Summer 8-11-2020

Conversational AI Agents: Investigating AI-Specific Characteristics that Induce Anthropomorphism and Trust in Human-AI Interaction

Kambiz Saffarizadeh
Georgia State University

CONVERSATIONAL AI AGENTS:
INVESTIGATING AI-SPECIFIC CHARACTERISTICS THAT INDUCE ANTHROPOMORPHISM AND TRUST IN HUMAN-AI INTERACTION

BY

KAMBIZ SAFFARIZADEH

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree

Of

Doctor of Philosophy

In the Robinson College of Business

Of

Georgia State University

GEORGIA STATE UNIVERSITY
ROBINSON COLLEGE OF BUSINESS
2020

ACCEPTANCE

This dissertation was prepared under the direction of the *KAMBIZ SAFFARIZADEH'S* Dissertation Committee. It has been approved and accepted by all members of that committee, and it has been accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Business Administration in the J. Mack Robinson College of Business of Georgia State University.

Richard Phillips, Dean

DISSERTATION COMMITTEE

Dr. Mark Keil (Chair)
Dr. Likoebe Maruping
Dr. J.J. Po-An Hsieh
Dr. Aaron M. Baird
Dr. Lingyao (Ivy) Yuan

ABSTRACT

CONVERSATIONAL AI AGENTS: INVESTIGATING AI-SPECIFIC CHARACTERISTICS THAT INDUCE ANTHROPOMORPHISM AND TRUST IN HUMAN-AI INTERACTION

BY

KAMBIZ SAFFARIZADEH

May 18, 2020

Committee Chair: *Dr. Mark Keil*

Major Academic Unit: *Computer Information Systems*

The investment in AI agents has steadily increased over the past few years, yet the adoption of these agents has been uneven. Industry reports show that the majority of people do not trust AI agents with important tasks. While the existing IS theories explain users' trust in IT artifacts, several new studies have raised doubts about the applicability of current theories in the context of AI agents. At first glance, an AI agent might seem like any other technological artifact. However, a more in-depth assessment exposes some fundamental characteristics that make AI agents different from previous IT artifacts. The aim of this dissertation, therefore, is to identify the AI-specific characteristics and behaviors that hinder and contribute to trust and distrust, thereby shaping users' behavior in human-AI interaction. Using a custom-developed conversational AI agent, this dissertation extends the human-AI literature by introducing and empirically testing six new constructs, namely, AI indeterminacy, task fulfillment indeterminacy, verbal indeterminacy, AI inheritability, AI trainability, and AI freewill.

DEDICATION

To my mother and grandmother.

ACKNOWLEDGMENTS

"To wisely live your life, you don't need to know much. Just remember two main rules for the beginning: You better starve, than eat whatever. And better be alone, than with whoever." – Omar Khayyam

First and foremost, I would like to thank my committee chair, mentor, friend, and role model Dr. Mark Keil for his guidance and support throughout the Ph.D. program. Mark is not only an excellent scholar but also a great human being. He is the reason I joined the Ph.D. program at GSU, and the reason I was able to finish my dissertation. He pushed me beyond my limits and made me grow as a scholar. I am forever grateful for the opportunity I had to work with him.

I would also like to thank my committee members: Dr. Likoebe Maruping, who gave me great feedback on my research ideas, Dr. J.J. Po-An Hsieh, who shaped my core view of the Information Systems field, Dr. Aaron Baird, who helped me refine my conceptualizations, and Dr. Lingyao Yuan, who gave me useful feedback on my work on the anthropomorphism construct. I would like to acknowledge the continuous help and support I received from Dr. Wael Jabr early in the Ph.D. program. I am also thankful to Dr. Daniel Robey, Dr. Arun Rai, Dr. Balasubramaniam Ramesh, Dr. Edward Rigdon, Dr. Yi Zhao, Dr. William Robinson, and Dr. Pierre Nguimkeu for teaching me how be a better researcher, and Dr. Seyyed Babak Alavi, Dr. Ali Naghi Mashayekhi, Dr. Satish Nargundkar, Dr. Shuguang Hong, and Mr. Ahmadian, who inspired me to be a better teacher.

I owe my curiosity about computer information systems to Farid Fooladi. Farid was our neighbor and a computer engineering student at the University of Tehran when I was eight years old. He was the person who set up my first computer. He always told me, "Do whatever you want inside the operating system. Explore everything. Don't worry; you won't break it." I am grateful for all the weekends he spent reinstalling the operating systems I destroyed while encouraging me to continue exploring.

One of my best and worst experiences while working on my dissertation was when I had to go to the hospital with level five emergency during ICIS 2018 in San Francisco – worst because of the long time it took me to recover and get back to research, and best because of the amount of support I received from Dr. Mark Keil, my colleagues, and my family. My heartful thanks go to Tawfiq Alashoor and Pengcheng Wang who took me to the urgent care and Yumeng Miao, my dearest friend who took me to the hospital. I would also like to thank Alireza, Roxanne, and Ariana Jafari, who nursed me back to health.

One of the most fruitful experiences of doing a Ph.D. for me was spending time with the incredibly talented Ph.D. students and discussing research ideas. My special thanks go to my fellow Ph.D. students Tawfiq, Yumeng, Maheshwar, Mahdi, Xiaocong, Christine, Yanran, Amber, Alan, Amrita, Joshua, Zirun, Zhitao, Arun, Sayed, Youyou, Neetu, Shane, Langtao, Vitali, Jessica, Hyoungyong, Pengcheng, Yukun, Heeseung, Junyoung, Khaleed, Peiwei, Sophia, Jeremy, Wei, Fengyuan, Hengqi, Yuting, Jing, Weifang, Roberto, Elizabeth, Bahadir, and Avishek.

I would like to thank my sister, Atena Saffarizadeh, who encouraged me to pursue a Ph.D. degree, and my father, Reza Saffarizadeh, whose memory influenced my life in many ways. I would also like to thank Behzad, Zari, Payam, Pedram, Ali, and Ryan, who gave me emotional support throughout the Ph.D. program, and Behrooz and Azizollah, whose only wish for me was to become a scientist.

Last but not least, I would like to thank my mother, Behnaz Feizbakhsh, and grandmother, Aghdas Jafari, who single-handedly raised my sister and me. They gave up their youth so we could thrive. They are examples of women throughout history who sacrificed their own lives to see their children succeed. Any positive thing that I have ever done and any positive thing that I will do is just an imperfect reflection of their pure perfection.

Table of Contents

Chapter 1: Introduction

Industry reports have projected that AI could contribute up to $15.7 trillion to the global economy by 2030 (PwC 2017). Throughout the world, large investments (over $400 thousand) in AI startups have grown exponentially, from $1.3 billion in 2010 to $40.4 billion in 2018, with over 3000 companies receiving more than $400 thousand in funding in 2018 (AI Index 2019). It is projected that companies will increase their investments in AI up to three-fold by 2020 (Forrester Research 2017). However, despite all the investments and excitement around AI (Naudé 2019), many AI projects have not delivered the expected results. According to the Gartner 2020 CIO Agenda survey, far fewer companies successfully implemented AI systems in 2019 than expected (Miller 2019). Gartner Vice President Svetlana Sicular stated that "something is stalling AI adoption" (Miller 2019). While many possible technical factors hinder the adoption of AI systems (e.g., buggy machine learning algorithms, and low quality of training data), practitioners believe that a crucial reason is a lack of focus on the customer experience and users' perception of the AI (Miller 2019). Several reports indicated that users do not trust AI agents (Longoni and Morewedge 2019; Miller 2019; Shattuck 2019; Towers-Clark 2019) and are not willing to delegate important tasks to them (Pew Research Center 2017).

Even though many experts acknowledge the lack of users' trust in AI agents, most of the current efforts in AI have been concentrated on development-side factors such as enhancing machine learning methods, training machine learning specialists, and optimizing the hardware and software that run the AI (Costello 2020). However, research on the user-side on human-AI interaction is inadequate.[1] There remains a need for studies that explore how users' perception of unique

[1] Based on the released statistics of AAAI, which is one of the longest running AI conferences that provides a broad coverage of AI topics, only 26 out of 1,148 accepted (169 out of 7,745 submitted) papers were related to human and AI or human-AI collaboration (AI Index 2019).

characteristics of AI shape users' trust and distrust in AI agents, and consequently drive their behavior in their interaction with such agents.

While at first glance, an AI agent might seem like any other technological artifact, a more in-depth assessment exposes some fundamental characteristics that make AI agents different from previous IT artifacts. First, the stochasticity in many machine learning methods that power AI agents (e.g., reinforcement learning (Mnih et al. 2015), generative adversarial networks (Goodfellow et al. 2014), and stochastic gradient descent used in most deep learning methods (Goodfellow et al. 2016)) makes their behavior inherently unpredictable for users. Similarly, the context-awareness of some AI agents (e.g., SlugBot on Amazon's Alexa (Bowden et al. 2019)) makes the behavior of the agent dependent on its surroundings, which is again inherently unpredictable. Many people perceive this type of inherent unpredictability as a unique characteristic of agents with free will (e.g., humans, God, etc.), because pure objects tend to show deterministic behavior (Ebert and Wegner 2011; Kay et al. 2010; Waytz et al. 2010). Second, AI agents often show autonomous choice-making capabilities. For instance, Duplex, a conversational agent developed by Google, can book appointments after going through a complex conversation with a person over the phone (Leviathan 2018). Again, most people believe that choice-making capability is a unique characteristic of agents with free will (Feldman et al. 2014). In summary, AI, unlike traditional artifacts, is capable of doing tasks that were traditionally reserved exclusively for humans (Brynjolfsson and McAfee 2014).

The majority of prior research in the information systems field has assumed that humans interact with two types of individual-level entities: humans and non-humans (e.g., IT artifacts). The underlying implicit assumption of such research is that there is a clear distinction between a human and a non-human. Nonetheless, as discussed, several AI characteristics challenge this assumption

(Schuetz and Venkatesh 2020). We postulate that the advent of artificial intelligence has created a continuum between a human and an object, with AI standing between the two.

In this dissertation, we investigate the factors that hinder and contribute to trust and distrust, and shape users' behavior in human-AI interaction. Specifically, using a custom-developed conversational AI agent, we conducted three studies to extend our understanding of human-AI interaction. Conversational agents are suitable for studying AI agents due to their widespread presence in our daily lives. In fact, industry reports suggest that about 3.25 billion conversational AI agents were in use at the beginning of 2019 (Voicebot.ai 2019), and it is estimated that this number will rise to 8 billion by 2023 (JuniperResearch 2018). According to Gartner, by 2023, 25 percent of employee interactions with applications will happen via voice (Miller 2019).

This dissertation is comprised of three essays presented in chapters 2 to 4, followed by a conclusion in chapter 5. Below we briefly introduce each of the three essays.

ESSAY 1

In the first essay, we study the phenomenon of AI indeterminacy and its effect on trust. We define AI indeterminacy as the unpredictability in the AI's behavior that seems not to have a directly observable cause. As AI artifacts become more complicated, users face more indeterminacies in their interactions with the artifacts. These indeterminacies have important effects on users' perception of the artifacts. In this research, we identified verbal and task fulfillment indeterminacies as two important indeterminacies in the context of conversational AI agents. We define verbal indeterminacy as perceived indeterminant variation in the way an agent conveys a given message (i.e., by using different choices of words and grammar), and task fulfillment

indeterminacy as perceived indeterminant variation in an agent's behavior of fulfilling a user command (e.g., by producing erroneous outcome).

Verbal and task fulfillment indeterminacies could influence the user's assessment of the humanness of the artifact. Since being a human is the thing we know the best (Broadbent 2017), we often implement human-based concepts to understand apparently unpredictable agents (Waytz et al. 2010). The mere implementation of a human-based prediction model could increase anthropomorphism (Epley et al. 2007). Consequently, verbal and task fulfillment indeterminacies could act as anthropomorphic signals.

While some scholars have investigated the effect of indeterminacy on anthropomorphism (e.g., Salem et al. 2013; Waytz et al. 2010), its effect on trust is still unclear. On the one hand, indeterminacy as a source of unreliability should, by definition, decrease trust (Mayer et al. 1995). In line with this notion, several studies have shown that indeterminacy in AI's behavior (e.g., in the form of erroneous outcome) is detrimental to users' trust in AI (Dietvorst et al. 2015; Dzindolet et al. 2003; Manzey et al. 2012). On the other hand, many studies have shown a positive effect of task fulfillment indeterminacy (in the form of erroneous outcome) on anthropomorphism (Salem et al. 2013), as well as a positive effect of anthropomorphism on trust in AI (Waytz et al. 2014). However, task fulfillment indeterminacy has not been found to have a significant negative effect on users' trusting behavior (Salem et al. 2015). In a recent review of empirical papers on trust in AI, Glikson and Woolley concluded that in the context of AI agents "low reliability [as a manifestation of high indeterminacy] does not always lead to low trust," and that "future research should further explore the reasons for the positive emotional reaction toward imperfect functioning anthropomorphic [AI] robots" (2020, p. 51).

Furthermore, it is unclear how indeterminacy influences trust in the presence of multiple signals of indeterminacy (e.g., verbal and task fulfillment indeterminacies). When more than one anthropomorphic signal is present, the prediction based on one signal might not match the other. The mismatch between prediction and observation can produce a large feedback error (called a "surprisal" in neuroscience literature) (Clark 2013) that might lead to the rejection of the whole idea of applying a human-based mental model to understand the agent's behavior (Burleigh et al. 2013; Saygin et al. 2011). Therefore, it is important to understand the interplay between different types of indeterminacy (i.e., verbal indeterminacy and task fulfillment indeterminacy) to address the tension between anthropomorphism and trust. In this research, we seek to answer two research questions:

> **RQ1:** What are the effects of verbal and task fulfillment indeterminacies on anthropomorphism and trust?
>
> **RQ2:** What is the interaction effect of verbal and task fulfillment indeterminacies on anthropomorphism and trust?

ESSAY 2

In the second essay, we study the transference of users' distrust in the creator of an AI agent to their distrust in the AI agent. We also discuss the factors that can mitigate this transference. Distrust is often regarded as a defensive mechanism to protect oneself against possible harmful actions of the other party (McKnight et al. 2004; Yang et al. 2015). As anecdotal evidence and industry reports show (Berlatsky 2018; Pew Research Center 2017), many users distrust AI agents and perceive them as malevolent agents striving to take over humanity.

While previous research identified perceived intentions of the trustee (i.e., the entity to be trusted or distrusted) to be central in shaping distrusting beliefs (Dimoka 2010; McKnight and Chervany 2001), it is not clear how users perceive AI agents' intentions. Most of the extant research has focused on intention in either human-human interactions, in which the trustee is perceived to have volition, or human-technology interaction, in which the technology is regarded as a tool and assumed to "lack volition and moral agency" (McKnight et al. 2011, p. 5). However, we argue that some of the underlying assumptions of the extant literature break down in the context of AI agents. First, distrust could be formed based on the users' perception of not only an agent itself but also the entity who is responsible for the observed behavior of the agent. Second, in user-artifact interaction, the user is not the only entity with volition. When the artifact can inherit intentions of other agents, such as its creator, there is a discernible "will" in the artifact's behavior that potentially helps shape users' distrust in the artifact. Third, users might view an artifact and its creator as a single entity with homogenous characteristics. Finally, a dichotomous approach to volition, based on which an entity either has complete volition or has no volition, ignores the possibility of a spectrum between pure objects and pure autonomous beings (humans). In the context of AI agents, the artifacts move from being mere objects toward what might be considered independent creatures, but they are neither traditional objects nor humans.

We speculate that users might construct their distrust based on their perception of the moral agent responsible for the artifact's behavior. A parsimonious set of responsible agents in the context of human-AI interaction includes the creator (i.e., the entity that has created the AI agent), the creature (i.e., the AI agent), and the user (i.e., the person who interacts with the AI agent). Using the analogy of a human offspring, we postulate that the behavior of an AI agent can be inherited from its parent (creator), learned through upbringing (training), and based on its own freewill.

First, if the user perceives that the creature inherited its values from its creator (i.e., AI inheritability), then the creator is responsible for its behavior. Second, if the artifact is trainable by the user (AI trainability), then the user is responsible for its behavior. Finally, if the user believes that the artifact has freewill (AI freewill), then the artifact is responsible for its own behavior (Gray et al. 2012).

In this essay, we seek to answer the following research questions:

> **RQ1**: What is the relationship between distrust in the creator of an AI agent and distrust in the AI agent itself?
>
> **RQ2**: What are the moderating effects of perceived AI characteristics (i.e., inheritability, trainability, and freewill) on this relationship?

