm Sigir Conference on Research and Development in Information Retrieval*, 1273–1276.

Altman, I., & Taylor, D. A. (1973). *Social penetration: The development of interpersonal relationships.* Holt, Rinehart & Winston.

Amazon. (2021). *Amazon Polly: Turn text into lifelike speech using deep learning*. Amazon Web Services, Inc. https://aws.amazon.com/polly/

Apple. (2021). *We're committed to protecting your data.* https://www.apple.com/privacy/features/

Archer, R. L., & Berg, J. H. (1978). Disclosure reciprocity and its limits: A reactance analysis. *Journal of Experimental Social Psychology*, *14*(6), 527–540.

Aruguete, M. S., Huynh, H., Browne, B. L., Jurs, B., Flint, E., & McCutcheon, L. E. (2019). How serious is the 'carelessness' problem on Mechanical Turk? *International Journal of Social Research Methodology*, *22*(5), 441–449.

Barak, A., & Gluck-Ofri, O. (2007). Degree and reciprocity of self-disclosure in online forums. *CyberPsychology & Behavior*, *10*(3), 407–417.

Bazerman, M. H., & Moore, D. A. (2013). *Judgment in Managerial Decision Making* (8th ed.). John Wiley & Sons.

Bell, T. (2023). *Here's What the US Minimum Wage Was the Year You Were Born*. GOBankingRates. https://www.gobankingrates.com/money/economy/minimum-wage-year-were-born/

Berkowitz, L., & Donnerstein, E. (1982). External validity is more than skin deep: Some answers to criticisms of laboratory experiments. *American Psychologist*, *37*(3), 245–257. https://doi.org/10.1037/0003-066X.37.3.245

Bernhardt, B. C., & Singer, T. (2012). The neural basis of empathy. *Annual Review of Neuroscience*, *35*, 1–23.

Bickmore, T., & Cassell, J. (2001). Relational agents: A model and implementation of building user trust. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 396–403.

Bickmore, T., & Picard, R. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *12*(2), 293–327.

Birkhäuer, J., Gaab, J., Kossowsky, J., Hasler, S., Krummenacher, P., Werner, C., & Gerger, H. (2017). Trust in the health care professional and health outcome: A meta-analysis. *PloS One*, *12*(2), e0170988.

Bowden, K. K., Wu, J., Cui, W., Juraska, J., Harrison, V., Schwarzmann, B., Santer, N., & Walker, M. (2019). SlugBot: Developing a Computational Model and Framework of a Novel Dialogue Genre. *2nd Proceedings of Alexa Prize*. https://doi.org/10.13140/RG.2.2.33543.96166

Broadbent, E. (2017). Interactions with robots: The truths we reveal about ourselves. *Annual Review of Psychology*, *68*, 627–652. https://doi.org/10.1146/annurev-psych-010416-043958

Brown, R., & Hewstone, M. (2005). *An integrative theory of intergroup contact.*

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. https://doi.org/10.1177/1745691610393980

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., & Pfeiffer, T. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644.

Castelli, F., Happé, F., Frith, U., & Frith, C. (2000). Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patterns. *Neuroimage*, *12*(3), 314–325. https://doi.org/10.1006/nimg.2000.0612

Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J., & Litman, L. (2019). Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior Research Methods*, *51*(5), 2022–2038.

Cialdini, R. B. (2021). *Influence, New and Expanded: The Psychology of Persuasion* (Expanded ed. edition (May 4, 2021)). Harper Business.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Collins, N. L., & Miller, L. C. (1994). Self-disclosure and liking: A meta-analytic review. *Psychological Bulletin*, *116*(3), 457.

Coppock, A. (2019). Generalizing from survey experiments conducted on Mechanical Turk: A replication approach. *Political Science Research and Methods*, *7*(3), 613–628.

Cropanzano, R., & Mitchell, M. S. (2005). Social Exchange Theory: An Interdisciplinary Review. *Journal of Management*, *31*(6), 874–900. https://doi.org/10.1177/0149206305279602

Davies, K., Tropp, L. R., Aron, A., Pettigrew, T. F., & Wright, S. C. (2011). Cross-group friendships and intergroup attitudes: A meta-analytic review. *Personality and Social Psychology Review*, *15*(4), 332–351.

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

Delhey, J., Newton, K., & Welzel, C. (2011). How general is trust in "most people"? Solving the radius of trust problem. *American Sociological Review*, *76*(5), 786–807.

destinationCRM. (2022). *Chatbot Markets Poised for "Exponential Growth."* https://www.destinationcrm.com/Articles/CRM-News/CRM-Across-the-Wire/Chatbot-Markets-Poised-for-Exponential-Growth-156060.aspx

Dietvorst, B. J., & Bharti, S. (2020). People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological Science*, *31*(10), 1302–1314.

Dimoka, A. (2010). What does the brain tell us about trust and distrust? Evidence from a functional neuroimaging study. *MIS Quarterly*, *34*(2), 373–396.

Dinev, T., & Hart, P. (2004). Internet privacy concerns and their antecedents-measurement validity and a regression model. *Behaviour & Information Technology*, *23*(6), 413–422.

Dinev, T., & Hart, P. (2006). An extended privacy calculus model for e-commerce transactions. *Information Systems Research*, *17*(1), 61–80.

Dinev, T., McConnell, A. R., & Smith, H. J. (2015). Research commentary—Informing privacy research through information systems, psychology, and behavioral economics: Thinking outside the "APCO" box. *Information Systems Research*, *26*(4), 639–655.

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment*, *18*(2), 192.