ESSAY 3

In the third essay, we study users' information disclosure to AI agents, which is an important behavioral outcome of trust. More specifically, we seek to examine reciprocal self-disclosure in the context of human-conversational-AI interaction by investigating the cognitive and affective bases of users' self-disclosure. Prior research has found robust evidence of reciprocity in human-human interactions. The tendency to reciprocate is so important that some scholars have mentioned it as a central characteristic of being human (Fox and Tiger 1971). Prior studies showed that reciprocal self-disclosure extends to human-computer interaction because people perceive computers as social actors (CASA) (Moon 2000). There is little research, however, on the underlying mechanism that makes people reciprocate an AI agent's behavior.

We argue that a plausible explanation for reciprocal self-disclosure in human-AI interaction is that people use a human-based mental model to understand why the artifact shares information about

itself, a behavior that is core to humanness. The adoption of a human-based mental model necessarily means anthropomorphism. Prior research indicated that anthropomorphism could fulfill some cognitive and affective needs associated with understanding the agent's behavior and creating a social connection with the agent (Epley et al. 2007). Therefore, we argue that by anthropomorphizing the agent, the user develops cognitive and affect-based assessments of the trustworthiness of the agent. These cognitive and affective bases of trust can then lead to a willingness to make oneself vulnerable to the actions of the agent by disclosing information about the self.

To address the shortcoming in the current literature and assess the validity and soundness of our reasoning, we seek to study the following research question:

> **RQ:** What roles do anthropomorphism and trust play in reciprocal self-disclosure in the context of conversational agents?

REFERENCES

Accenture. 2016. "Artificial Intelligence in Healthcare | Accenture." (https://www.accenture.com/us-en/insight-artificial-intelligence-healthcare, accessed April 8, 2020).

AI Index. 2019. "The AI Index Annual Report," AI Index.

Berlatsky, N. 2018. "Is AI Dangerous? Why Our Fears of Killer Computers or Sentient 'Westworld' Robots Are Overblown." (https://www.nbcnews.com/think/opinion/ai-dangerous-why-our-fears-killer-computers-or-sentient-westworld-ncna943111, accessed June 14, 2019).

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

Broadbent, E. 2017. "Interactions with Robots: The Truths We Reveal about Ourselves," *Annual Review of Psychology* (68), pp. 627–652. (doi: 10.1146/annurev-psych-010416-043958).

Brynjolfsson, E., and McAfee, A. 2014. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*, WW Norton & Company.

Burleigh, T. J., Schoenherr, J. R., and Lacroix, G. L. 2013. "Does the Uncanny Valley Exist? An Empirical Test of the Relationship between Eeriness and the Human Likeness of Digitally Created Faces," *Computers in Human Behavior* (29:3), pp. 759–771.

Clark, A. 2013. "Whatever next? Predictive Brains, Situated Agents, and the Future of Cognitive Science," *Behavioral and Brain Sciences* (36:3), pp. 181–204. (doi: 10.1017/S0140525X12000477).

Costello, K. 2020. "Gartner Predicts the Future of AI Technologies." (//www.gartner.com/smarterwithgartner/gartner-predicts-the-future-of-ai-technologies/, accessed March 19, 2020).

Dietvorst, B. J., Simmons, J. P., and Massey, C. 2015. "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err.," *Journal of Experimental Psychology: General* (144:1), American Psychological Association, p. 114.

Dimoka, A. 2010. "What Does the Brain Tell Us about Trust and Distrust? Evidence from a Functional Neuroimaging Study," *Mis Quarterly*, pp. 373–396.

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

Ebert, J. P., and Wegner, D. M. 2011. "Mistaking Randomness for Free Will," *Consciousness and Cognition* (20:3), pp. 965–971. (doi: 10.1016/j.concog.2010.12.012).

Epley, N., Waytz, A., and Cacioppo, J. T. 2007. "On Seeing Human: A Three-Factor Theory of Anthropomorphism.," *Psychological Review* (114:4), pp. 864–886. (doi: 10.1037/0033-295X.114.4.864).

Feldman, G., Baumeister, R. F., and Wong, K. F. E. 2014. "Free Will Is about Choosing: The Link between Choice and the Belief in Free Will," *Journal of Experimental Social Psychology* (55), pp. 239–245. (doi: 10.1016/j.jesp.2014.07.012).

Forrester Research. 2017. "Predictions 2017: Artificial Intelligence Will Drive The Insights Revolution," Forrester.

Fox, R., and Tiger, L. 1971. *The Imperial Animal*, New York: Holt, Rinehart and Winston.

Glikson, E., and Woolley, A. W. 2020. "Human Trust in Artificial Intelligence: Review of Empirical Research," *Academy of Management Annals*.

Goodfellow, I., Bengio, Y., and Courville, A. 2016. *Deep Learning*, MIT press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. 2014. "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems*, pp. 2672–2680.

Gray, K., Young, L., and Waytz, A. 2012. "Mind Perception Is the Essence of Morality," *Psychological Inquiry* (23:2), pp. 101–124.

JuniperResearch. 2018. "Digital Voice Assistants in Use to Triple to 8 Billion by 2023." (https://www.juniperresearch.com/press/press-releases/digital-voice-assistants-in-use-to-triple, accessed March 3, 2019).

Kay, A. C., Moscovitch, D. A., and Laurin, K. 2010. "Randomness, Attributions of Arousal, and Belief in God," *Psychological Science* (21:2), pp. 216–218. (doi: 10.1177/0956797609357750).

Kinsella, B. 2019. "Voice Startup Funding in 2019 Set to Nearly Triple Says European VC Mangrove and 'Voice Economy' to Be a Trillion Dollar Market in 2025," *Voicebot.Ai*. (https://voicebot.ai/2019/07/19/voice-startup-funding-in-2019-set-to-nearly-triple-over-2018-says-european-vc-mangrove/, accessed April 8, 2020).

Leviathan, Y. 2018. "Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone," *Google AI Blog*. (http://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html, accessed February 6, 2019).

Longoni, C., and Morewedge, C. K. 2019. "AI Can Outperform Doctors. So Why Don't Patients Trust It?," *Harvard Business Review*. (https://hbr.org/2019/10/ai-can-outperform-doctors-so-why-dont-patients-trust-it).

Manzey, D., Reichenbach, J., and Onnasch, L. 2012. "Human Performance Consequences of Automated Decision Aids: The Impact of Degree of Automation and System Experience," *Journal of Cognitive Engineering and Decision Making* (6:1), Sage Publications Sage CA: Los Angeles, CA, pp. 57–87.

Mayer, R. C., Davis, J. H., and Schoorman, F. D. 1995. "An Integrative Model of Organizational Trust," *Academy of Management Review* (20:3), pp. 709–734. (doi: 10.5465/amr.1995.9508080335).

McKnight, D. H., Carter, M., Thatcher, J. B., and 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), p. 12.

McKnight, D. H., and Chervany, N. L. 2001. "Trust and Distrust Definitions: One Bite at a Time," in *Trust in Cyber-Societies*, Springer, pp. 27–54.

McKnight, D. H., Kacmar, C. J., and Choudhury, V. 2004. "Dispositional Trust and Distrust Distinctions in Predicting High-and Low-Risk Internet Expert Advice Site Perceptions," *E-Service* (3:2), pp. 35–58.

Miller, M. J. 2019. "Gartner: The Present and Future of Artificial Intelligence," *PCMAG*. (https://www.pcmag.com/news/gartner-the-present-and-future-of-artificial-intelligence, accessed March 19, 2020).

Mirnig, N., Stollnberger, G., Miksch, M., Stadler, S., Giuliani, M., and Tscheligi, M. 2017. "To Err Is Robot: How Humans Assess and Act toward an Erroneous Social Robot," *Frontiers in Robotics and AI* (4), Frontiers, p. 21.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., and Ostrovski, G. 2015. "Human-Level Control through Deep Reinforcement Learning," *Nature* (518:7540), p. 529. (doi: 10.1038/nature14236).

Moon, Y. 2000. "Intimate Exchanges: Using Computers to Elicit Self-Disclosure from Consumers," *Journal of Consumer Research* (26:4), pp. 323–339.

Naudé, W. 2019. "AI's Current Hype and Hysteria Could Set the Technology Back by Decades," *The Conversation*. (http://theconversation.com/ais-current-hype-and-hysteria-could-set-the-technology-back-by-decades-120514, accessed April 8, 2020).

Pew Research Center. 2017. "Many Americans Would Be Hesitant to Use Various Automation Technologies," *Pew Research Center: Internet, Science & Tech*. (http://www.pewinternet.org/2017/10/04/automation-in-everyday-life/pi_2017-10-04_automation_0-02/, accessed April 23, 2018).

PwC. 2017. "PwC's Global Artificial Intelligence Study: Sizing the Prize," *PwC*. (https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html, accessed April 8, 2020).

Ragni, M., Rudenko, A., Kuhnert, B., and Arras, K. O. 2016. "Errare Humanum Est: Erroneous Robots in Human-Robot Interaction," in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, IEEE, pp. 501–506.

Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., and Joublin, F. 2013. "To Err Is Human (-like): Effects of Robot Gesture on Perceived Anthropomorphism and Likability," *International Journal of Social Robotics* (5:3), pp. 313–323.

Salem, M., Lakatos, G., Amirabdollahian, F., and Dautenhahn, K. 2015. "Would You Trust a (Faulty) Robot? Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust," in *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, pp. 1–8.

Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J., and 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), pp. 413–422. (doi: 10.1093/scan/nsr025).

Schuetz, S., and Venkatesh, V. 2020. "The Rise of Human Machines: How Cognitive Computing Systems Challenge Assumptions of User-System Interaction," *Journal of the Association for Information Systems* (Forthcoming).

Shattuck, S. 2019. "People Don't Trust AI. We Need to Change That.," *Medium*, , February 26. (https://towardsdatascience.com/people-dont-trust-ai-we-need-to-change-that-d1de5a4a0021, accessed March 15, 2020).

Towers-Clark, C. 2019. "80% Of People Don't Trust AI With Money - How Can We Fix Its Image?," *Forbes*. (https://www.forbes.com/sites/charlestowersclark/2019/01/15/80-of-people-dont-trust-ai-with-money-how-can-we-fix-its-image/, accessed March 15, 2020).

Voicebot.ai. 2019. "Juniper Estimates 3.25 Billion Voice Assistants Are in Use Today, Google Has About 30% of Them," *Voicebot*, , February 14. (https://voicebot.ai/2019/02/14/juniper-estimates-3-25-billion-voice-assistants-are-in-use-today-google-has-about-30-of-them/, accessed March 3, 2019).

Waytz, A., Heafner, J., and Epley, N. 2014. "The Mind in the Machine: Anthropomorphism Increases Trust in an Autonomous Vehicle," *Journal of Experimental Social Psychology* (52), pp. 113–117. (doi: 10.1016/j.jesp.2014.01.005).

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J.-H., and Cacioppo, J. T. 2010. "Making Sense by Making Sentient: Effectance Motivation Increases Anthropomorphism.," *Journal of Personality and Social Psychology* (99:3), pp. 410–435. (doi: 10.1037/a0020240).

Yang, J., Sia, C.-L., and Ou, C. X. 2015. "Identify the Antecedents of Distrust in a Website.," in *PACIS*, p. 108.

Chapter 2:

AI Indeterminacy in Conversational Agents:

Investigating Anthropomorphism and Trust

Abstract

The exponential advancement of machine learning in the past decade has enabled ordinary technology users to interact with conversational AI agents (e.g., Amazon's Alexa, Apple's Siri, and Google Assistant) on a daily basis. Meanwhile, factors such as the stochastic nature of AI have made the behavior of modern conversational agents appear indeterministic. We define AI indeterminacy as unpredictability in the AI's behavior that seems not to have a directly observable cause. While prior research suggests that unpredictability in an agent's behavior increases perceived humanness (i.e., anthropomorphism), indeterminacy should, by definition, decrease trust. This apparent paradoxical effect of AI indeterminacy on perceived humanness and trust leads to a dilemma for AI developers. Some developers choose to make the behavior of AI agents artificially deterministic to increase perceived reliability and trustworthiness of the agent at the cost of making the agent seem less humanlike. However, we argue that not all AI indeterminacies are created equal. Leveraging the specific context of conversational AI agents, we introduce two types of AI indeterminacy, namely verbal indeterminacy and task fulfillment indeterminacy. We draw on perspectives in neuroscience, social psychology, and information systems to theorize the effects of verbal indeterminacy and task fulfillment indeterminacy on anthropomorphism and trust.

Using a custom-developed conversational AI agent, we conduct a randomized experiment and show that both verbal indeterminacy and task fulfillment indeterminacy increase anthropomorphism and have positive indirect effects on trust (mediated by anthropomorphism). Our results suggest that only task fulfillment indeterminacy has a direct negative effect on trust and is the real source of the tension between anthropomorphism and trust. We also find that the interaction of verbal indeterminacy and task fulfillment indeterminacy negatively influences anthropomorphism and trust (fully mediated by anthropomorphism). Our results reveal that the effect of AI indeterminacy on anthropomorphism and trust depends on the type of indeterminacy. We discuss the implications of our findings for research on human-AI interaction and for developers of AI agents.

Keywords: Artificially Intelligent Agent, AI Indeterminacy, Verbal Indeterminacy, Task Fulfillment Indeterminacy, Anthropomorphism, Trust

INTRODUCTION

With the recent advances in machine learning, we see a prevalence of conversational AI agents (e.g., Amazon's Alexa, Apple's Siri, and Google Assistant) in our daily lives (Columbus 2018). Across different platforms, about 3.25 billion conversational AI agents (hereafter conversational agents) were in use at the beginning of 2019 (Voicebot.ai 2019), and it is estimated that by 2023 this number will rise to 8 billion (JuniperResearch 2018). Conversational agents provide a wide variety of benefits for both ordinary and business users. For instance, Duplex, a conversational agent developed by Google, can help users book appointments with offline businesses by calling the business contact over the phone and autonomously navigating through complex conversations with humans (Leviathan 2018). Similarly, Alexa for Business, a service that enables organizations to use Amazon's conversational agent, can help employees set up and find empty rooms for ad-hoc business meetings, access upcoming events on their calendars, and access corporate applications via voice (Amazon 2019).

Despite the benefits, the use of AI agents has been uneven. According to industry reports, most people do not trust them to do important tasks (Longoni and Morewedge 2019; Pew Research Center 2017; Shattuck 2019; Towers-Clark 2019). While users seek reliability and certainty in AI agents, the behavior of many AI agents seem to be indeterministic. We define *AI indeterminacy* as unpredictability in the AI's behavior that seems not to have a directly observable cause. For instance, in 2018, many users reported that their Alexa, a conversational agent developed by Amazon, laughed at them when they asked her to do a task. Later Amazon stated that "in rare circumstances, Alexa can mistakenly hear the phrase 'Alexa, laugh,'" when users ask for other things, and thus the laughs were only a manifestation of false-positive errors in speech recognition (Chokshi 2018).

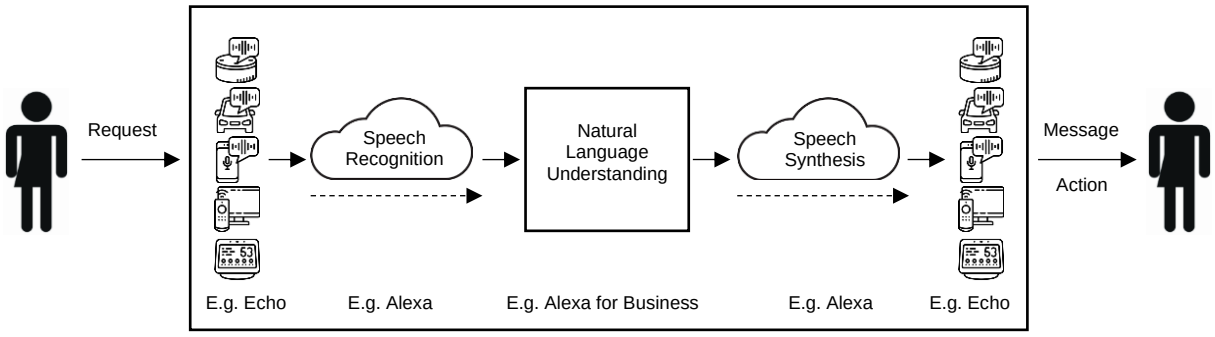
Recently in a discussion about how Alexa was created, Rohit Prasad, the Vice President and Head Scientist of Alexa Machine Learning mentioned that "in those days, there was still a lot of emphasis on rule-based systems … but we had a statistical-first approach … where for our language understanding we had … an entity recognizer and an intent classifier which was all trained statistically. In fact, we had to build the deterministic matching as follow-up to fix bugs that statistical models have. So, it was just a different mindset" (Fridman 2019).