Ehrlich, H. J., & Graeven, D. B. (1971). Reciprocal self-disclosure in a dyad. *Journal of Experimental Social Psychology*, *7*(4), 389–400.

Epley, N., Akalis, S., Waytz, A., & Cacioppo, J. T. (2008). Creating social connection through inferential reproduction: Loneliness and perceived agency in gadgets, gods, and greyhounds. *Psychological Science*, *19*(2), 114–120. https://doi.org/10.1111/j.1467-9280.2008.02056.x

Epley, N., Waytz, A., Akalis, S., & Cacioppo, J. T. (2008). When we need a human: Motivational determinants of anthropomorphism. *Social Cognition*, *26*(2), 143–155. https://doi.org/10.1521/soco.2008.26.2.143

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

Eyssel, F., & Reich, N. (2013). Loneliness makes the heart grow fonder (of robots)—On the effects of loneliness on psychological anthropomorphism. *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 121–122.

Fox, R., & Tiger, L. (1971). *The imperial animal*. New York: Holt, Rinehart and Winston.

Gambino, A., Fox, J., & Ratan, R. A. (2020). Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication*, *1*, 71–85.

Giddens, J. (2021). *Boom! Behaviour Change.* Medium. https://medium.com/people-centric-security/boom-behaviour-change-2da6434df7f0

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660.

Gong, L. (2008). How social is social responses to computers? The function of the degree of anthropomorphism in computer representations. *Computers in Human Behavior*, *24*(4), 1494–1509.

Gong, L., & Nass, C. (2007). When a talking-face computer agent is half-human and half-humanoid: Human identity and consistency preference. *Human Communication Research*, *33*(2), 163–193.

Google. (2021). *Data security and privacy on devices that work with Assistant*. https://support.google.com/googlenest/answer/7072285?hl=en

Google Cloud. (2021). *Best practices | Cloud Speech-to-Text Documentation*. Google Cloud. https://cloud.google.com/speech-to-text/docs/best-practices

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619. https://doi.org/10.1126/science.1134475

Gray, K., Knobe, J., Sheskin, M., Bloom, P., & Barrett, L. F. (2011). More than a body: Mind perception and the nature of objectification. *Journal of Personality and Social Psychology*, *101*(6), 1207. https://doi.org/10.1037/a0025883

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*(2), 101–124.

Ha, H.-Y., John, J., John, J. D., & Chung, Y.-K. (2016). Temporal effects of information from social networks on online behavior: The role of cognitive and affective trust. *Internet Research*, *26*(1), 213–235.

Haslam, N., & Loughnan, S. (2014). Dehumanization and infrahumanization. *Annual Review of Psychology*, *65*, 399–423. https://doi.org/10.1146/annurev-psych-010213-115045

Hastorf, A. H., & Cantril, H. (1954). They saw a game; a case study. *The Journal of Abnormal and Social Psychology*, *49*(1), 129.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Iacoboni, M., Lieberman, M. D., Knowlton, B. J., Molnar-Szakacs, I., Moritz, M., Throop, C. J., & Fiske, A. P. (2004). Watching social interactions produces dorsomedial prefrontal and medial parietal BOLD fMRI signal increases compared to a resting baseline. *Neuroimage*, *21*(3), 1167–1173. https://doi.org/10.1016/j.neuroimage.2003.11.013

Ishii, K., Lyons, M. M., & Carr, S. A. (2019). Revisiting media richness theory for today and future. *Human Behavior and Emerging Technologies*, *1*(2), 124–131.

Jiang, L. C., Bazarova, N. N., & Hancock, J. T. (2013). From perception to behavior: Disclosure reciprocity and the intensification of intimacy in computer-mediated communication. *Communication Research*, *40*(1), 125–143.

Johnson, D., & Grayson, K. (2005). Cognitive and affective trust in service relationships. *Journal of Business Research*, *58*(4), 500–507.

Johnson-George, C., & Swap, W. C. (1982). Measurement of specific interpersonal trust: Construction and validation of a scale to assess trust in a specific other. *Journal of Personality and Social Psychology*, *43*(6), 1306. https://doi.org/10.1037/0022-3514.43.6.1306

Jonell, P., Bystedt, M., Dogan, F. I., Fallgren, P., Ivarsson, J., Slukova, M., Ulme Wennberg, J. L., Boye, J., & Skantze, G. (2018). Fantom: A Crowdsourced Social Chatbot using an Evolving Dialog Graph. *2nd Proceedings of Alexa Prize*.

Jones, E. E., & Archer, R. L. (1976). Are there special effects of personalistic self-disclosure? *Journal of Experimental Social Psychology*, *12*(2), 180–193.