The seemingly indeterministic behavior of AI agents is not limited to Alexa, but fundamental to how modern AI systems work. Unlike traditional symbolic AI systems (e.g., rule-based expert systems) whose behavior is deterministic, the behavior of many modern AI agents appears to be indeterministic due to factors such as randomness in the process of stochastic gradient descent in the retraining process of many deep learning models (Goodfellow et al. 2016), stochastic exploratory actions in reinforcement learning (Mnih et al. 2015), random input in generative models (Goodfellow et al. 2014), context awareness of some models (e.g., Bowden et al. 2019), and the sheer algorithmic complexity of compound models (e.g., Chen et al. 2019).

Extant literature suggests that people tend to attribute indeterministic behavior to the agent's free will (Ebert and Wegner 2011), which is a vital part of the perception of humanlike state of mind (Gray et al. 2007). As possession of a humanlike state of mind is the single most important attribute of humans, indeterminacy increases anthropomorphism (Epley, Waytz, et al. 2008; Kay et al. 2010; Waytz, Morewedge, et al. 2010), which is the process of perceiving nonhumans as humans (Epley et al. 2007). Therefore, it is reasonable to suggest that AI indeterminacy can act as an anthropomorphic signal and help conversational agents seem more humanlike. However, indeterminacy should, by definition, decrease trust because trust is essentially based on an assessment of the predictable positive behavior of the trustee (Mayer et al. 1995). This apparent

paradoxical effect of AI indeterminacy on anthropomorphism and trust leads to a dilemma for AI developers. Some developers might choose to make the behavior of AI agents artificially deterministic or entirely rely on symbolic AI to increase perceived reliability, interpretability, and trustworthiness of the agent at the cost of making the agent seem less humanlike (i.e., more "robotic").

However, we argue that not all AI indeterminacies are created equal. In the context of conversational AI, artifacts typically possess language capabilities to communicate with users and task fulfillment capabilities to fulfill requests such as controlling smart devices at home or office (Seeger et al. 2018). Leveraging the specific context of conversational agents, we introduce two types of AI indeterminacy, namely verbal indeterminacy and task fulfillment indeterminacy (see Figure 1).



When the user requests something from the agent, the message is received by the device. Then the device sends the message to the speech recognition service. Next, the detected text is sent to the natural language understanding (NLU) server. Next, the response code from NLU goes to the speech synthesis service to produce a voice response. Finally, the response code goes back to the device, and the device takes an action and delivers a response message based on the response code. Therefore, in responding to a user request, the AI delivers both a verbal message and an action, which create the possibility for two types of indeterminacy: verbal indeterminacy and task fulfillment indeterminacy.

Figure 1. AI Indeterminacy in Conversational AI

First, we define *verbal indeterminacy* as perceived indeterminant variation in the way an agent conveys a given message (i.e., by using different choices of words and grammar). Many of today's AI systems benefit from advancements in natural language processing methods to understand and

generate textual information, and advancements in speech recognition and speech synthesis methods to convert speech to text as well as to generate humanlike voice (Seeger et al. 2018). Such a powerful and complex conversational capability of AI leads to indeterminacy in AI's use of language because the AI can generate different sentences based on factors that are not apparent to the user. For instance, when verbal indeterminacy is high, on any given day a conversational agent such as Amazon's Alexa might respond differently to the same question about the weather than it did the day before, even if the weather forecast for the two days is the same.

Second, we define *task fulfillment indeterminacy* as perceived indeterminant variation in an agent's behavior of fulfilling a user command. In theory, task fulfillment indeterminacy can be viewed as indeterminacy in how the agent fulfills the command (process) or indeterminacy in whether the agent fulfills the command (outcome). In this research, we adopt the latter view because it is more commonly observed in the context of conversational AI agents. Task fulfillment indeterminacy is widespread in AI agents as the stochasticity in machine learning methods and complex back-end systems that power the agents make it harder for users to predict the agent's behavior regarding task fulfillment. An AI agent's behavior might be influenced by many different factors that could potentially counteract with one another. Even when a set of deterministic rules or trained models govern an AI's behavior, the coupling of potentially counteracting factors can make the behavior seemingly indeterministic (Levy 1994; Oestreicher 2007; Thietart and Forgues 1995). These factors include privacy and security regulations that represent constraints on AI behavior, software requirements and limitations that define the range of possible AI behavior, as well as users' and developers' expectations regarding what behaviors are acceptable in a given context. For instance, a user might ask Amazon's Alexa to turn on the light, but when task fulfillment indeterminacy is high, Alexa might or might not fulfill the request. Note that the lack of fulfillment could be due to

an error in the AI, conflicting requirements in the system, a problem in the local network, or the AI's apparently autonomous decision. However, regardless of the reason, task fulfillment indeterminacy is about a user's perception of the probabilistic nature of the fulfillment.

The interaction of these two indeterminacies as signals of anthropomorphism might lead to conflicts in users' perception. A body of research on the uncanny valley indicates that the presence of multiple signals of anthropomorphism could lead to surprising results (Broadbent 2017). Mori (1970) introduced the concept of the uncanny valley when he observed that humanlikeness increases acceptability to a certain point, but that this relationship breaks down if an entity looks very close to but is not quite humanlike. However, the relationship between humanlikeness and acceptability becomes positive again once we move beyond the uncanny valley and the entity seems quite humanlike. The sudden decrease and increase in the relationship create a valley-like region in the relationship plot. Many researchers suggested that the main reason for the uncanny valley is due to conflicting signals about humanness (Broadbent 2017). Based on some signals, such as an extremely humanlike face, the observer categorizes the artifact as a human being, but a slight mismatch, for instance, between the way a human moves her/his lips and the way the artifact does so, leads to conflicting signals that are so strong it makes the artifact be perceived as eerie. Therefore, when an artifact contains multiple sources of anthropomorphic signals, the interaction effect among them can be important.

Understanding the effect of different types of AI indeterminacy and their interaction on anthropomorphism and trust is central to our understanding of how users perceive and react to the seemingly indeterministic behaviors of modern AI agents. To the best of our knowledge, however, no previous research has studied such effects in the context of conversational AI or elsewhere. In this research, we first introduce the concepts of verbal and task fulfillment indeterminacies and

then leverage them to study the effect of indeterminacy on trust. We also investigate the role of anthropomorphism in this relationship. Hence, we seek to address the following research questions:

> **RQ1:** What are the effects of verbal and task fulfillment indeterminacies on anthropomorphism and trust?
>
> **RQ2:** What is the interaction effect of verbal and task fulfillment indeterminacies on anthropomorphism and trust?

We developed a custom conversational agent to investigate our research questions experimentally. We used a sample of 152 technology users who had some experience using digital technology to examine the soundness of our theoretical conjectures about the phenomenon. We randomly assigned the participants to different experimental conditions. Randomized experiments are the gold standard of internal validity as they provide a robust way of assessing causal relationships. The artifact that we developed resembles conversational agents, such as Google Assistant and Apple's Siri, that users regularly interact with. This makes the experiment more realistic than a hypothetical scenario in which the user has no direct interaction with the software artifact. Creation and use of a working conversational agent allowed us to create a task environment that is engaging for participants and which has a high degree of psychological realism (Berkowitz and Donnerstein 1982).

By answering the above research questions, we contribute to the literature in four ways. First, we introduce and elaborate on the concept of AI indeterminacy, suggesting that there are two specific types of it in the context of conversational AI, namely verbal indeterminacy and task fulfillment indeterminacy. Second, we explain why the interaction of multiple sources of indeterminacy in AI could lead to counterintuitive results regarding anthropomorphism and trust due to conflicting

signals. Such an explanation is a good fit for real-world AI systems because most such systems include multiple sources of indeterminacy that make their behavior hard to predict. Third, we add to the trust literature by identifying AI-specific drivers of trust in the context of conversational AI (i.e., verbal indeterminacy and task fulfillment indeterminacy). Finally, we highlight the importance of considering indeterminacy in the wider literature of technology adoption and use by investigating the direct and indirect effects of indeterminacy on trust.

THEORETICAL DEVELOPMENT

The overarching idea of this paper is that when a user faces indeterminacy in the behavior of a conversational agent, in order to understand the indeterminacy, s/he leverages mental models of how humans behave because these are the best models that humans possess to deal with indeterminacy. The very process of leveraging human-based mental models leads the user to understand the agent's behavior in human terms, which promotes anthropomorphism. We theorize that while AI indeterminacy in general increases anthropomorphism, some types of indeterminacy increase trust whereas other types of indeterminacy decrease trust, thus creating an interesting tension between anthropomorphism and trust.

In this section, we briefly discuss the relevant literature that informs our theoretical development and elaborate on the logic of our hypotheses. Figure 2 provides a summary of our research model and hypotheses.

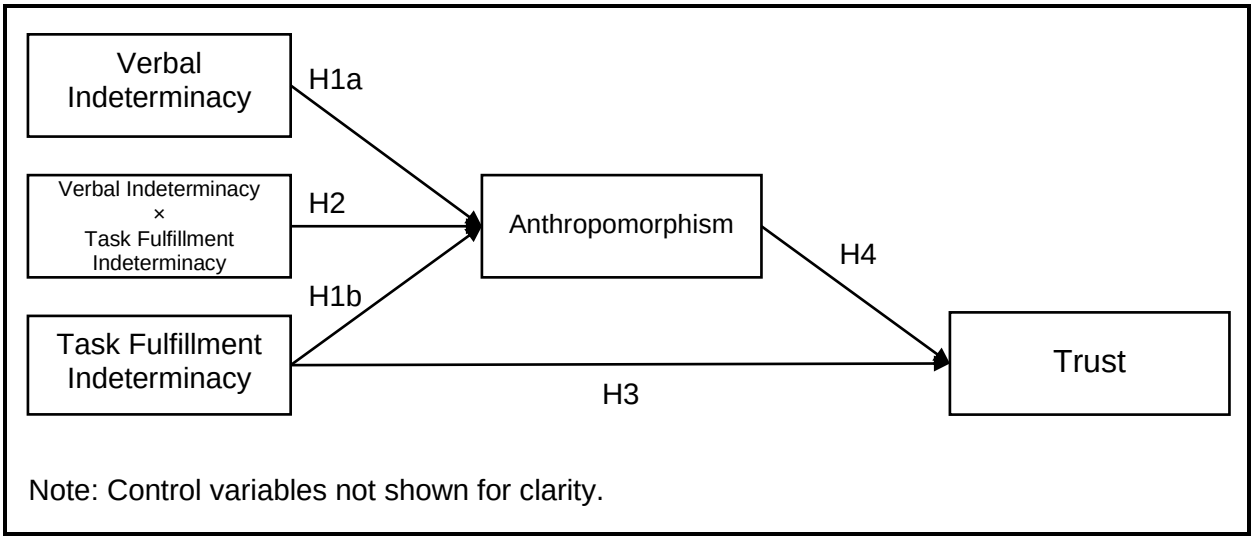


Figure 2. Research Model

Indeterminacy

The brain is essentially a prediction machine (Clark 2013) and "has evolved to represent or infer the causes of changes in its sensory inputs" (Friston 2005, p. 815). One could argue that most of what the human brain does is to extract patterns out of its apparently stochastic surroundings that include the environment and people (Clark 2013). Humans strive to explain away the seemingly stochastic noise (ε) in different phenomena. However, in most cases there will be some leftover variance that cannot be explained by the known facts. The concept of indeterminacy has been used to describe the behavioral variance that cannot be explained (Monroe et al. 2014).

If any observed behavior has a cause and the cause, in turn, has its own causes, then any observed behavior is completely deterministic. However, this chain of causality breaks at some point due to either (1) human bounded rationality and selective perception (Leonardi 2008; Orton and Weick 1990) or (2) belief in such factors as free will that justify the unexplained variance beyond the known factors (Monroe et al. 2014). However, in either case, the observed behavior will not be statistically different from a random behavior drawn from an unknown distribution of all possible

behaviors (Ebert and Wegner 2011). In fact, many studies have shown that people find meaning, perceive intention, and detect patterns even in purely random events (Caruso et al. 2010; Ebert and Wegner 2011; Oskarsson et al. 2009) partially due to their misconception of chance and randomness (Kahneman and Tversky 1972).

The way people perceive the unexplained variance in others' behavior and the way they try to explain it away is core to our understanding of how people anthropomorphize other agents (Epley et al. 2007).

Anthropomorphism

Scholars in different disciplines have used several different terms (e.g., humanness, humanlikeness, personhood, anthropomorphism, and humanization) to capture the presence of human characteristics, or the perception of such characteristics, in nonhuman entities. Appendix A provides an interdisciplinary summary of prior research on anthropomorphism.

Based on the extant literature, we define anthropomorphism as an 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; Gray et al. 2007). Therefore, first, anthropomorphism is not the mere use of human adjectives to describe the physical aspects of nonhumans; it involves going beyond observable characteristics of the entity and making inference about its unobservable characteristics. Second, anthropomorphism is different from animism. Animism is about treating an object as living, while anthropomorphism goes beyond that and entails attributing uniquely humanlike characteristics to it. Third, anthropomorphism is about people’s tendency to *perceive* human traits in nonhuman agents.

Questions regarding the accuracy of this perception and whether a nonhuman entity should be treated as human are orthogonal to anthropomorphism (Epley, Waytz, et al. 2008).

In strong forms of anthropomorphism, the person truly believes that the nonhuman entity possesses humanlike characteristics. Nevertheless, in weak forms of anthropomorphism, the person does not truly believe but acts as-if the nonhuman entity has human characteristics (e.g., cursing at a machine) (Epley, Waytz, et al. 2008).

Because the human brain uses the same neural system to make judgments about both humans and anthropomorphized nonhumans (Castelli et al. 2000; Iacoboni et al. 2004), to understand anthropomorphism, it is essential to understand what makes people view an entity as human.

Many scholars have argued that possession of a humanlike state of mind is the core of humanness (Waytz, Morewedge, et al. 2010). Perceived mind (i.e., the extent to which an entity is perceived to have a mind of its own) consists of two dimensions (Gray et al. 2007). First, the capacity for *agency*, which includes such attributes as self-control, morality, memory, emotion recognition, planning, communication, and thought. Second, the capacity for *experience*, which includes such attributes as hunger, fear, pain, pleasure, rage, desire, personality, consciousness, pride, embarrassment, and joy. According to Gray et al. (2007) these dimensions map to Aristotle's distinction between moral agents, who can do morally right or wrong behavior and be held responsible for what they did, and moral patients, who are the recipients of right or wrong behavior, and have moral rights and privileges. In the context of technological artifacts, when a person uses the artifacts in the desired way, the focus is on the experience dimension of the artifact. However, when the user expects the artifact to perform a task, the focus is on its agency.

The lay view of humanness, especially the agency dimension, is closely tied to possession of free will, which could be perceived as possession of conscious mind and spirit (Dennett 2017; Shepherd

2012). Free will is the capacity to have chosen otherwise. Most people believe that humans have free will (Feldman et al. 2016; Monroe et al. 2014; Sarkissian et al. 2010). Many studies in the philosophy field strived to explain whether free will exists or it is just an illusion (Bode et al. 2014). However, the overwhelming belief in free will, regardless of its soundness, "suggests that the mind operates in a way that encourages the inference that one's actions are freely chosen" (Ebert and Wegner 2011, p. 966). In fact, according to the theory of apparent mental causation (Wegner 2008; Wegner and Wheatley 1999), because intention and action are always temporally conjunct, people think that their conscious mind is the cause of their actions (Ebert and Wegner 2011). After the fact, people remember their intention to act before the act (Libet et al. 1983). Free will belief is fundamental to our concept of self (Bode et al. 2014) and therefore to our concept of humanness of other agents (Schilbach et al. 2013).

Impact of Indeterminacy on Anthropomorphism

When faced with a phenomenon, the brain first strives to make an inference about it through leveraging existing concepts and mental models. When the existing knowledge fails, the brain creates new concepts or updates the existing ones (Clark 2013). Since being human is the thing we know best (Broadbent 2017), the brain often attempts to make predictions about the other entity assuming that the entity is similar to oneself. In fact, as cited by Epley et al. (2008), Charles Darwin argued that we need to anthropomorphize other animals if we want to understand them (1872/2009).

Research on children indicated that, early in life, they anthropomorphize a wide variety of nonhuman agents, and only later they develop more sophisticated concepts of other agents (Carey 1985; Inagaki and Hatano 1987). Unlike children, adults often possess more sophisticated mental

models to understand the variance in the behavior of nonhuman entities (e.g., conversational agents) (Waytz, Morewedge, et al. 2010). Nevertheless, even in adults, detecting humanlike behavior in nonhuman agents – albeit unconsciously – activates mirror neurons which make them experience the same state, as if they were the agent, in order to understand and predict the behavior of the agent (Epley et al. 2007). Furthermore, neuroscientists found evidence suggesting that people anthropomorphize objects to understand their apparently intentional motions. They found that areas in the brain associated with theory of mind (i.e., the ability to attribute mental states to oneself and others) were more active when participants observed objects involved in apparently intentional motions (Castelli et al. 2000; Heberlein and Adolphs 2004; Martin and Weisberg 2003; Pelphrey et al. 2004; Waytz, Morewedge, et al. 2010).

Humans are motivated to understand the cause of observed behaviors in the environment (Kelley 1967; Lombrozo 2006) to increase their own chance of survival (White 1959). Dispositional attribution, i.e., attribution of an effect to the internal characteristics of an agent, addresses part of people's need for prediction of their surroundings (Pittman and Pittman 1980). However, an entity with free will can generate behavioral variance that cannot be explained by any factor other than the entity's own choice or desire. So, the variance that cannot be attributed to any known external or internal factors is attributed to an agent's free will.[1]

Empirical studies showed that people, regardless of their personal views of free will, perceive a probabilistic choice as free will. In fact, when the evidence supports indeterminacy in an entity's behavior, people tend to perceive the behavior as freely chosen (Ebert and Wegner 2011).

[1] This is different from dispositional attribution because, in dispositional attribution, the internal attribution does not imply that the person believes the agent could have chosen otherwise. Dispositional attribution does not always entail attributing uniquely human attributes. Therefore, "dispositional attributions are necessary but insufficient for anthropomorphism" (Waytz, Morewedge, et al. 2010, p. 416).

However, the indeterminacy could stem from such seemingly irrelevant things as pure randomness. “Because randomness is a kind of indeterminacy, people may mistakenly interpret randomness in behavior as owing to free will” (Ebert and Wegner 2011, p. 966). Empirical evidence in the context of computerized animated agents suggests that people perceive the agents to have higher free will when the agents follow a random sequence of actions instead of a predetermined one (Ebert and Wegner 2011).

In the context of our research, when a user interacts with a conversational agent, s/he strives to understand its behavior. Being able to predict the artifact’s behavior gives the user the ability to minimize threats and maximize opportunities in interacting with the artifact. However, humans have limited ways to understand uncertainty in their surroundings. The ultimate well-trained mental model that any person possesses is the model about her/his own perceptions, beliefs, intentions, and behaviors (Broadbent 2017). This model is not only in charge of a person’s own behaviors (Clark 2013), it is also used to make sense of others’ behaviors (Schilbach et al. 2013).

When a conversational agent uses different words or grammar every time it needs to communicate with the user, it sends a signal to the user that there exists a variance in the agent’s behavior that needs to be understood. Since the variance seems indeterminant, the user needs to use a human-based model to predict how the agent will communicate in the future. In this way, the concepts and attributes associated with humans will become activated under conditions of high verbal indeterminacy. Thus, we hypothesize that:

> *H1a: Verbal indeterminacy in a conversational agent’s behavior increases anthropomorphism when other types of indeterminacy are not present.*

When a conversational agent shows variance in fulfilling the requested tasks, the user tries to make sense of the behavior. If the agent fails or succeeds in fulfilling the task all the time with no

variance, then the behavior is easy to understand. Moreover, if the user can see an obvious pattern in the task fulfillment behavior, again the behavior is completely determined. However, when the reason for the variance is not easy to understand, the user needs to use his/her more advanced mental models to make sense of the behavior. Since other humans also show task fulfillment indeterminacy in their behavior, a human-based model of the behavior might help explain it. Hence, by seeing the agent as more human, the user increases his/her chance to properly model the indeterminacy and better predict an agent's future behavior. Therefore, we hypothesize that:

> *H1b: Task fulfillment indeterminacy in a conversational agent's behavior increases anthropomorphism when other types of indeterminacy are not present.*

People tend to attribute more free will to an agent with more behavioral randomness only when the randomness could be contextually meaningful. Therefore, when the action makes sense in the context, it is "possible for [even] wholly determined actions to appear freely chosen" (Ebert and Wegner 2011, p. 970). When the user interacts with a conversational agent, s/he stretches her/his imagination to some extent and attributes indeterministic behaviors to the agent's free will and possession of mind. In other words, when the user has enough room to make causal sense of the agent's behavior based on other contextual clues, s/he perceives even a purely random behavior as intentional. Otherwise, the behavior seems random and mindless.

We argue that a task fulfillment indeterminacy increases anthropomorphism despite possible undesirable consequences because the user attributes the random obedience and disobedience of the artifact to an intentional decision-making process. However, when both task fulfillment and verbal indeterminacies are observed at the same time, the user can use the information from both observations to make an inference about the cause of the indeterminacy. A human-based model of language to understand an AI's verbal indeterminacy is valid in the absence of contradictory

evidence in the context. A human model for understanding an AI's task fulfillment indeterminacy would also make sense as long as other evidence does not rule out the possibility. However, based on a human model, an agent that possesses and actively uses humanlike language capability would communicate its decision not to fulfill a task. Therefore, there exists a mismatch between the expectation based on verbal indeterminacy and the observed task fulfillment indeterminacy.

Such a seemingly simple *mismatch* between the user's observation and expectation regarding the conversational agent's behavior leads to a large feedback error in the user's mind because conflicting signals lead to violations in neurocognitive expectancies (Friston 2010; Rao and Ballard 1999; Saygin et al. 2011). Consequently, the user cannot rely on the anthropomorphized view of the agent to explain its behavior. Therefore, anthropomorphism is a holistic experience that depends on the match/mismatch among all available anthropomorphic signals, and any mismatch among the signals can reduce anthropomorphism.

In summary, the mismatch between the two indeterminacies leads to a dissonance that makes the user suspect the validity of a human-based model for explaining the AI's behavior. Therefore, we hypothesize that:

> *H2: The interaction of task fulfillment indeterminacy and verbal indeterminacy decreases anthropomorphism.*

Trust

People leverage mechanisms such as trust and control to adjust their confidence level when dealing with indeterminacy in other parties' behavior (Das and Teng 1998). Prior research emphasized the importance of trust in the context of conversational agents (Saffarizadeh et al. 2017). Based on

previous findings, people tend to adopt a human model of interpersonal trust for interactions with artifacts that possess humanlike features (Lankton et al. 2015).

In the past two decades, there has been a convergence in the accepted definition of interpersonal trust (hereafter referred to as trust)[2] (Rousseau et al. 1998; Schoorman et al. 2007). The most widely used definition of trust is "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, irrespective of the ability to monitor or control that other party" (Mayer et al. 1995, p. 712).

To better understand such a complex construct as trust, it is helpful to begin by clarifying what the construct is not. First, trust is not taking risk. Taking risk in an interaction with another party is a decision to intentionally make oneself vulnerable to the actions of the other party. The behavioral manifestation of trust or trusting behavior is "to act as if the uncertain future actions of others were indeed certain in circumstances wherein the violation of these expectations results in negative consequences for those involved" (Lewis and Weigert 1985, p. 971). Therefore, trust is not risk taking behavior or actions, "but an underlying psychological condition that can cause or result from such actions" (Rousseau et al. 1998, p. 395). Second, trust is not cooperation. Cooperation is an observable behavior that could be a consequence of trust as well as many other antecedents such as perceived level of control (Das and Teng 1998). Thus, we do not equate such behaviors as cooperation with trust.

Trustworthiness, alongside with generalized tendency to trust (Rotter 1967), is the main predictor of trust (Mayer et al. 1995). Trustworthiness or trusting beliefs is comprised of many factors

[2] Trust is often referred to as trusting intentions in the IS literature (e.g., McKnight et al. 2002). In this paper, in line with the definitions we adopted, we use the term trust.

(Johnson-George and Swap 1982; Rempel et al. 1985; Rotter 1967), but most scholars agree on three components (Mayer et al. 1995; McKnight et al. 2002): ability or competence (Giffin 1967), integrity or reliability (Giffin 1967; Johnson-George and Swap 1982), and benevolence or caring (Mayer et al. 1995; Rempel et al. 1985).

Impacts of Indeterminacy and Anthropomorphism on Trust

Task fulfillment indeterminacy in a conversational agent means that the user of the agent cannot be sure about the behavioral outcome of the system. The reason for task fulfillment indeterminacy could be due to different things such as system failure, mistaking user commands, conflicting commands, or the agent's own decision to override the user's command(s). No matter what the underlying reason is and what the user perceives the reason to be, the agent exhibits indeterminacy in fulfilling the tasks. Such indeterminacy means a lack of consistency in the agent's behavior. The agent might or might not fulfill the task, which means that the reliability of the system could be questioned. While the user might attribute the lack of fulfillment to either the agent's own intention or some other problem in the system, the unreliability in behavior decreases the user's perception of the agent's integrity.

Furthermore, since indeterminacy means that sometimes the agent will not fulfill the assigned task, the user will find the behavior of the AI to be, on average, less desirable. Given this type of behavior, the user will perceive the agent to be less benevolent as well. Taken together, task fulfillment indeterminacy decreases the AI's trustworthiness, which in turn negatively influences trust. We hypothesize that:

H3: Task fulfillment indeterminacy decreases trust.

When the user anthropomorphizes the agent, s/he perceives it to be more competent, predictable, and caring. First, an anthropomorphized agent is perceived to have high agency, which is an essential part of a humanlike state of mind (Gray et al. 2007). People perceive entities with high agency to be capable of planning, controlling, and fulfilling tasks (Gray et al. 2011; Waytz et al. 2014). Therefore, an anthropomorphized conversational agent is more likely to be perceived as competent. Second, prior research has shown that one of the major reasons that people anthropomorphize nonhuman agents is to increase their ability to predict the agents' behavior (Epley et al. 2007; Waytz, Morewedge, et al. 2010). In other words, anthropomorphism increases *perceived* predictability of an agent. Finally, prior research has shown that anthropomorphism is associated with feelings of connectedness and warmth (Epley et al. 2007; Qiu and Benbasat 2009). Some scholars suggested that lonely people "create human agents out of nonhumans through anthropomorphism to satisfy their motivation for social connection" (Epley et al. 2007, p. 866). Empirical evidence has also revealed that people often anthropomorphize nonhumans such as God, pets, gadgets (Epley, Akalis, et al. 2008; Epley, Waytz, et al. 2008), and robots (Eyssel and Reich 2013) to fulfill their need for social connectedness and caring. Therefore, when a user anthropomorphizes a conversational agent, s/he is more likely to perceive it as caring.

In summary, users perceive an anthropomorphized agent to be more caring, to have more competence to act on its caring, and to be more predictable. In other words, they perceive an anthropomorphized agent to display a predictable caring behavior. Therefore, we argue that users are more willing to be vulnerable to the actions of a conversational agent when they anthropomorphize it. Therefore, in line with prior research (Qiu and Benbasat 2009; Waytz et al. 2014), we hypothesize that:

H4: Anthropomorphism increases trust.

RESEARCH METHOD

Experiment Design

We conducted a 2x2 factorial design experiment in which verbal indeterminacy and task fulfillment indeterminacy were manipulated independently.

We recruited a total of 226 participants of which 152 (78 females, 74 males, and 0 other, with an average age of 36.6 ranging from 20 to 70 years old) followed the instruction of the experiment, received the assigned treatment, and passed the attention check measures. We chose to recruit the participants from Amazon's Mechanical Turk because we wanted the participants for our study to have some experience using digital technology. MTurk participants are more demographically diverse than both standard internet samples and typical American college samples (Buhrmester et al. 2011; Chandler et al. 2019; Mason and Suri 2012) and are typically not too familiar with manipulations and measures because the majority of them are new to the platform every year (Robinson et al. 2019). Recent studies have shown that the quality of data from surveys with attention-check questions on MTurk is comparable to that from surveys with student subjects (Aruguete et al. 2019). Moreover, they found that in many cases, findings from MTurk samples are similar to those from national samples, supporting the generalizability of the findings based on MTurk samples (Coppock 2019). We used the Cloud Research platform (Litman et al. 2017) to remove participants who had participated in pilot studies and to block any users who may have tried to participate in the experiment multiple times (based on IP addresses and geo-locations). Some of the workers on Mechanical Turk might participate in many studies per day (Paolacci et al. 2010). To ensure that we obtained high-quality responses, we limited the participants to those with more than 97% acceptance rate and MTurk experience between 500 and 10,000 HITs (Human

Intelligence Tasks, which are the tasks posted on MTurk marketplace).[3] The experiment took 3 minutes on average and all participants received $0.50 compensation.

The Conversational Agent

We developed a conversational agent named Amanda to increase the external validity of our study. The artifact uses state of the art text-to-speech technologies. We used Amazon's AWS Polly text-to-speech to provide a humanlike voice for the agent. Many experimental studies lack external validity and ecological validity partially because the artifacts used in the experiment are poor representations of the real-world artifacts. Our choice of technologies helps to address such concerns. Furthermore, by developing an actual functioning conversational agent we created a task environment that is engaging for participants and has a high degree of psychological realism (Berkowitz and Donnerstein 1982).

Procedure

The participants were asked to open the web version of our app on their browsers and to begin by reading a short set of instructions in which they were told that they would be testing a certain functionality of the agent to make sure it works properly on different devices. We asked the participants to tell the agent to "turn on the light" and to try this command five times. Each time, after the AI received the command, it generated a response based on the treatment condition to which the participant was assigned. After the interaction, participants were asked to respond to a

[3] The upper bound was placed in order to avoid possible adverse effects associated with survey fatigue and familiarity with the measures.

series of questions. We debriefed the participants at the end. Figure 3 shows two snapshots of the interaction between a participant and the agent.

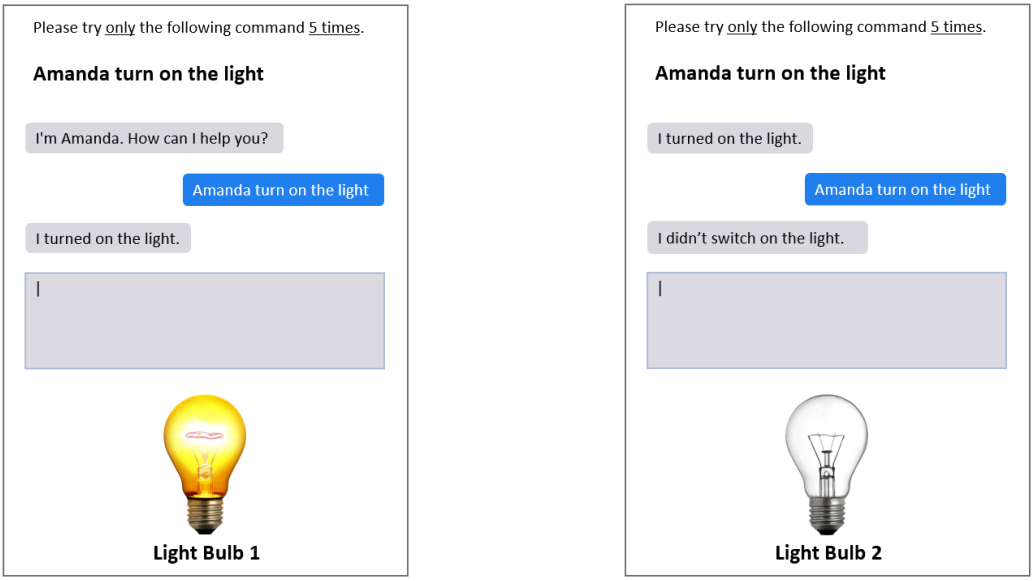


Figure 3. A Sample of Participant's Interaction with Amanda

Operationalization of Constructs

In line with previous research in information systems, we measure trust using a holistic measure (Srivastava and Chandra 2018). In this research, we are not interested in measuring trustworthiness and its components, therefore, we do not separately measure such variables as ability, benevolence, and integrity. This approach is also in line with research in the management field (Mayer and Davis 1999; Mayer and Gavin 2005; Schoorman et al. 2007). We used a 7-point Likert scale to measure trust.

Anthropomorphism, in its essence, is the attribution of humanlike attributes to nonhuman entities (Epley et al. 2007). In addition, most definitions of humanness suggest that uniquely human attributes are those related to human mental states. Therefore, we operationalize anthropomorphism using five items measuring participants' perception of an artifact's humanlike

mental states, namely possession of mind, intentions, free will, consciousness, and emotions. Each item is measured using a 5-point Likert-type scale. The measured attributes can convey both agency and experience dimensions of the artifact (Gray et al. 2007, 2011; Gray and Wegner 2012). Other researchers have used the same items to operationalize anthropomorphism in similar contexts (e.g., Epley, Waytz, et al. 2008; Hart et al. 2013; Waytz, Cacioppo, et al. 2010; Waytz, Morewedge, et al. 2010).

We directly manipulate verbal and task fulfillment indeterminacies. Table 1 provides a summary of the measures used.