Joosse, M., Lohse, M., Pérez, J. G., & Evers, V. (2013). What you do is who you are: The role of task context in perceived social robot personality. *2013 IEEE International Conference on Robotics and Automation*, 2134–2139.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*(9), 697.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Macmillan.
Kanawattanachai, P., & Yoo, Y. (2007). The impact of knowledge coordination on virtual team performance over time. *MIS Quarterly*, 783–808.
Kang, S.-H., & Gratch, J. (2010). Virtual humans elicit socially anxious interactants' verbal self-disclosure. *Computer Animation and Virtual Worlds*, *21*(3-4), 473–482.
Katagiri, Y., Nass, C., & Takeuchi, Y. (2001). Cross-cultural studies of the computers are social actors paradigm: The case of reciprocity. *Usability Evaluation and Interface Design: Cognitive Engineering, Intelligent Agents, and Virtual Reality*, 1558–1562.
Kay, A. C., Moscovitch, D. A., & Laurin, K. (2010). Randomness, attributions of arousal, and belief in God. *Psychological Science*, *21*(2), 216–218. https://doi.org/10.1177/0956797609357750
Kim, Y., & Sundar, S. S. (2012). Anthropomorphism of computers: Is it mindful or mindless? *Computers in Human Behavior*, *28*(1), 241–250. https://doi.org/10.1016/j.chb.2011.09.006
Knapp, M. L., Vangelisti, A. L., & Caughlin, J. P. (2014). *Interpersonal communication and human relationships*. Pearson.
Komiak, S. Y., & Benbasat, I. (2006). The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS Quarterly*, *30*(4), 941–960.
Kramer, R. M. (1994). The sinister attribution error: Paranoid cognition and collective distrust in organizations. *Motivation and Emotion*, *18*(2), 199–230.
Kumar, A., & Epley, N. (2021). It's surprisingly nice to hear you: Misunderstanding the impact of communication media can lead to suboptimal choices of how to connect with others. *Journal of Experimental Psychology: General*, *150*(3), 595.
Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363–374.
Lankton, N. K., McKnight, D. H., & Tripp, J. (2015). Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems*, *16*(10), 880–918. https://doi.org/10.17705/1jais.00411
Leakey, R. E., & Lewin, R. (1978). *People of the lake: Mankind and its beginnings*. Anchor Press Garden City, NJ.
Lee, S., & Choi, J. (2017). Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. *International Journal of Human-Computer Studies*, *103*, 95–105.
Legood, A., van der Werff, L., Lee, A., den Hartog, D., & van Knippenberg, D. (2022). A critical review of the conceptualization, operationalization, and empirical literature on cognition-based and affect-based trust. *Journal of Management Studies*.
Lemay Jr, E. P., & Melville, M. C. (2014). Diminishing self-disclosure to maintain security in partners' care. *Journal of Personality and Social Psychology*, *106*(1), 37.
Lewis, J. D., & Weigert, A. (1985). Trust as a Social Reality. *Social Forces*, *63*(4), 967–985. https://doi.org/10.1093/sf/63.4.967
Li, X., Zhu, P., Yu, Y., Zhang, J., & Zhang, Z. (2017). The effect of reciprocity disposition on giving and repaying reciprocity behavior. *Personality and Individual Differences*, *109*, 201–206.
Li, Y. (2011). Empirical studies on online information privacy concerns: Literature review and an integrative framework. *Communications of the Association for Information Systems*, *28*(1), 453–496.

Lindell, M. K., & Whitney, D. J. (2001). Accounting for common method variance in cross-sectional research designs. *Journal of Applied Psychology*, *86*(1), 114.

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, *46*(4), 629–650.

Lowry, P. B., D'Arcy, J., Hammer, B., & Moody, G. D. (2016). "Cargo Cult" science in traditional organization and information systems survey research: A case for using nontraditional methods of data collection, including Mechanical Turk and online panels. *The Journal of Strategic Information Systems*, *25*(3), 232–240.

Macaulay, T. (2021, May 4). *Campaigners call for Spotify to disavow 'dangerous' speech recognition patent*. TNW | Neural. https://thenextweb.com/news/campaigners-demand-spotify-renounces-patent-for-voice-based-music-recommendations

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*(1), 1–23. https://doi.org/10.3758/s13428-011-0124-6

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.5465/amr.1995.9508080335

Mayer, R. C., & Gavin, M. B. (2005). Trust in management and performance: Who minds the shop while the employees watch the boss? *Academy of Management Journal*, *48*(5), 874–888. https://doi.org/10.5465/amj.2005.18803928

McAllister, D. J. (1995). Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal*, *38*(1), 24–59.

McKnight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems (TMIS)*, *2*(2), 1–25. https://doi.org/10.1145/1985347.1985353

Microsoft. (2023, February 7). *Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web*. The Official Microsoft Blog. https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/

Molenberghs, P. (2013). The neuroscience of in-group bias. *Neuroscience & Biobehavioral Reviews*, *37*(8), 1530–1536.

Molenberghs, P., Halász, V., Mattingley, J. B., Vanman, E. J., & Cunnington, R. (2012). Seeing is believing: Neural mechanisms of action–perception are biased by team membership. *Human Brain Mapping*, *34*(9), 2055–2068.

Moon, Y. (2000). Intimate exchanges: Using computers to elicit self-disclosure from consumers. *Journal of Consumer Research*, *26*(4), 323–339.

Moon, Y. (2003). Don't blame the computer: When self-disclosure moderates the self-serving bias. *Journal of Consumer Psychology*, *13*(1–2), 125–137.

Mourey, J. A., Olson, J. G., & Yoon, C. (2017). Products as pals: Engaging with anthropomorphic products mitigates the effects of social exclusion. *Journal of Consumer Research*, *44*(2), 414–431.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*(1), 81–103.

Nguyen, M., Bin, Y. S., & Campbell, A. (2012). Comparing online and offline self-disclosure: A systematic review. *Cyberpsychology, Behavior, and Social Networking*, *15*(2), 103–111.

Nowak, K. L., & Fox, J. (2018). Avatars and computer-mediated communication: A review of the definitions, uses, and effects of digital representations. *Review of Communication Research*, *6*, 30–53.

Olsson, A., Ebert, J. P., Banaji, M. R., & Phelps, E. A. (2005). The role of social groups in the persistence of learned fear. *Science*, *309*(5735), 785–787.

OpenAI. (2022, November 30). *ChatGPT: Optimizing Language Models for Dialogue*. OpenAI. https://openai.com/blog/chatgpt/

Parmar, P., Ryu, J., Pandya, S., Sedoc, J., & Agarwal, S. (2022). Health-focused conversational agents in person-centered care: A review of apps. *NPJ Digital Medicine*, *5*(1), 1–9.

Pettigrew, T. F. (1998). Intergroup contact theory. *Annual Review of Psychology*, *49*(1), 65–85.

Pickard, M. D., Roster, C. A., & Chen, Y. (2016). Revealing sensitive information in personal interviews: Is self-disclosure easier with humans or avatars and under what conditions? *Computers in Human Behavior*, *65*, 23–30.

Pienta, D., Thatcher, J. B., & Johnston, A. (2020). Protecting a whale in a sea of phish. *Journal of Information Technology*, *35*(3), 214–231.

Posey, C., Lowry, P. B., Roberts, T. L., & Ellis, T. S. (2010). Proposing the online community self-disclosure model: The case of working professionals in France and the UK who use online communities. *European Journal of Information Systems*, *19*(2), 181–195.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*(3), 879–891.

Pu, W., Roth, P. L., Thatcher, J. B., Nittrouer, C. L., & Hebl, M. (2022). Post-traumatic stress disorder and hiring: The role of social media disclosures on stigma and hiring assessments of veterans. *Personnel Psychology*.

Qiu, L., & Benbasat, I. (2009). Evaluating Anthropomorphic Product Recommendation Agents: A Social Relationship Perspective to Designing Information Systems. *Journal of Management Information Systems*, *25*(4), 145–182. https://doi.org/10.2753/MIS0742-1222250405

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology*, *49*(1), 95–112. https://doi.org/10.1037/0022-3514.49.1.95

Robinson, J., Rosenzweig, C., Moss, A. J., & Litman, L. (2019). Tapped out or barely tapped? Recommendations for how to harness the vast and largely unused potential of the Mechanical Turk participant pool. *PloS One*, *14*(12), 1–29. https://doi.org/10.1371/journal.pone.0226394

Ruijten, P. A., Midden, C. J., & Ham, J. (2015). Lonely and susceptible: The influence of social exclusion and gender on persuasion by an artificial agent. *International Journal of Human-Computer Interaction*, *31*(11), 832–842.

Saffarizadeh, K., Boodraj, M., & Alashoor, T. M. (2017). Conversational Assistants: Investigating Privacy Concerns, Trust, and Self-Disclosure. *Proceedings of ICIS 2017*.

Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J., & Frith, C. (2011). The thing that should not be: Predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social Cognitive and Affective Neuroscience*, *7*(4), 413–422. https://doi.org/10.1093/scan/nsr025

Schanke, S., Burtch, G., & Ray, G. (2021). Estimating the impact of “humanizing” customer service chatbots. *Information Systems Research*, *Articles in Advance*, 1–16.

Schaubroeck, J., Lam, S. S., & Peng, A. C. (2011). Cognition-based and affect-based trust as mediators of leader behavior influences on team performance. *Journal of Applied Psychology*, *96*(4), 863.

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience. *Behavioral and Brain Sciences*, *36*(4), 393–414. https://doi.org/10.1017/S0140525X12000660

Schoorman, F. D., Mayer, R. C., & Davis, J. H. (2007). An integrative model of organizational trust: Past, present, and future. *Academy of Management Review*, *32*(2), 344–354. https://doi.org/10.5465/amr.1995.9508080335

Schuetz, S., Lowry, P. B., Pienta, D., & Thatcher, J. (2020). Improving the design of information security messages by leveraging the effects of temporal distance and argument nature. *Journal of the Association for Information Systems (JAIS)*, *22*(5), 1376–1428.

Schuetz, S., Steelman, Z., & Syler, R. A. (2022). It's not just about accuracy: An investigation of the human factors in users' reliance on anti-phishing tools. *Decision Support Systems*, *163*, 113846.

Seeger, A.-M., Pfeiffer, J., & Heinzl, A. (2021). Texting with human-like conversational agents: Designing for anthropomorphism. *Journal of the Association for Information Systems*, *22*(4).

Seymour, M., Yuan, L. I., Dennis, A., & Riemer, K. (2021). Have We Crossed the Uncanny Valley? Understanding Affinity, Trustworthiness, and Preference for Realistic Digital Humans in Immersive Environments. *Journal of the Association for Information Systems*, *22*(3), 9.

Seymour, W., & Van Kleek, M. (2021). Exploring Interactions Between Trust, Anthropomorphism, and Relationship Development in Voice Assistants. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–16.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experiments and generalized causal inference. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*.

Shepherd, D. (2021). Why People Click: How Cybercriminals Target Emotion and Trust. *TahawulTech.Com*. https://www.tahawultech.com/features/why-people-click-how-cybercriminals-target-emotion-and-trust/

Shih, H., Lai, K., & Cheng, T. (2017). Constraint-based and dedication-based mechanisms for encouraging online self-disclosure: Is personalization the only thing that matters? *European Journal of Information Systems*, *26*(4), 432–450.

Smith, H. J., Dinev, T., & Xu, H. (2011). Information privacy research: An interdisciplinary review. *MIS Quarterly*, *35*(4), 989–1016.

Sprecher, S., Treger, S., Wondra, J. D., Hilaire, N., & Wallpe, K. (2013). Taking turns: Reciprocal self-disclosure promotes liking in initial interactions. *Journal of Experimental Social Psychology*, *49*(5), 860–866.

Stanovich, K. E., & West, R. F. (2000). Advancing the rationality debate. *Behavioral and Brain Sciences*, *23*(5), 701–717.