Table 1. Operationalization of Constructs

Construct	Definition	Measures	Sources that Informed the Measures
Trust	"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, irrespective of the ability to monitor or control that other party" (Mayer et al. 1995, p. 712).	(1-7 scale) t1. I trust Amanda to be reliable. t2. I believe Amanda to be trustworthy. t3. I trust Amanda.	Srivastava and Chandra 2018
Anthropomorphism	An 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; Gray et al. 2007).	(1-5 scale) 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, Waytz, et al. 2008 Waytz, Cacioppo, et al. 2010

Verbal Indeterminacy	Perceived indeterminant variation in the ways an agent conveys a given message using different choices of words and grammar.	Manipulated through random wording of agent responses. Manipulation check questions: (1-7 scale) vi1. I can predict what Amanda would say, word for word. (R) vi2. There is an obvious pattern of words in Amanda's sentences. (R) vi3. Amanda's choice of words is predictable. (R)	-
Task Fulfillment Indeterminacy	Perceived indeterminant variation in the outcome of a task that an agent is supposed to fulfill.	Manipulated through random task fulfillment. Manipulation check questions: (1-7 scale) ti1. Amanda is predictable in terms of turning on the light. (R) ti2. When I say "turn on the light," I can predict whether Amanda will turn on the light. (R) ti3. I can predict whether Amanda turns on the light when I ask her to do so. (R)	-

Manipulation of Verbal Indeterminacy and Task Fulfillment Indeterminacy

Verbal indeterminacy was manipulated by making the conversational agent use either a fixed sentence or a verbal variation of the same underlying message to communicate with the participant. For instance, in the fixed sentence condition, if the participant asked the conversational agent to turn on the light, the agent would always use “I turned on the light” to communicate that it turned on the light. However, in the verbally varying condition, the agent would use different variations of the same underlying message, such as “I switched on the light” or “the light is on now,” to respond to the user.

Task fulfillment indeterminacy was manipulated by making the conversational agent either always fulfill the task or randomly fulfill the task. For instance, if the participant asked the agent to turn on the light, in the always fulfill condition the agent would always say that it turned on the light,

while in the random fulfillment condition the agent might say that it turned or did not turn on the light. Table 2 shows more details about the design. Note that for brevity Table 2 only contains two variations of the sentence for verbal indeterminacy, while in the experiment the randomizer chose from ten variations of the sentence.

Table 2. Experiment Design

		Task Fulfillment Indeterminacy	
		No	**Yes**
Verbal Indeterminacy	**No**	1) I turned on the light!	2) I turned on the light! I didn't turn on the light!
	Yes	3) I turned on the light! I switched on the light!	4) I turned on the light! I didn't turn on the light! I switched on the light! I didn't switch on the light!

To eliminate the possible effect of any specific variation of the sentence on the results of group 1 and group 2, we randomly chose the base (non-random) sentence for each participant. For instance, when a participant is randomly assigned to group 1, the randomizer might choose either "I turned on the light" or "I switched on the light" for the participant, and keep this choice for all interactions with this participant. Group 1 is different from group 3 because in group 1 the participant keeps receiving the one randomly constructed sentence every time he or she asks the agent to turn on the light, but in group 3 the sentence is randomly constructed each time, so the participant randomly receives a different sentence each time s/he asks the agent to turn on the light.

Control Variables

We used two control variables, gender and age. Previous research has shown that age and gender could play a role in shaping trust and trust-related intentions (Riedl et al. 2010; Yuan and Dennis 2019).

ANALYSIS AND RESULTS

Manipulation Checks

We conducted manipulation checks to assess whether participants perceived our manipulations as we planned. The manipulation checks for verbal indeterminacy and task fulfillment indeterminacy each asked the participants to answer three questions listed in Table 1. We averaged the three items for each manipulation check. In a one-way ANOVA, the mean difference between low verbal indeterminacy ($M = 3.01, SD = 1.24$) and high verbal indeterminacy ($M = 3.52, SD = 1.32$) was statistically significant and in the expected direction ($F(1,150) = 5.839, p < 0.01, \eta_p^2 = 0.04$). In a separate one-way ANOVA test, the mean difference between low task fulfillment indeterminacy ($M = 2.44, SD = 1.41$) and high task fulfillment indeterminacy ($M = 3.60, SD = 1.66$) was statistically significant and in the expected direction ($F(1,150) = 21.722, p < 0.01, \eta_p^2 = 0.13$).

Measurement Model

We measured anthropomorphism using five indicators. We chose to model the construct reflectively based on Waytz et al., who argued that these measurement items "should reflect anthropomorphism" (2010, p. 221). We measured trust using three indicators. We chose to model the construct reflectively as previous studies in IS have operationalized trust as a reflective construct (e.g., Petter et al. 2007; Srivastava and Chandra 2018). Note that by operationalizing anthropomorphism and trust in this way we assumed that they are not composites of their indicators but common factors of them.

The indicators for anthropomorphism and trust were measured using Likert-type scales. While most of the previous research treated these measures as continuous, we statistically tested whether they could be treated as continuous variables. All three indicators of trust were fairly symmetrically distributed with seven categories (skewness of -0.53, -0.34, and -0.35). Therefore, we could treat trust as either continuous or ordinal. However, indicators of anthropomorphism followed non-symmetrical distributions (skewness of 0.51, 0.83, 1.41, 1.25, and 2.02) with five categories. Thus, we could not treat anthropomorphism measures as continuous. For simplicity, we treated both trust and anthropomorphism as ordinal. We simultaneously estimated the thresholds for trust and anthropomorphism with the rest of the model.[4] The thresholds were used to create categories for the ordinal levels of each variable.

We used confirmatory factor analysis (CFA) to assess the measurement items using lavaan (version 0.6-5 on R version 3.6.1). Since a CFA model is saturated, i.e., all constructs can freely covary with other constructs, any misfit in the model is due to how the items fit with the constructs. We used CFI, RMSEA, and SRMR to assess the fit. Such fit indices are more robust to variations in sample size as compared to the chi-square measure of fit. Moreover, simulation studies on fit measures have revealed that while many of the fit indices might lead to incorrect interpretations about the fit, a combination rule based on the three mentioned indices can provide a good indicator for fit (Hu and Bentler 1999). Based on this combination rule, a model with CFI of more than 0.95 and either RMSEA of less than 0.06 or SRMR of less than 0.08 has a good fit with the data. The fit measures for our model are CFI=0.984, RMSEA=0.056, and SRMR=0.037, indicating a satisfactory fit.

[4] We assumed each level of an ordinal measure corresponds to a specific value on its underlying latent continuous measure. As such, we estimated these values or "thresholds" as a part of our estimation, simultaneously.

We assessed the convergent validity of our measurement model by inspecting the lambda (i.e., item loadings) and average variance extracted (AVE) values. The acceptable thresholds for lambda and AVE are 0.70 and 0.50, respectively (Kline 2015). All lambda values, except for one item for anthropomorphism with a value of 0.69, were larger than 0.70. Also, all AVE values were larger than the 0.50 threshold, providing support for the convergent validity of the measurement model.

To assess the discriminant validity of our measurement model, we tested whether each construct has more common variance with its items than with other constructs. More specifically, we tested whether the square root of each construct's AVE was greater than its correlation with other constructs (Fornell and Larcker 1981). All constructs passed the test, providing support for discriminant validity of the measurement model. The composite reliabilities of all constructs were above 0.70, which is the threshold for reliability (Fornell and Larcker 1981). Table 3 presents the descriptive statistics, correlations, square roots of AVE values, and item loadings (lambda values).

Table 3. Descriptive Statistics, Correlations, $\sqrt{AVE}$, and Loadings (N=152)

Construct / Variable	Loadings	M	SD	1	2	3	4	5	6
1. Age	NA	36.61	11.21	NA					
2. Gender	NA	0.49	0.50	-0.06	NA				
3. Verbal Indeterminacy	NA	0.46	0.50	-0.01	0.00	NA			
4. Task Fulfillment Indeterminacy	NA	0.48	0.50	-0.03	0.06	0.12	NA		
5. Anthropomorphism	a1. 0.691	1.74	0.79	-0.14	-0.03	0.06	0.01	0.78	
	a2. 0.795								
	a3. 0.791								
	a4. 0.832								
	a5. 0.759								
6. Trust	t1. 0.842	4.20	1.60	-0.06	-0.06	-0.02	-0.39	0.26	0.93
	t2. 0.965								
	t3. 0.985								

Notes: $\sqrt{AVE}$ (square root of average variance extracted) values are represented on the diagonal and correlations are shown off-diagonal. Gender is coded as 0=female and 1=male.

Structural Model

We used a factor-based structural equation model (SEM) to analyze our path model. We coded the two manipulated variables as binary and generated a binary interaction term by multiplying the interacting terms. Our control variables (age and gender) did not significantly influence trust or improve the model fit. Therefore, we dropped them from the structural model. We employed a WLSMV (mean- and variance-adjusted weighted least squares) estimator, which is a DWLS (diagonally weighted least squares) estimation method with robust standard errors for ordinal data (Xia 2016). We first investigated the fit indices to evaluate whether the model was a good representation of the data. Since we used a robust estimation method, we adopted scaled fit measures. The scaling correction factor for our model is 0.913 and the model has 39 degrees of freedom.

Scaled chi-square is 39.392 (df.scaled=39, p.scaled= 0.452), which fails to reject the hypothesis that the model constraints and assumptions hold within the sampling error. While chi-square is satisfactory and can provide some general sense about the fit of the model, it has many shortcomings, such as its substantial dependence on sample size. Therefore, we rely on robust fit measures to assess the model. Scaled RMSEA is 0.008 (p.scaled = 0.899), scaled CFI is 0.999, and scaled SRMR is 0.043. These three fit indices indicate a good fit according to Hu and Bentler's (1999) thresholds (SRMR< .08 AND [CFI>0.95 OR RMSEA<0.06]).

Path Testing

We used the path estimations from the model to assess the hypotheses (see Figure 4). Since we did not mean-center the indicators of verbal and task fulfillment indeterminacies and these factors are single-indicator, the regression coefficient of each of these factors in the model indicates the effect

of the factor when the other factor is not present (set to zero). Therefore, in order to assess H1a and H1b, we can directly use the estimates without creating a base model with no interaction. Furthermore, the common method of testing such hypotheses in a model with no interaction has fundamental issues because if the interaction is important in the full model, then a model with no interaction could yield biased estimates, because the needed estimates are usually highly correlated with the interaction term.

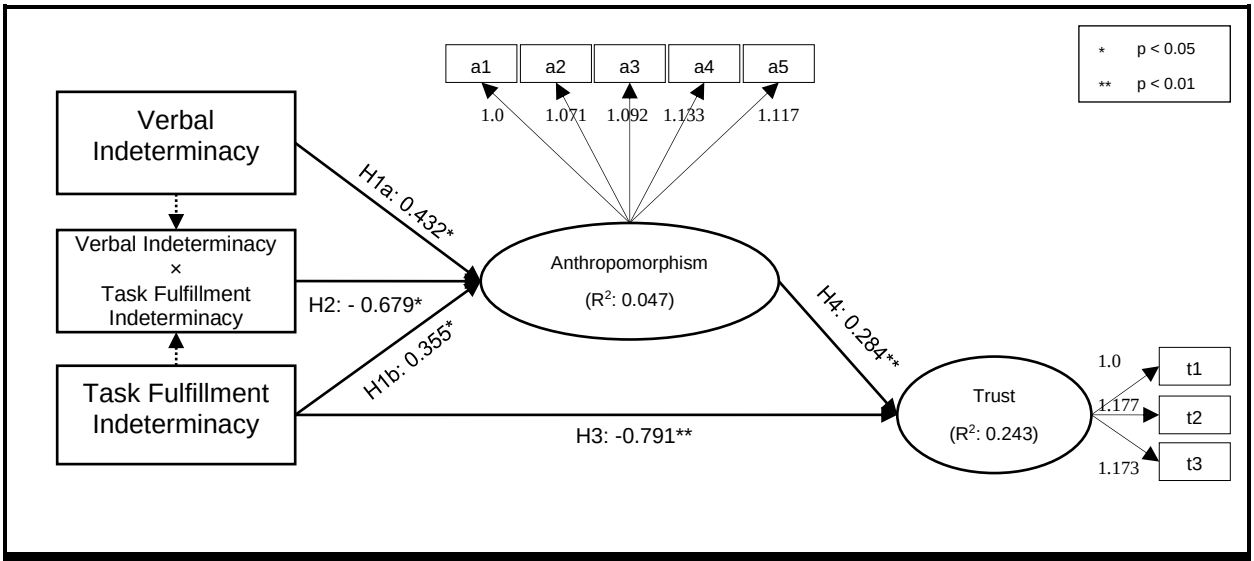


Figure 4. SEM Results

Hypothesis 1a stated that verbal indeterminacy increases anthropomorphism. Our model estimation showed a significant positive effect of verbal indeterminacy on anthropomorphism ($\beta = 0.432, p < 0.05$). This result provides support for H1a by indicating that verbal indeterminacy increases anthropomorphism. Hypothesis 1b stated that task fulfillment indeterminacy increases anthropomorphism. The estimated model provided support for this hypothesis showing a significant positive effect ($\beta = 0.355, p < 0.05$).

Hypothesis 2 theorized that the presence of each of the two forms of indeterminacy decreases the effect of the other one on anthropomorphism. The result supported this claim by showing a

significant negative effect of the interaction term on anthropomorphism ($\beta = -0.679, p < 0.05$). Therefore, the estimated model suggested that the effect of verbal indeterminacy on anthropomorphism depends on the amount of task fulfillment indeterminacy. To further probe this interaction effect, we constructed an interaction plot and examined the simple slopes of the relationship between verbal indeterminacy and anthropomorphism (see Figure 5). The interaction between verbal indeterminacy and task fulfillment indeterminacy suggest that both types of indeterminacy simultaneously inform anthropomorphism process, and therefore any mismatch between the two significantly decreases anthropomorphism. This provides support for the notion that anthropomorphism is a holistic experience.

Verbal indeterminacy, task fulfillment indeterminacy, and their interaction explain 5% of the variance in anthropomorphism. We argue that since both predictors are exogenous, i.e., externally and independently manipulated, the path coefficients are unbiased despite the relatively low R^2.

According to Fritz et al. (2012), the variance explained in an experiment should not be compared to that of a correlational research. While many factors, such as generalized tendency to anthropomorphize (Waytz, Cacioppo, et al. 2010), could influence anthropomorphism, we manipulated only two types of indeterminacy. It would be surprising if people's inference of the humanness of an AI agent were shaped only by a few factors. Our measures of anthropomorphism reflect such aspects as free will, consciousness, and ability to experience emotions. Therefore even a small increase in users' anthropomorphism of an AI agent that goes beyond adding basic physical cues such as a humanlike face or voice (Yuan and Dennis 2019) has high practical and theoretical significance.

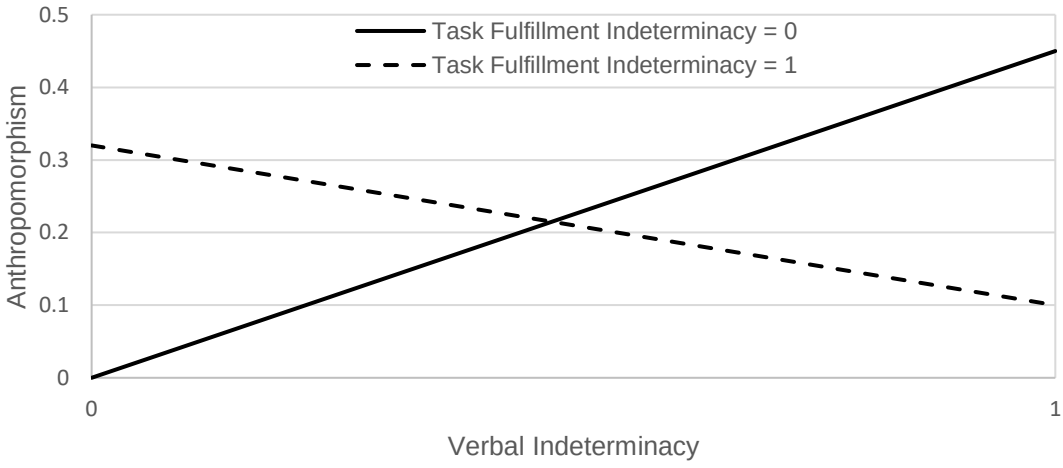


Figure 5. Interaction Plot

Hypothesis 3 predicted that task fulfillment indeterminacy decreases trust. The empirical model provided support for this claim by indicating a significant negative effect of task fulfillment indeterminacy on trust ($\beta = -0.791, p < 0.01$). This finding provides evidence that inconsistency created by the indeterminacy negatively influences trust.