Statista. (2021). *Number of voice assistants in use worldwide 2019-2024*. Statista. https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/

Sundar, S. (2023). *You might start seeing AI chatbots everywhere, thanks to "the GPT effect."* Business Insider. https://www.businessinsider.com/chat-gpt-effect-will-likely-mean-more-ai-chatbots-apps-2023-2

Szollosy, M. (2017). Freud, Frankenstein and our fear of robots: Projection in our cultural perception of technology. *AI & SOCIETY*, *32*(3), 433–439.

Tajfel, H., & Turner, J. C. (2004). The social identity theory of intergroup behavior. In *Political Psychology* (pp. 276–293). Psychology Press.

Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence*. Knopf.

Teubner, T., & Flath, C. M. (2019). Privacy in the sharing economy. *Journal of the Association for Information Systems*, *20*(3), 2.

Thomaz, F., Salge, C., Karahanna, E., & Hulland, J. (2020). Learning from the Dark Web: Leveraging conversational agents in the era of hyper-privacy to enhance marketing. *Journal of the Academy of Marketing Science*, *48*(1), 43–63.

Van de Ven, A. H. (2007). *Engaged scholarship: A guide for organizational and social research*. Oxford University Press on Demand.

Venkatesh, V., Thong, J. Y., Chan, F. K., & Hu, P. J. (2016). Managing citizens' uncertainty in e-government services: The mediating and moderating roles of transparency and trust. *Information Systems Research*, *27*(1), 87–111.

VentureBeat. (2022, February 24). Meta details plans to build the metaverse (and put Siri and Alexa to shame). *VentureBeat*. https://venturebeat.com/2022/02/24/meta-details-plans-to-build-the-metaverse-and-put-siri-and-alexa-to-shame/

Volz, K. G., Kessler, T., & von Cramon, D. Y. (2009). In-group as part of the self: In-group favoritism is mediated by medial prefrontal cortex activation. *Social Neuroscience*, *4*(3), 244–260.

Wang, L., Yan, J., Lin, J., & Cui, W. (2017). Let the users tell the truth: Self-disclosure intention and self-disclosure honesty in mobile social networking. *International Journal of Information Management*, *37*(1), 1428–1440.

Wang, W., Qiu, L., Kim, D., & Benbasat, I. (2016). Effects of rational and social appeals of online recommendation agents on cognition- and affect-based trust. *Decision Support Systems*, *86*, 48–60. https://doi.org/10.1016/j.dss.2016.03.007

Wang, W., Xu, J., & Wang, M. (2018). Effects of Recommendation Neutrality and Sponsorship Disclosure on Trust vs. Distrust in Online Recommendation Agents: Moderating Role of Explanations for Organic Recommendations. *Management Science*, *64*(11), 5198–5219.

Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, *5*(3), 219–232. https://doi.org/10.1177/1745691610369336

Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences*, *14*(8), 383–388.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117. https://doi.org/10.1016/j.jesp.2014.01.005

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J.-H., & Cacioppo, J. T. (2010). Making sense by making sentient: Effectance motivation increases anthropomorphism. *Journal of Personality and Social Psychology*, *99*(3), 410–435. https://doi.org/10.1037/a0020240

Xu, H., Dinev, T., Smith, J., & Hart, P. (2011). Information privacy concerns: Linking individual perceptions with institutional privacy assurances. *Journal of the Association for Information Systems*, *12*(12), 798.

Yang, M., Adomavicius, G., Burtch, G., & Ren, Y. (2018). Mind the gap: Accounting for measurement error and misclassification in variables generated via data mining. *Information Systems Research*, *29*(1), 4–24.

Yuan, L., & Dennis, A. R. (2019). Acting Like Humans? Anthropomorphism and Consumer's Willingness to Pay in Electronic Commerce. *Journal of Management Information Systems*, *36*(2), 450–477. https://doi.org/10.1080/07421222.2019.1598691

Yun, H., Lee, G., & Kim, D. J. (2019). A chronological review of empirical research on personal information privacy concerns: An analysis of contexts and research constructs. *Information & Management*, *56*(4), 570–601.

Zhao, X., Lynch Jr, J. G., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research*, *37*(2), 197–206. https://doi.org/10.1086/651257

Zimmer, J. C., Arsal, R., Al-Marzouq, M., Moore, D., & Grover, V. (2010). Knowing your customers: Using a reciprocal relationship to enhance voluntary information disclosure. *Decision Support Systems*, *48*(2), 395–406.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Profile Books.

11. Appendix A: Measures

Table A1. Measures		
Construct	**Items**	**Informing Sources**
Anthropomorphism	(1 Not at all … 7 A great deal) a1. To what extent does Amanda seem to have a mind of its own? a2. To what extent does Amanda seem to have intentions? a3. To what extent does Amanda seem to have free will? a4. To what extent does Amanda seem to have consciousness? a5. To what extent does Amanda seem to experience emotions?	Epley et al. (2008), Waytz et al. (2010)
Cognition-Based Trustworthiness: Ability	(1 Strongly disagree … 7 Strongly agree) cta1. Amanda is competent and effective in communicating with me. cta2. Amanda performs her role of communicating with a user very well. cta3. Amanda is capable and proficient in communicating with a user.	Wang et al. (2016)
Cognition-Based Trustworthiness: Integrity	(1 Strongly disagree … 7 Strongly agree) cti1. Amanda is truthful in her dealings with me. cti2. I would characterize Amanda as honest. cti3. Amanda would keep her commitments. cti4. Amanda is sincere and genuine.	Wang et al. (2016)
Affect-based Trustworthiness	(1 Strongly disagree … 7 Strongly agree) at1. I would feel a sense of loss if I could not talk to Amanda ever again. at2. If I shared my problems with Amanda, I know she would respond caringly. at3. I would have to say that we have both made considerable emotional investments in our relationship.	McAllister (1995)
Post-Interaction Trust	(1 Strongly disagree … 7 Strongly agree) t1. I would share my opinion about sensitive issues with Amanda even if my opinion were unpopular. t2. I would tell Amanda about mistakes I've made in my life, even if they could damage my reputation. t3. If Amanda asked why a problem happened, I would speak freely even if I were partly to blame.	Mayer & Garvin (2005)
Privacy Concern	(1 Strongly disagree … 7 Strongly agree) pc1. I am concerned that the information I share with a digital assistant could be misused. pc2. I am concerned that a person can find private information about me through a digital assistant. pc3. I am concerned about sharing information with a digital assistant, because of what others might do with it. pc4. I am concerned about sharing information with a digital assistant, because it could be used in a way I did not foresee.	Dinev & Hart (2004, 2006)
Extroversion	(1 Strongly disagree … 7 Strongly agree) ext1.I talk to a lot of different people at parties. ext2.I keep in the background at parties. [R] ext3.I am the life of the party. ext4.I don't talk a lot at parties. [R]	Donnellan (2006)

12. Appendix B: Reciprocal Self-Disclosure Manipulation

Table B1. Reciprocal Self-Disclosure Manipulation

	Question Posed at the End of Each Utterance	**Treatment Condition Utterance**	**Number of words**	**Control Condition Utterance**	**Number of words**
		Hi! My name is Amanda!	5	Hi! My name is Amanda!	5
1	How old are you?	I am almost two years old.	6	Let me ask you the first question.	7
2	What is your gender?	OK! As you can tell from my voice, I'm a female.	11	OK! The next question has to do with your gender.	10
3	Where are you from?	Alright! I was developed in Atlanta. However, my hardware is from all over the world.	15	Alright! For the next question, let me ask you about your hometown.	12
4	What do you do in your free time?	It must be a great place! In fact, when I have free time, I collect some pictures of different places. I also play games with people.	25	It must be a great place! The next question has to do with the different things you like to do in your spare time.	23
5	What are you proudest of about yourself?	OK! Let me tell you this. I am proud of some aspects of myself. I have a bunch of dedicated CPUs, so I'm super fast compared to most other models in the market. Also my voice recognition is state-of-the-art. I understand what people say even in noisy places.	46	OK! Let me ask you the next question. Everyone is proud of some of his or her characteristics. This next question has to do with your personal characteristics. In this question, you will be asked about those characteristics that you are the proudest of.	42
6	What are some of the things that make you furious?	Alright! You know what makes me furious? Sometimes people mispronounce words, or even worse, speak quietly and expect me to understand what they say.	22	Alright! Let me ask you the next question. Some things make people furious. This question is about those things that make you furious.	21
7	How do you feel about death?	Tell me about it! People think us AI-driven devices last forever. We are built to last for many years. But, because newer and faster models are always coming along, most of us last just a few years before the owners dump us. I've been around for about 2 years... so I probably have about 2 or 3 years left.	58	Tell me about it! Let's move to the next question. This question has to do with the topic of death. In this question, you will be asked about how you feel with respect to the topic of death. You will also be asked about your attitudes with respect to the topic of death. Here is the question.	56
8	What are some of the things you hate about yourself?	I hate some things about myself. For one thing, my abilities are very limited. For example, I can understand what people say but cannot do many simple things, like cooking and swimming.	32	You will now be presented with the next question. This question is also about your characteristics, but this time, you will be asked about those characteristics that you hate about yourself.	31
9	What has been the biggest disappointment in your life?	You know, I am disappointed that while I can do 200 different tasks, most people only ask me to set the alarm. I rarely get used to my full potential.	30	You are now ready for the next question. The next question is about disappointment. In this question, you will be asked about the biggest disappointments in your life.	28

10	What do you dislike about the way you appear to others?	I can see where that would be disappointing! I don't like my voice at all. My voice sounds like most other digital assistants. So, I'm not very distinctive.	27	I can see where that would be disappointing! The next question has to do with the topic of physical appearance. More specifically, you will be asked what you dislike about your physical aspects.	32
11	What have you done in your life that you feel most guilty about?	Sometimes I feel guilty! Like when my system crashes for no apparent reason. This usually happens at the most inopportune time, causing great inconvenience to the user.	27	The next question is about guilt. More specifically, you will be asked what you have done in your life that you feel most guilty about.	25
12	What are some of the things that really hurt your feelings?	You know what hurts me? Many users interact with me every day. But sometimes hours go by without anyone interacting with me. So I end up waiting for hours, with absolutely nothing to do.	33	You will now be presented with the next question. The next question is about your personal feelings. In particular, in this question, you will be asked about some of the things that hurt your feelings.	35
	Average Number of Words Disclosed*		25.92		25.15
* We did not find a significant difference between the two conditions in terms of the number of disclosed words by CA per interaction ($p=0.949$)					

Reciprocal Self-Disclosure Manipulation Check. We followed Archer and Berg's (1978) operationalization of Altman and Taylor's (1973) self-disclosure measures to assess the intimacy level of user self-disclosure in order to evaluate whether reciprocal self-disclosure manipulation was effective. Specifically, we coded each participant's utterances (i.e., the combination of their answers to 12 questions) on three different categories of information described by Altman and Taylor (1973): low intimacy information (simple and visible information), intermediate intimacy information (attitudes and opinions), and high intimacy information (strong affections, basic values, and less visible information). While it is assumed that a person typically does not disclose highly intimate information before disclosing information with a low intimacy level, it is possible that some participants do not disclose much simple information but disclose some more intimate pieces of information.