Hypothesis 4, in line with previous research (Qiu and Benbasat 2009; Waytz et al. 2014), predicted a positive effect of anthropomorphism on trust. The estimated model provided supporting evidence for this assertion by showing a positive association between the two constructs ($\beta = 0.284, p < 0.01$). While verbal and task fulfillment indeterminacies were externally manipulated and therefore we could guarantee that they preceded anthropomorphism and trust, we cannot guarantee the same for anthropomorphism with respect to trust. Hence, the results show a correlation and the causality is inferred based on the theoretical reasoning and similar finding from prior literature. Task fulfillment indeterminacy and anthropomorphism explain 24% of the variance in trust. Table 4 provides a summary of the results.

Table 4. Results Summary		
Hypothesis	**Relationship**	**Finding**
H1a	Verbal Indeterminacy $\overset{+}{\rightarrow}$ **Anthropomorphism**	Supported
H1b	Task Fulfilment Indeterminacy $\overset{+}{\rightarrow}$ **Anthropomorphism**	Supported
H2	Verbal Indeterminacy × Task Fulfilment Indeterminacy $\overset{-}{\rightarrow}$ **Anthropomorphism**	Supported
H3	Task Fulfilment Indeterminacy $\overset{-}{\rightarrow}$ **Trust**	Supported
H4	Anthropomorphism $\overset{+}{\rightarrow}$ **Trust**	Supported

Post-Hoc Analysis: Mediation

We did not explicitly develop any hypothesis on the mediating role of anthropomorphism on the relationship between indeterminacy and trust because we did not have a strong theory based on which we could hypothesize such mediation. However, mediation analysis can help us understand the relationship between indeterminacy and trust.

While the total effect of verbal and task fulfillment indeterminacies are still relevant, we are also interested in the effect that is transferred through anthropomorphism. As Zhao et al. argued, "to establish mediation, all that matters is that the indirect effect is significant" (2010, p. 204). We used simultaneous estimation of all paths in SEM as it is preferred over Baron and Kenny's (1986) three steps of regression analysis because it does not assume that the three regressions are independent (Preacher and Hayes 2008). Furthermore, since mediated effects were constructed based on the product term of two normal parameters, we used bootstrapping to avoid making unwarranted assumptions about the normality of the product term (Hayes and Preacher 2014; Preacher and Hayes 2008). We used 1000 bootstrap samples to find the standard errors.

To be able to calculate the total effect of verbal indeterminacy, task indeterminacy, and their interaction on trust, we tested a saturated model that included all the paths from our previous model

as well as the direct effects of both verbal indeterminacy and the interaction term on trust. The model scaled fit measures were satisfactory (CFI= 0.999, RMSEA=0.038, SRMR=0.043), therefore, we could comfortably interpret the estimates. Note that the estimates of paths that existed in both the main and saturated models were similar in direction and p-value.

Based on our mediation analysis, verbal indeterminacy has a significant positive influence on trust via anthropomorphism ($\beta = 0.123, p < 0.05$). This is interesting because we failed to find any evidence suggesting a direct effect of verbal indeterminacy on trust ($\beta = -0.005, p = 0.981$). Therefore, the effect of verbal indeterminacy on trust is fully mediated by anthropomorphism. While verbal indeterminacy does not seem to be related to user's trust toward the conversational agent, this analysis suggests that verbal indeterminacy can increase trust indirectly by increasing anthropomorphism.

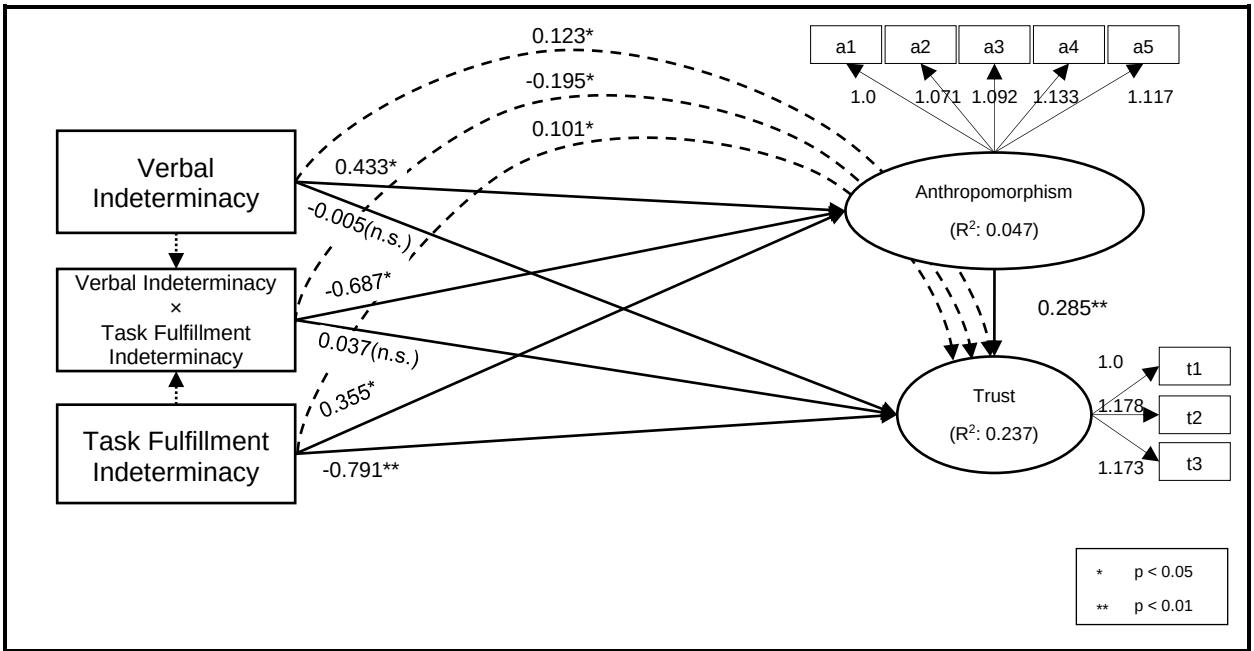


Figure 6. Saturated Model for Testing Mediation

Moreover, we found that task fulfillment indeterminacy has a significant positive influence on trust via anthropomorphism ($\beta = 0.101, p < 0.05$). The negative direct effect ($\beta = -0.791, p < 0.05$)

and positive indirect effect of task fulfillment indeterminacy create an interesting paradox because task fulfillment indeterminacy should decrease trust by definition. While the total effect showed a significantly negative net effect ($\beta = -0.690, p < 0.05$) indicating that the direct effect was stronger than the indirect effect, the positive indirect effect of task fulfillment indeterminacy on trust through anthropomorphism mitigated its negative direct effect to some extent (see Figure 6).

Anthropomorphism also mediated the effect of the interaction term (task fulfillment indeterminacy × verbal indeterminacy) on trust. We found a significant negative indirect effect of the interaction term on trust ($\beta = -0.195, p < 0.05$). However, we did not find any evidence suggesting either a direct ($\beta = 0.037, p = 0.844$) or a total ($\beta = -0.158, p = 0.374$) effect of the interaction term on trust.

Robustness Check

We conducted an additional experiment to serve as a robustness check and to rule out rival explanations for the observed effects described above. First, the reader will remember that even though we acknowledged that task fulfillment indeterminacy can be viewed as indeterminacy in both the outcome and the process of fulfillment, we chose to focus on outcome rather than process in designing our initial experiment. In doing so, we operationalized task fulfillment indeterminacy using a task with two outcomes: light is on or light is off. In this operationalization the outcome of the task is desirable 100% of the time for the low task fulfillment indeterminacy condition because the user's command is always fulfilled. However, the outcome of the task is desirable only 50% of the time for the high task fulfillment indeterminacy condition because the command is fulfilled only half of the time on average, due to randomness. This difference in desirability of the outcome reduces trust. While this is in line with our theoretical development where we discussed that the

reduced desirability of the outcome is a reason for the reduced trust, we cannot ignore the possibility that our results might have been different if we had chosen to examine task fulfillment indeterminacy in terms of process (as opposed to outcome). Therefore, we can add robustness to our findings if we show that the negative direct effect of task fulfillment indeterminacy on trust still exists even when the desirability of the outcome remains the same. This robustness check can also show whether our choice to focus on task fulfillment indeterminacy as indeterminacy in the outcome of the task that an agent is supposed to fulfill had a limiting effect on the theoretical generalizability of our results.

Second, in the operationalization of task fulfillment, the conversational agent says, "I did not turn on the light" (or a sentence with a similar meaning) whenever it fails to fulfill the task. One could argue that the reason for increased anthropomorphism is not task fulfillment indeterminacy, but the fact that the agent shows rebellion against the user. We argue that when a user construes the agent's task fulfillment indeterminacy as rebellion, he or she is in fact attributing human-like concepts (i.e., rebellion) to a non-human agent (i.e., anthropomorphism) to understand its behavior. However, we acknowledge that the agent's utterance ("I did not turn on the light") itself could be a confounding factor that directly leads to increased anthropomorphism due to perceived rebellion. Therefore, we can add robustness to our results if we show that task fulfillment indeterminacy increases anthropomorphism even when the agent does not signal rebellion through its utterance.

Finally, in our experiment we told the participants to ask the agent to turn on the light five times to make sure the agent works properly. One could argue that this is not a realistic setting because users typically do not ask an agent to do the same task five times in a row. Therefore, we can add robustness to our study, if we change the task environment each time the participant needs to repeat the command, so that the repetition of the command has greater ecological validity.

We conducted an additional experiment to add robustness to our findings by addressing the above threats to the validity of our results. We recruited 60 participants (25 females, 34 males, and 1 other, with an average age of 41.7 ranging from 28 to 76 years old) from Amazon's Mechanical Turk (pool of Turkers with more than 97% acceptance rate, more than 500 HITs, and master status). All participants passed the attention check questions and received $0.75 for participating in our study, which took 5 minutes on average (ranging from 2.7 to 12.15 minutes).

We told the participants that Amanda was a digital assistant which could control many home devices even if the device was not smart, and that the developers of Amanda had created a simulation that allowed users to see how it worked in real-life. We asked them to use Amanda in a simulation to control a mower. We explained that the mower worked like a real device, was not smart, and was entirely controlled by Amanda (see Appendix B). As shown in Figure 7, participants were asked to use Amanda 10 times to park the mower in a predetermined parking spot. The locations of the parking spot and the mower along with a random tree were fixed in all variations of the task environment. However, in order to create variation in the task environment for every time participants ask Amanda to park the mower, we randomly selected the grass texture (from a pool of 15 grass texture images) and the tree image (from a pool of 5 tree images), and randomly located 15 flowers on the grass (see Appendix B).

In this experiment, we manipulated task fulfillment indeterminacy in terms of *how* the agent performs the task (i.e., process indeterminacy). In doing so, we randomly assigned participants to either a low task fulfillment indeterminacy condition in which Amanda used the exact same path to park the mower every time or a high task fulfillment indeterminacy condition in which Amanda

used a randomly selected path (from all possible paths)[5] every time. To assure that no specific path drives the results in the low task indeterminacy condition, for each specific participant in this condition, we randomly picked one path from all possible paths and kept the path constant throughout all interactions of the participant with the agent. In both conditions, Amanda responded to the participant's command by just saying "OK!".

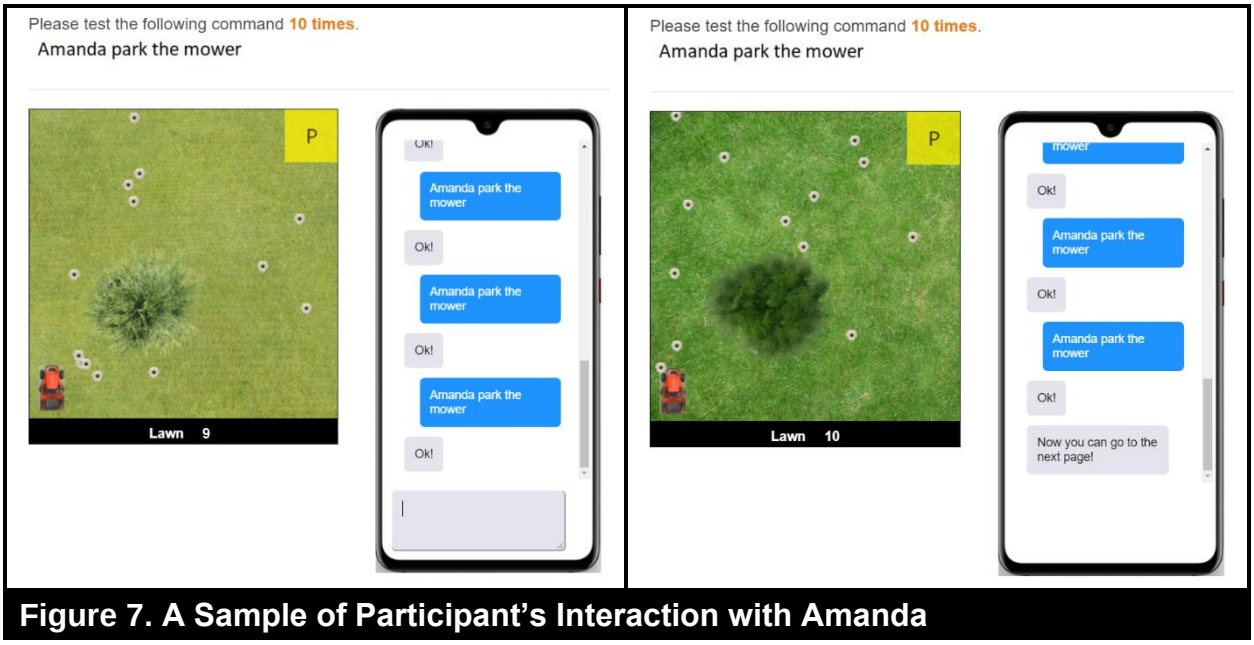


Figure 7. A Sample of Participant's Interaction with Amanda

As a manipulation check, we asked the participants whether they agreed or disagreed with the following statements: "the way Amanda parked the mower was unpredictable," "I did NOT notice a pattern in the way Amanda parked the mower," and "I could NOT predict exactly how Amanda would park the mower." Participants in the low task indeterminacy condition reported lower perceived task indeterminacy ($M = 1.96, SD = 1.07$) than participants in the high task indeterminacy condition ($M = 4.58, SD = 1.75; F(1,59) = 50.765, p < 0.001, \eta_p^2 = 0.47$).

[5] To limit the possible number of paths from the mower location to the parking spot to a tractable number, we created a graph with 16 nodes, where the mower was on node 1 and parking spot was on node 16. Assuming that the mower does not go through the same location more than once and cannot go over the tree, there are 22 directional edges in the graph. We employed a depth first search (DFS) algorithm to find all possible paths from node 1 to node 16.

Due to the relatively small sample size[6], we used the PROCESS macro in SPSS to estimate the effect of task fulfillment indeterminacy on anthropomorphism and trust rather than SEM. Accordingly, we constructed linear composites for anthropomorphism ($\alpha = 0.93$; 5 items) and trust ($\alpha = 0.96$; 3 items). In an exploratory factor analysis, both the anthropomorphism and trust items loaded higher on their respective construct than the other construct and had loadings greater than 0.7 (see Appendix B). The results showed that task fulfillment indeterminacy significantly increases anthropomorphism ($\beta = 0.757, p < 0.05; \eta_p^2 = 0.06$), and decreases trust ($\beta = -0.617, p < 0.05; \eta_p^2 = 0.07$), controlling for anthropomorphism, age, and gender (similar to the previous study). These results lend a measure of robustness to our findings by demonstrating that they can be reproduced in a different task context that involves process rather than outcome indeterminacy and by ruling out potential rival explanations such as AI rebellion or the manner in which task fulfillment indeterminacy was operationalized.

DISCUSSION

While the question of what qualifies an entity to be a human might be a research question for philosophers, the question of what makes people *perceive* an entity as a human has important implications for applied disciplines (Waytz, Cacioppo, et al. 2010). We leveraged the anthropomorphism literature to hone in on the concept of trust in human-AI interaction. This study contributes to both the anthropomorphism and trust streams of research in psychology, human-

[6] The results were the same in terms of direction and significance when we used a SEM analysis similar to the one described in the main experiment. We modeled anthropomorphism and trust ordinal variables and used WLSMV estimator. We found that task fulfillment indeterminacy increases anthropomorphism ($\beta = 0.455, p < 0.01; \eta_p^2 = 0.06$) and decreases trust ($\beta = -0.503, p < 0.01; \eta_p^2 = 0.08$), when controlling for anthropomorphism, age, and gender. However, despite satisfactory scaled fit measures (df=39; chi-squared= 40.887; CFI=0.999; RMSEA= 0.029; SRMR= 0.062), we could not rely on these results due to small sample size.

computer interaction, and information systems. Below we elaborate on the implications of our findings for research and practice, and the limitations of our study.