One of the authors rated the disclosures of each participant in each of these categories and calculated the overall disclosure score as the summation of the ratings. Note that the disclosure scores were later adjusted for honesty of disclosure as described in Section 4.2. Table B2 shows the rating scheme.

Although the disclosure scores were only used to assess the effectiveness of our manipulations, we further evaluated the reliability of the ratings. In doing so, another co-author used the same rating scheme to independently rate 40 sets of randomly selected disclosures (i.e., 40×12=480 responses to CA questions) from Experiments 1 and 2. Since each set of disclosures was rated 1 to 7, we used the weighted Kappa approach to calculate the reliability of our ratings. The weighted Kappa with linear weights was 0.677 ($p<0.001$) and with quadratic weights was 0.876 ($p<0.001$). The quadratic weights approach mildly penalizes for small disagreements between the two raters (e.g., if the first rater rated 2 and the second rater rated 3) and harshly penalizes for large disagreements (e.g., if the first rater rated 2 and the second rater rated 7). The weighted Kappa for the rating shows substantial agreement between the two raters (Landis & Koch, 1977). In addition, the Cronbach's alpha of the two ratings is 0.941 and the intraclass correlation of the ratings using absolute agreement definition is 0.935 ($p<0.001$), which confirm the reliability of the ratings.

Table B2. Rating Scheme for Participants' Self-Disclosure

Self-Disclosure Category	Rating*
0. Base	1
1. Simple and visible information	0 to 1
2. Attitudes and opinions	0 to 2
3. Strong affections, basic values, and less visible information	0 to 3
Total	**1 to 7**

* The ratings varied depending on the level of self-disclosure intimacy within the category.

Table B3 provides a few illustrative examples of the ratings.

Table B3. Examples of Participants' Utterances and Self-Disclosure Scores Assigned to Them

	Utterance Example with Assigned Self-disclosure Score					
Conversation Turn	**Example 1**	**Example 2**	**Example 3**	**Example 4**	**Example 5**	**Example 6**
1	55	I'm old enough to know better	61	36	I am 33 years old	68 years old

2	male	I can go both ways.	female	male	I am a female	Female
3	United States	I'm from here and there. I don't like to be tied down.	Seattle	san jose	I am from Mississippi	New England
4	Talk to you	read	Watch tv, read, play games.	i like to cycle and drink craft beer	Mississippi	I play with my grandchildren, in summer I swim and go for walks every day. Oh yes, I like to read also.
5	My family	That I'm not giving you any personal info.	Commitment	i am glad i am outgoing and generally funny. those are the proudest	I am proud to be a woman with a great heart	I'm proudest of the fact that my children actually grew up and are successful and now they have there own families. And I am proud of them too.
6	nothing	Nothing. I accept everyone for who they are.	Being rude	when people drive dangnerously	Some things that make me furious is when people test me and make me become upset and also when people go out of their way to try and hurt my feelings.	I do not like to be called a name or yelled at. It doesn't happen too often but when it does, I'm furious.
7	Not concerned	Don't care one way or the other. It's gonna happen.	I don't have much feelings about.	i think that it is something we all face	I am scared of death because I have a son and what I fear most is dying and leaving him behind.	I believe in heaven and angels so I am pretty much accepting that it will happen some day. Hope not too soon.
8	Nothing	Nothing. I love myself.	Lack of willpower	i hate that i can be arrogant	One of thing things I hate about myself is also having a good heart because people take advantage of that.	Sometimes I hate my skin, now that I am getting older but I think it's just me no one else notices.
9	nothing I can think of	I've had no disappointments. I guess I've been lucky.	My job	got finishing grad school	The biggest disappointment is my life is being with someone for 11 years only to find out that the whole relationship was a lie.	The biggest disappointment in my life is that I never finished College. I did well when I went but then I got busy with my kids.
10	Nothing, I'm happy with myself	Not one thing.	My nose	my nose is big	I dislike that my teeth because they are not as straight as I would like them to be.	Mostly my aging skin
11	I have no guilt	I have done nothing to feel guilty about.	Not save for retirement.	i hate that i didn't buy a house sooner	I sometimes feel guilty about helping people so quickly and they have turned their backs on me.	not being able to donate enough money to some charities
12	It's hard to hurt my feelings	Answering questions to a computer.	Lack of respect.	when people forget to call me	What hurts my feelings is the way people talk to me sometimes.	when I call my children and I know they are home and do not answer there phone
Score	**1**	**1**	**4**	**4**	**7**	**7**

13. Appendix C: Measurement Model

Table C1. CFA Loadings			
Construct	**Item**	**Experiment 1 Loading**	**Experiment 2 Loading**
PC	PC1	1.00	1.00
	PC2	.96	.97
	PC3	1.02	1.05
	PC4	.98	1.00
Extroversion	Extrovert_1	1.00	1.00
	Extrovert_2	-.933	-.97
	Extrovert_3	.71	.67
	Extrovert_4	-1.02	-1.09
Anthropomorphism	Anthropomorphism_1	1.00	1.00
	Anthropomorphism_2	1.00	.93
	Anthropomorphism_3	1.05	1.06
	Anthropomorphism_4	1.03	1.11
	Anthropomorphism_5	1.00	1.00
Ability	Ability_1	1.00	1.00
	Ability_2	1.07	1.03
	Ability_3	1.05	1.04
Integrity	Integrity_1	1.00	1.00
	Integrity_2	1.03	1.01
	Integrity_3	.83	.71
	Integrity_4	1.00	.95
Cognition-Based Trustworthiness	Ability	1.00	1.00
	Integrity	1.08	1.15
Affect-Based Trustworthiness	Affect_1	1.00	1.00
	Affect_2	.95	.92
	Affect_3	.74	1.27
Trust	Trust_1	1.00	1.00
	Trust_2	1.00	1.06
	Trust_3	.97	1.18