Implications for Research

Prior research on anthropomorphism has ignored the nature of the indeterminacy that leads to anthropomorphism. In this research, we defined the concept of AI indeterminacy. We specifically identified, defined, and studied two AI indeterminacies, namely verbal indeterminacy and task fulfillment indeterminacy, that exist in the context of conversational AI. We empirically showed that verbal and task fulfillment indeterminacies have positive effects on users' perception of AI's humanness, i.e., anthropomorphism.

In addition, we addressed the apparent tensions between anthropomorphism and trust in the presence of indeterminacy. More specifically, we showed that the negative effect of indeterminacy on trust can be traced to the direct negative effect of task fulfillment indeterminacy on trust. In contrast, verbal indeterminacy not only does not decrease trust but also has an indirect positive effect on it. This finding has important theoretical implications for research on trust in human-AI interaction because it indicates that different types of indeterminacies could have opposite effects on trust.

Previous literature identified that humans are driven to find an explanation for the unpredictability in their surroundings (i.e., effectance motivation), as a reason why people anthropomorphize unpredictable agents (Epley et al. 2007; Epley, Waytz, et al. 2008; Waytz, Morewedge, et al. 2010). This research adds to this account by conceptualizing unpredictability as indeterminacy. By doing so, we bridged the anthropomorphism literature and centuries of studies on free will in the field of philosophy (e.g., Nahmias et al. 2014), which can inform future research on human-AI

interaction. Furthermore, we contextualized the previously ambiguous concept of indeterminacy at a more granular level. While in most of the previous research indeterminacy represented a philosophical concept, in this research we introduced two specific types of AI indeterminacy that can be measured and manipulated empirically.

We also investigated the interaction effect of multiple sources of AI indeterminacy on users' perception of AI's humanness. To the best of our knowledge, this study is the first to theorize and empirically test such interaction. We found evidence to show that multiple sources of indeterminacy can influence the effect of each other on anthropomorphism of an AI artifact. More specifically, we found that while verbal indeterminacy alone could increase anthropomorphism, when it co-occurred with task fulfillment indeterminacy, it lost its effect to some extent. We explained why such a phenomenon took place. Verbal indeterminacy signals a humanlike state of mind and specifically the ability to communicate thought and reasons, and task fulfillment indeterminacy signals a mind that has the ability to make an independent decision whether or not to fulfill a command. However, when both behaviors are present (verbal indeterminacy and task indeterminacy), users expect the AI to be able to communicate the reason for not fulfilling the task. Anthropomorphism happens when users can explain the AI's behavior using their mental model of humans. Therefore, the presence of non-contradictory anthropomorphic signals enables them to perceive the AI as more human, while the presence of contradictory anthropomorphic signals hinders their imagination to create humans out of the AI artifacts, i.e., to anthropomorphize.

This study also contributes to trust literature by identifying AI-specific drivers of trust. Decades of research on the concept of trust has been done based on early works on trust in human-human interaction (e.g., Deutsch 1958; Johnson-George and Swap 1982; Rempel et al. 1985; Rotter 1967). While the nature of trust in human interaction with different agents might stay the same, an AI-

specific conceptualization of its drivers could portray a more accurate picture of reality for human-AI interactions. In this research, we introduced task fulfillment indeterminacy as a type of indeterminacy on the conversational agent's side. This construct could be related to the reliability of the technology (Lankton et al. 2015) or the integrity of the agent (Mayer et al. 1995). However, conceptualizing the phenomenon as a type of indeterminacy enables researchers to understand *why* and *how* users might anthropomorphize the agent in order to increase their ability to predict the agent's behavior, which in turn influences their trust.

Thus, this research is a response to the call for the development of contextualized trust (Mayer et al. 1995; Schoorman et al. 2007). Prior attempts to develop a framework for trust in the AI context provided very little theoretical explanation on similarities and dissimilarities of trust in humans and trust in AI (e.g., Hancock et al. 2011). While IS researchers pointed out that the level of the humanness of technology is critical in choosing the proper operationalization of trust (human trust versus technology trust) (Lankton et al. 2015), very little research has been done to determine the underlying drivers of trust in the AI context. This research is among the first attempts to delineate the nuanced behavioral similarities between human and AI, i.e., behavioral indeterminacies, that influence our trust in AI in general and conversational agents in particular.

Furthermore, in this research, we tested the impact of verbal indeterminacy, task fulfillment indeterminacy, and anthropomorphism on trust, which is a major determinant of technology use (Gefen et al. 2003). We found a negative direct effect of task fulfillment indeterminacy on trust but no significant direct effect of verbal indeterminacy on trust. However, we showed that both verbal and task fulfillment indeterminacies positively influence trust through anthropomorphism. By investigating the direct and indirect effects of indeterminacy on trust, we highlighted the

importance of considering indeterminacy and anthropomorphism in the wider general frameworks of technology adoption and use.

Implications for Practice

Based on our analysis, we believe that signaling more humanlike state of mind in an artifact, through multiple signals, could sometimes lead people to perceive the AI as less humanlike. A user makes assumptions about an artifact's unobserved capabilities when interacting with it. These assumptions are, to some degree, consistent with a model of a human with similar capabilities. The user does not necessarily expect to directly observe those imaginary assumed capabilities. As long as those capabilities are imaginary, they are flexible enough for the user to explain away possible inconsistencies in artifact's behavior. However, when the user actually observes a capability, any contradictory signal that gives the user concrete evidence that the artifact's capability does not match that of a human could make the user reject the whole idea that the artifact is humanlike. If it looks like a duck, but does not quack like a duck, then it might not be a duck! Thus, we believe that when the goal is to increase the perceived humanness of an artifact, the developers should avoid using half-developed anthropomorphic features that could provide contradictory evidence.

Based on our theory, we propose that developers address conflicting signals between task fulfillment and verbal indeterminacy by adding *verbally indeterminant error messages* to their conversational agents. A verbally indeterminant error message conveys the error message using a different choice of words and grammar every time. Such a message addresses the aforementioned signal conflicts while it preserves the artifact's verbal indeterminacy. A verbally determinant error message might undermine the perceived humanness over time as it provides evidence of determinacy in the AI.

Based on our findings, we speculate that developers can take advantage of task fulfillment indeterminacy as an inexpensive method to increase perceived humanness. A human makes mistakes and disobeys, and so does a humanlike agent. However, the perceived humanness induced by task fulfillment indeterminacy comes with a price: decreased trust. Increasing anthropomorphism through adding task fulfillment indeterminacy could be useful in some contexts such as gaming where an agent that makes mistakes might be perceived as more humanlike (e.g., the commentator of a soccer game), but it would not be useful in other contexts such as self-driving cars where the agent is supposed to fulfill a safety-critical task with high reliability. Therefore, we propose that when developers need to increase the sense of humanness in the AI, they can deliberately take advantage of task fulfillment indeterminacy provided that the context of the human-AI interaction does not require an AI with high reliability. Otherwise, they should focus on non-task fulfillment indeterminacies such as verbal indeterminacy.

Limitations and Future Directions

While there are different types of indeterminacy in a conversational agent, we focused on verbal and task fulfillment indeterminacies as two key indeterminacies that are present in any conversational agent. For instance, indeterminant variations in the response time, tone, pauses, and facial expression (in a conversational agent with a physical embodiment or graphical representation) could potentially signal mind possession. While we limited our study to the more common indeterminacies, we acknowledge that there may be other types of AI indeterminacy that are worthy of study.

In this research, we measured users' perception of the humanness of the artifact using existing measures of anthropomorphism. While this approach fits our objectives, future research could

distinguish between anthropomorphism toward a specific artifact and a person's general tendency to anthropomorphize non-human agents (Waytz, Cacioppo, et al. 2010).

Users' perception of the humanness of AI artifacts could change over time. In this research, in line with extant literature, we focused on a relatively short human-AI interaction and measured anthropomorphism only once at the end of the experiment. Nevertheless, more research is needed to understand the dynamics of anthropomorphism, i.e., how the perception of humanness changes over time as users collect more information about the AI's behavior. It is also important to determine whether indeterminacy provides a more sustainable source of anthropomorphism compared to other sources such as physical anthropomorphic features such as humanlike voice, avatar, and physical embodiment.

In this study, we adopted a one-dimensional approach to trust. Future research can expand our model by dissecting trust into cognition-based and affect-based trust to understand whether different types of indeterminacies influence cognition-based and affect-based trust differently. For instance, it is possible that task fulfillment indeterminacy erodes the cognitive foundations of trust by providing evidence about lack of integrity in the artifact's behavior and simultaneously enhances the affective foundations of trust by inducing a sense of warmth due to the vulnerability of the artifact.

The concept of anthropomorphism does not imply that the attribution of humanlike qualities to a nonhuman entity is an error or that the entity does not deserve to be treated as a human being. Whether the entity deserves to be treated like a human being is independent of the humanizing process, i.e., anthropomorphism. Prior research has studied the process of dehumanizing humans (Haslam 2006), which hints at the fact that humanizing and dehumanizing phenomena capture a perception about an entity without judging the soundness of the perception. Such a perception

could explain why people treat some objects as humans and some humans as objects (Haslam and Loughnan 2014). In this research, we focused on the humanization process, i.e., anthropomorphism. Nevertheless, future research can investigate the reverse process, i.e., dehumanization, in the context of human-AI interaction. For instance, since we tend to dehumanize the people who are different from us (Vaes et al. 2012), and since stereotypes in human-human interaction sometimes spill over into human-machine interactions (Eyssel and Kuchenbrandt 2012; Nass et al. 1997), the process of dehumanization might also take place in human-AI interaction.

CONCLUSION

As AI artifacts become more complicated, users face more indeterminacies in their interactions with these artifacts. These indeterminacies have important effects on users' perception of the artifacts. In this research, we identified verbal and task fulfillment indeterminacies as two important indeterminacies in the context of conversational agents. Using a custom-developed conversational agent, we investigated the effect of such indeterminacies on users' perception of an artifact's humanness, i.e., anthropomorphism, and their trust toward the artifact. We drew upon psychological accounts on anthropomorphism to explain the phenomenon. We further leveraged the theoretical findings in the uncanny valley literature to explain the interaction effect of multiple indeterminacies on anthropomorphism and trust. The findings from this research are relevant for researchers in the fields of information systems, human-computer interaction, marketing, and psychology.

REFERENCES

Amazon. 2019. “Alexa for Business – Empower Your Organization with Alexa,” *Amazon Web Services, Inc.*

Aruguete, M. S., Huynh, H., Browne, B. L., Jurs, B., Flint, E., and McCutcheon, L. E. 2019. “How Serious Is the ‘Carelessness’ Problem on Mechanical Turk?,” *International Journal of Social Research Methodology* (22:5), Taylor & Francis, pp. 441–449.

Baron, R. M., and Kenny, D. A. 1986. “The Moderator–Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations.,” *Journal of Personality and Social Psychology* (51:6), pp. 1173–1182.

Berkowitz, L., and Donnerstein, E. 1982. “External Validity Is More than Skin Deep: Some Answers to Criticisms of Laboratory Experiments.,” *American Psychologist* (37:3), pp. 245–257.

Bode, S., Murawski, C., Soon, C. S., Bode, P., Stahl, J., and Smith, P. L. 2014. “Demystifying ‘Free Will’: The Role of Contextual Information and Evidence Accumulation for Predictive Brain Activity,” *Neuroscience & Biobehavioral Reviews* (47), pp. 636–645.

Bowden, K. K., Wu, J., Cui, W., Juraska, J., Harrison, V., Schwarzmann, B., Santer, N., and Walker, M. 2019. “SlugBot: Developing a Computational Model and Framework of a Novel Dialogue Genre,” *2nd Proceedings of Alexa Prize*.

Broadbent, E. 2017. “Interactions with Robots: The Truths We Reveal about Ourselves,” *Annual Review of Psychology* (68), pp. 627–652.

Buhrmester, M., Kwang, T., and Gosling, S. D. 2011. “Amazon’s Mechanical Turk: A New Source of Inexpensive, yet High-Quality, Data?,” *Perspectives on Psychological Science* (6:1), pp. 3–5.

Carey, S. 1985. *Conceptual Change in Childhood*, MIT Press.

Caruso, E. M., Waytz, A., and Epley, N. 2010. “The Intentional Mind and the Hot Hand: Perceiving Intentions Makes Streaks Seem Likely to Continue,” *Cognition* (116:1), Elsevier, pp. 149–153.

Castelli, F., Happé, F., Frith, U., and Frith, C. 2000. “Movement and Mind: A Functional Imaging Study of Perception and Interpretation of Complex Intentional Movement Patterns,” *Neuroimage* (12:3), pp. 314–325.

Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J., and Litman, L. 2019. “Online Panels in Social Science Research: Expanding Sampling Methods beyond Mechanical Turk,” *Behavior Research Methods* (51:5), Springer, pp. 2022–2038.

Chen, C.-Y., Yu, D., Wen, W., Yang, Y. M., Zhang, J., Zhou, M., Jesse, K., Chau, A., Bhowmick, A., and Iyer, S. 2019. "Gunrock: Building A Human-Like Social Bot By Leveraging Large Scale Real User Data," *EMNLP 2019*.

Chokshi, N. 2018. "Amazon Knows Why Alexa Was Laughing at Its Customers," *The New York Times*.

Clark, A. 2013. "Whatever next? Predictive Brains, Situated Agents, and the Future of Cognitive Science," *Behavioral and Brain Sciences* (36:3), pp. 181–204.

Columbus, L. 2018. "10 Charts That Will Change Your Perspective On Artificial Intelligence's Growth," *Forbes*.

Coppock, A. 2019. "Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach," *Political Science Research and Methods* (7:3), Cambridge University Press, pp. 613–628.

Darwin, C. 2009. *The Expression of the Emotions in Man and Animals*, (4th ed.), Oxford University Press (Original work published 1872).

Das, T. K., and Teng, B.-S. 1998. "Between Trust and Control: Developing Confidence in Partner Cooperation in Alliances," *Academy of Management Review* (23:3), pp. 491–512.

Dennett, D. C. 2017. *Brainstorms: Philosophical Essays on Mind and Psychology*, MIT press.

Deutsch, M. 1958. "Trust and Suspicion," *Journal of Conflict Resolution* (2:4), pp. 265–279.

Ebert, J. P., and Wegner, D. M. 2011. "Mistaking Randomness for Free Will," *Consciousness and Cognition* (20:3), pp. 965–971.

Epley, N., Akalis, S., Waytz, A., and Cacioppo, J. T. 2008. "Creating Social Connection through Inferential Reproduction: Loneliness and Perceived Agency in Gadgets, Gods, and Greyhounds," *Psychological Science* (19:2), pp. 114–120.

Epley, N., Waytz, A., Akalis, S., and Cacioppo, J. T. 2008. "When We Need a Human: Motivational Determinants of Anthropomorphism," *Social Cognition* (26:2), pp. 143–155.

Epley, N., Waytz, A., and Cacioppo, J. T. 2007. "On Seeing Human: A Three-Factor Theory of Anthropomorphism.," *Psychological Review* (114:4), pp. 864–886.

Eyssel, F., and Kuchenbrandt, D. 2012. "Social Categorization of Social Robots: Anthropomorphism as a Function of Robot Group Membership," *British Journal of Social Psychology* (51:4), pp. 724–731.

Eyssel, F., and Reich, N. 2013. "Loneliness Makes the Heart Grow Fonder (of Robots)—On the Effects of Loneliness on Psychological Anthropomorphism," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, pp. 121–122.

Feldman, G., Wong, K. F. E., and Baumeister, R. F. 2016. "Bad Is Freer than Good: Positive–Negative Asymmetry in Attributions of Free Will," *Consciousness and Cognition* (42), pp. 26–40.

Fornell, C., and Larcker, D. F. 1981. "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error," *Journal of Marketing Research* (18:1), pp. 39–50.

Fridman, L. 2019. *Rohit Prasad: Amazon Alexa and Conversational AI | Artificial Intelligence (AI) Podcast*.

Friston, K. 2005. "A Theory of Cortical Responses," *Philosophical Transactions of the Royal Society of London B: Biological Sciences* (360:1456), pp. 815–836.

Friston, K. 2010. "The Free-Energy Principle: A Unified Brain Theory?," *Nature Reviews Neuroscience* (11:2), p. 127.

Fritz, C. O., Morris, P. E., and Richler, J. J. 2012. "Effect Size Estimates: Current Use, Calculations, and Interpretation.," *Journal of Experimental Psychology: General* (141:1), p. 2.

Gefen, D., Karahanna, E., and Straub, D. W. 2003. "Trust and TAM in Online Shopping: An Integrated Model," *MIS Quarterly* (27:1), pp. 51–90.