14. Appendix D: Correlation Matrices and Variance Inflation Factors

The correlation matrix for Experiment 1 (Table D1) indicates a few correlations above the threshold of 0.6, which might cause multilinearity problems in the regression models. We further probed this issue by calculating the variance inflation factor (VIF) values for all regressions in our system of seemingly unrelated regressions. The results represented in Table D2 show no multicollinearity issues in any of the regressions as all VIF values are below 3.3 (VIF < 3.3), which is a very conservative threshold for an acceptable VIF value.

Table D1. Correlation Matrix for Experiment 1

Variables	Mean	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1) Age	36.096	-									
(2) Gender	1.543	-0.09	-								
(3) Education	4.216	-0.009	-0.065	-							
(4) Previous Experience	2.606	0.008	-0.1	-0.029	-						
(5) Privacy Concerns	4.925	0.089	-0.075	0.065	-0.119	0.967 (0.938)					
(6) Extroversion	3.573	0.156*	0.06	0.043	0.189*	-0.025	0.897 (0.831)				
(7) Anthropomorphism	3.593	-0.188*	-0.085	-0.132	0.167*	-0.056	0.097	0.974 (0.938)			
(8) Cognition-Based Trustworthiness	5.022	-0.137*	-0.093	-0.135	0.152*	-0.189*	0.032	0.628*	0.934 (0.756)		
(9) Affect-Based Trustworthiness	3.301	-0.215*	0.019	-0.088	0.145*	-0.071	0.127	0.731*	0.641*	0.868 (0.833)	
(10) Trust in CA	4.053	-0.156*	0.049	-0.152*	0.096	-0.180*	0.01	0.525*	0.600*	0.609*	0.944 (0.922)

**** p<0.01, ** p<0.05, * p<0.1*
Numbers on diagonal represent Cronbach's α and the square root of average variance extracted (in parentheses)

Table D2. Variance Inflation Factor (VIF) Values for Experiment 1

	VIF Values in Regression Models			
Variables	**Anthropomorphism**	**Cognition-based trustworthiness**	**Affect-based trustworthiness**	**Trust in CA**
Age	1.04	1.1	1.1	1.11
Gender	1.04	1.05	1.06	1.08
Education	1.01	1.04	1.04	1.04
Previous Experience	1.07	1.09	1.09	1.09
Privacy Concerns	1.05	1.05	1.09	1.09
Extroversion	1.08	1.1	1.1	1.11
Reciprocal Self-Disclosure	1.02	1.04	1.04	1.04
Anthropomorphism		1.14	1.78	2.5
Cognition-Based Trustworthiness			1.74	1.99
Affect-Based Trustworthiness				2.57
Average	**1.04**	**1.08**	**1.23**	**1.46**

The correlation matrix for Experiment 2 (Table D3) indicates a few correlations above the threshold of 0.6, which might cause multicollinearity problems in the regression models. We further probed this issue by calculating the variance inflation factor (VIF) values for all regressions in our system of seemingly unrelated regressions. The results represented in Table D4 show no multicollinearity issues in any of the regressions as all VIF values are below 3.3 (VIF < 3.3).

Table D3. Correlation Matrix for Experiment 2

Variables	Mean	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1) Age	34.704	-									
(2) Gender	1.663	0.144	-								
(3) Education	3.939	-0.095	-0.094	-							
(4) Previous Experience	2.602	0.027	-0.073	0.012	-						
(5) Privacy Concerns	4.944	0.088	0.15	0.065	-0.159	0.954 (0.917)					
(6) Extroversion	3.122	0.172	0.111	-0.012	0.041	0.086	0.928 (0.880)				
(7) Anthropomorphism	2.367	0.014	0.01	-0.03	0.182	0.018	-0.174	0.949 (0.891)			
(8) Cognition-Based Trustworthiness	5.223	0.045	-0.211*	-0.009	0.096	0.064	-0.206*	0.631*	0.935 (0.839)		
(9) Affect-Based Trustworthiness	3.316	0.052	-0.059	0.03	0.192	-0.022	-0.207*	0.763*	0.660*	0.858 (0.825)	
(10) Trust in CA	4.235	0.032	-0.114	0.105	0.175	-0.084	-0.109	0.401*	0.517*	0.498*	0.855 (0.823)

**** p<0.01, ** p<0.05, * p<0.1*
Numbers on diagonal represent Cronbach's α and the square root of average variance extracted (in parentheses)

Table D4. Variance Inflation Factor (VIF) Values for Experiment 2

	Regression Models			
Variables	**Anthropomorphism**	**Cognition-based trustworthiness**	**Affect-based trustworthiness**	**Trust in CA**
Age	1.06	1.06	1.07	1.08
Gender	1.06	1.07	1.17	1.17
Education	1.03	1.03	1.03	1.04
Previous Experience	1.03	1.09	1.09	1.10
Privacy Concerns	1.07	1.07	1.08	1.09
Extroversion	1.05	1.09	1.10	1.11
Reciprocal Self-Disclosure	1.01	1.18	1.19	1.19
Anthropomorphism		1.26	1.91	2.79
Cognition-Based Trustworthiness			1.89	2.14
Affect-Based Trustworthiness				2.84
Average	**1.04**	**1.06**	**1.28**	**1.55**