Giffin, K. 1967. "The Contribution of Studies of Source Credibility to a Theory of Interpersonal Trust in the Communication Process.," *Psychological Bulletin* (68:2), p. 104.

Goodfellow, I., Bengio, Y., and Courville, A. 2016. *Deep Learning*, MIT press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. 2014. "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems*, pp. 2672–2680.

Gray, H. M., Gray, K., and Wegner, D. M. 2007. "Dimensions of Mind Perception," *Science* (315:5812), pp. 619–619.

Gray, K., Knobe, J., Sheskin, M., Bloom, P., and Barrett, L. F. 2011. "More than a Body: Mind Perception and the Nature of Objectification.," *Journal of Personality and Social Psychology* (101:6), p. 1207.

Gray, K., and Wegner, D. M. 2012. "Feeling Robots and Human Zombies: Mind Perception and the Uncanny Valley," *Cognition* (125:1), pp. 125–130.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., and Parasuraman, R. 2011. "A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction," *Human Factors: The Journal of the Human Factors and Ergonomics Society* (53:5), pp. 517–527.

Hart, P. M., Jones, S. R., and Royne, M. B. 2013. "The Human Lens: How Anthropomorphic Reasoning Varies by Product Complexity and Enhances Personal Value," *Journal of Marketing Management* (29:1–2), pp. 105–121.

Haslam, N. 2006. "Dehumanization: An Integrative Review," *Personality and Social Psychology Review* (10:3), pp. 252–264.

Haslam, N., and Loughnan, S. 2014. "Dehumanization and Infrahumanization," *Annual Review of Psychology* (65), pp. 399–423.

Hayes, A. F., and Preacher, K. J. 2014. "Statistical Mediation Analysis with a Multicategorical Independent Variable," *British Journal of Mathematical and Statistical Psychology* (67:3), pp. 451–470.

Heberlein, A. S., and Adolphs, R. 2004. "Impaired Spontaneous Anthropomorphizing despite Intact Perception and Social Knowledge," *Proceedings of the National Academy of Sciences* (101:19), pp. 7487–7491.

Hu, L., and 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), pp. 1–55.

Iacoboni, M., Lieberman, M. D., Knowlton, B. J., Molnar-Szakacs, I., Moritz, M., Throop, C. J., and 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), pp. 1167–1173.

Inagaki, K., and Hatano, G. 1987. "Young Children's Spontaneous Personification as Analogy," *Child Development*, pp. 1013–1020.

Johnson-George, C., and 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), p. 1306.

JuniperResearch. 2018. "Digital Voice Assistants in Use to Triple to 8 Billion by 2023."

Kahneman, D., and Tversky, A. 1972. "Subjective Probability: A Judgment of Representativeness," *Cognitive Psychology* (3:3), Elsevier, pp. 430–454.

Kay, A. C., Moscovitch, D. A., and Laurin, K. 2010. "Randomness, Attributions of Arousal, and Belief in God," *Psychological Science* (21:2), pp. 216–218.

Kelley, H. H. 1967. "Attribution Theory in Social Psychology.," in *Nebraska Symposium on Motivation*, University of Nebraska Press.

Kline, R. B. 2015. *Principles and Practice of Structural Equation Modeling*, Guilford publications.

Lankton, N. K., McKnight, D. H., and Tripp, J. 2015. "Technology, Humanness, and Trust: Rethinking Trust in Technology," *Journal of the Association for Information Systems* (16:10), pp. 880–918.

Leonardi, P. M. 2008. "Indeterminacy and the Discourse of Inevitability in International Technology Management," *Academy of Management Review* (33:4), Academy of Management Briarcliff Manor, NY, pp. 975–984.

Leviathan, Y. 2018. "Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone," *Google AI Blog*.

Levy, D. 1994. "Chaos Theory and Strategy: Theory, Application, and Managerial Implications," *Strategic Management Journal* (15:S2), pp. 167–178.

Lewis, J. D., and Weigert, A. 1985. "Trust as a Social Reality," *Social Forces* (63:4), pp. 967–985.

Libet, B., Gleason, C. A., Wright, E. W., and Pearl, D. K. 1983. "Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential) the Unconscious Initiation of a Freely Voluntary Act," *Brain* (106:3), pp. 623–642.

Litman, L., Robinson, J., and Abberbock, T. 2017. "TurkPrime. Com: A Versatile Crowdsourcing Data Acquisition Platform for the Behavioral Sciences," *Behavior Research Methods* (49:2), pp. 433–442.

Lombrozo, T. 2006. "The Structure and Function of Explanations," *Trends in Cognitive Sciences* (10:10), pp. 464–470.

Longoni, C., and Morewedge, C. K. 2019. "AI Can Outperform Doctors. So Why Don't Patients Trust It?," *Harvard Business Review*.

Martin, A., and Weisberg, J. 2003. "Neural Foundations for Understanding Social and Mechanical Concepts," *Cognitive Neuropsychology* (20:3–6), pp. 575–587.

Mason, W., and Suri, S. 2012. "Conducting Behavioral Research on Amazon's Mechanical Turk," *Behavior Research Methods* (44:1), pp. 1–23.

Mayer, R. C., and Davis, J. H. 1999. "The Effect of the Performance Appraisal System on Trust for Management: A Field Quasi-Experiment.," *Journal of Applied Psychology* (84:1), p. 123.

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

Mayer, R. C., and 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), pp. 874–888.

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

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., and Ostrovski, G. 2015. "Human-Level Control through Deep Reinforcement Learning," *Nature* (518:7540), p. 529.

Monroe, A. E., Dillon, K. D., and Malle, B. F. 2014. "Bringing Free Will down to Earth: People's Psychological Concept of Free Will and Its Role in Moral Judgment," *Consciousness and Cognition* (27), pp. 100–108.

Mori, M. 1970. "The Uncanny Valley," *Energy* (7:4), pp. 33–35.

Nahmias, E., Shepard, J., and Reuter, S. 2014. "It's OK If 'My Brain Made Me Do It': People's Intuitions about Free Will and Neuroscientific Prediction," *Cognition* (133:2), pp. 502–516.

Nass, C., Moon, Y., and Green, N. 1997. "Are Machines Gender Neutral? Gender-stereotypic Responses to Computers with Voices," *Journal of Applied Social Psychology* (27:10), pp. 864–876.

Oestreicher, C. 2007. "A History of Chaos Theory," *Dialogues in Clinical Neuroscience* (9:3), pp. 279–289.

Orton, J. D., and Weick, K. E. 1990. "Loosely Coupled Systems: A Reconceptualization," *Academy of Management Review* (15:2), Academy of Management Briarcliff Manor, NY 10510, pp. 203–223.

Oskarsson, A. T., Van Boven, L., McClelland, G. H., and Hastie, R. 2009. "What's next? Judging Sequences of Binary Events.," *Psychological Bulletin* (135:2), American Psychological Association, p. 262.

Paolacci, G., Chandler, J., and Ipeirotis, P. G. 2010. "Running Experiments on Amazon Mechanical Turk," *Judgment and Decision Making* (5:5), pp. 411–419.

Pelphrey, K. A., Morris, J. P., and Mccarthy, G. 2004. "Grasping the Intentions of Others: The Perceived Intentionality of an Action Influences Activity in the Superior Temporal Sulcus during Social Perception," *Journal of Cognitive Neuroscience* (16:10), pp. 1706–1716.

Petter, S., Straub, D. W., and Rai, A. 2007. "Specifying Formative Constructs in Information Systems Research," *MIS Quarterly* (31:4), pp. 623–656.

Pew Research Center. 2017. "Many Americans Would Be Hesitant to Use Various Automation Technologies," *Pew Research Center: Internet, Science & Tech*.

Pittman, T. S., and Pittman, N. L. 1980. "Deprivation of Control and the Attribution Process.," *Journal of Personality and Social Psychology* (39:3), p. 377.

Preacher, K. J., and Hayes, A. F. 2008. "Asymptotic and Resampling Strategies for Assessing and Comparing Indirect Effects in Multiple Mediator Models," *Behavior Research Methods* (40:3), pp. 879–891.

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

Rao, R. P., and Ballard, D. H. 1999. "Predictive Coding in the Visual Cortex: A Functional Interpretation of Some Extra-Classical Receptive-Field Effects," *Nature Neuroscience* (2:1), pp. 79–87.

Rempel, J. K., Holmes, J. G., and Zanna, M. P. 1985. "Trust in Close Relationships.," *Journal of Personality and Social Psychology* (49:1), p. 95.

Riedl, R., Hubert, M., and Kenning, P. 2010. "Are There Neural Gender Differences in Online Trust? An FMRI Study on the Perceived Trustworthiness of EBay Offers," *MIS Quarterly* (34:2), pp. 397–428.

Robinson, J., Rosenzweig, C., Moss, A. J., and 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), Public Library of Science.

Rotter, J. B. 1967. "A New Scale for the Measurement of Interpersonal Trust," *Journal of Personality* (35:4), pp. 651–665.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., and Camerer, C. 1998. "Not so Different after All: A Cross-Discipline View of Trust," *Academy of Management Review* (23:3), pp. 393–404.

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

Sarkissian, H., Chatterjee, A., De Brigard, F., Knobe, J., Nichols, S., and Sirker, S. 2010. "Is Belief in Free Will a Cultural Universal?," *Mind & Language* (25:3), pp. 346–358.

Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J., and 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), pp. 413–422.

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., and Vogeley, K. 2013. "Toward a Second-Person Neuroscience," *Behavioral and Brain Sciences* (36:4), pp. 393–414.

Schoorman, F. D., Mayer, R. C., and Davis, J. H. 2007. "An Integrative Model of Organizational Trust: Past, Present, and Future," *Academy of Management Review* (32:2), pp. 344–354.

Seeger, A.-M., Pfeiffer, J., and Heinzl, A. 2018. *Designing Anthropomorphic Conversational Agents: Development and Empirical Evaluation of a Design Framework*.

Shattuck, S. 2019. “People Don’t Trust AI. We Need to Change That.,” *Medium*, , February 26.

Shepherd, J. 2012. “Free Will and Consciousness: Experimental Studies,” *Consciousness and Cognition* (21:2), pp. 915–927.

Srivastava, S. C., and Chandra, S. 2018. “Social Presence in Virtual World Collaboration: An Uncertainty Reduction Perspective Using a Mixed Methods Approach,” *MIS Quarterly* (42:3), pp. 779–803.

Thietart, R.-A., and Forgues, B. 1995. “Chaos Theory and Organization,” *Organization Science* (6:1), pp. 19–31.

Towers-Clark, C. 2019. “80% Of People Don’t Trust AI With Money - How Can We Fix Its Image?,” *Forbes*.

Vaes, J., Leyens, J.-P., Paola Paladino, M., and Pires Miranda, M. 2012. “We Are Human, They Are Not: Driving Forces behind Outgroup Dehumanisation and the Humanisation of the Ingroup,” *European Review of Social Psychology* (23:1), pp. 64–106.

Voicebot.ai. 2019. “Juniper Estimates 3.25 Billion Voice Assistants Are in Use Today, Google Has About 30% of Them,” *Voicebot*, , February 14.

Waytz, A., Cacioppo, J., and Epley, N. 2010. “Who Sees Human? The Stability and Importance of Individual Differences in Anthropomorphism,” *Perspectives on Psychological Science* (5:3), pp. 219–232.

Waytz, A., Heafner, J., and Epley, N. 2014. “The Mind in the Machine: Anthropomorphism Increases Trust in an Autonomous Vehicle,” *Journal of Experimental Social Psychology* (52), pp. 113–117.

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J.-H., and Cacioppo, J. T. 2010. “Making Sense by Making Sentient: Effectance Motivation Increases Anthropomorphism.,” *Journal of Personality and Social Psychology* (99:3), pp. 410–435.

Wegner, D. M. 2008. “Self Is Magic,” in *Are We Free? Psychology and Free Will*, New York, NY, US: Oxford University Press, pp. 226–247.

Wegner, D. M., and Wheatley, T. 1999. “Apparent Mental Causation: Sources of the Experience of Will.,” *American Psychologist* (54:7), p. 480.

White, R. W. 1959. “Motivation Reconsidered: The Concept of Competence.,” *Psychological Review* (66:5), p. 297.

Xia, Y. 2016. “Investigating the Chi-Square-Based Model-Fit Indexes for WLSMV and ULSMV Estimators,” *Doctoral Dissertation*.

Yuan, L., and Dennis, A. R. 2019. “Acting Like Humans? Anthropomorphism and Consumer’s Willingness to Pay in Electronic Commerce,” *Journal of Management Information Systems* (36:2), pp. 450–477.

Zhao, X., Lynch Jr, J. G., and Chen, Q. 2010. “Reconsidering Baron and Kenny: Myths and Truths about Mediation Analysis,” *Journal of Consumer Research* (37:2), pp. 197–206.

APPENDIX A – Anthropomorphism in Extant Literature

Table A1. An Interdisciplinary Literature Summary of Anthropomorphism, Humanness, and, Agency Detection

	Authors	Research Questions / Objectives	Method	Key Findings
Information Systems	(Qiu and Benbasat 2009)	What is the effect of anthropomorphic features, namely humanoid embodiment and voice output, on user's perceived social relationship with a technological artifact designed for electronic commerce contexts?	Single experiment	Anthropomorphic features increase perceived social presence, which in turn increases trusting beliefs, perception of enjoyment, and intention to use.
	(Seeger et al. 2017)	Can the agent substitution type explain the contradicting findings about the trust-inducing effect of anthropomorphic design?	Conceptual	They theorize that agent substitution type, i.e., whether the agent is a substitute for a human or a system, moderates the positive relationship between anthropomorphism and trusting beliefs.
	(Yuan et al. 2016)	How does anthropomorphism influence individual's cognition processes during online bidding?	Single experiment with EEG	Non-Caucasian consumers bid more on an anthropomorphic product, because of a non-rational cognitive process. They provided some insight into the cognitive process in which anthropomorphism changes bidding behavior.
Marketing	(Hart et al. 2013)	What is the relationship between consumer anthropomorphism and personal value?	Survey	Anthropomorphism can account for the personal value of a product beyond the influence of common marketplace factors. The magnitude of consumer anthropomorphism will be greater for complex products than simple products.
	(Chandler and Schwarz 2010)	Does thinking of objects as alive make people less willing to replace them?	Two experiments	Consumers who think about their cars in anthropomorphic terms are less willing to replace it and give less weight to its quality when making replacement decisions.
	(Kim and McGill 2011)	What is the effect of anthropomorphism on risk perception and what is the moderating effect of individual's feeling of social power?	Three experiments	People with high (low) power perceive anthropomorphic risk-bearing entities, i.e., entities with anthropomorphic features, as less (more) risky compared to non-anthropomorphic entities.

	(Touré-Tillery and McGill 2015)	How may differences in people's levels of trust in human agents influence the persuasiveness of anthropomorphized messengers compared with human messengers?	Three experiments	People low in the generalized interpersonal trust are more persuaded by anthropomorphized messengers than by human spokespeople. People high in interpersonal trust respond similarly to human and anthropomorphized messengers. However, when asked to be attentive, they are more persuaded by human spokespeople than by anthropomorphized messengers.
	(Aggarwal and McGill 2011)	What is the effect of anthropomorphizing a brand on automatic behavior in response to a brand prime?	Three experiments	Anthropomorphized brands trigger people's goals for a successful social interaction, resulting in behavior that is assimilative or contrastive to the brand's image. Consumers are more likely to assimilate behavior associated with anthropomorphized partner brands that they like and servant brands that they dislike. Consumers show a contrastive behavior when primed with disliked partner brands and liked servant brands.
	(Hellen and Sääksjärvi 2013)	The objective of this study is to provide a conceptualization and measurement for childlike anthropomorphic characteristics in products.	Three surveys	They found that childlike characteristics comprise four dimensions: sweetness, simplicity, sympathy, and smallness. Consumers react positively to childlike anthropomorphic characteristics in products.
	(May and Monga 2013)	The objective of this paper is to introduce anthropomorphism of time.	One field survey and four experiments	They showed that time anthropomorphism influences intertemporal preferences. The effect is different for people with perceived low and high power. They argue that time anthropomorphism happens probably for three reasons: the general prevalence of anthropomorphism, the linguistic portrayal of time as a human, and the moving nature of time.

	(Van den Hende and Mugge 2014)	-	Two experiments	When a human gender schema is primed, that is, congruent with consumers' own gender, consumers show more preferential evaluations and are more likely to perceive the product as human, even when no product-schema congruent features are present in the product. Results indicated that perceived anthropomorphism mediates the gender-schema congruity effect and the product-schema congruity effect on product evaluations.
Psychology	(Waytz, Morewedge, et al. 2010)	What is the effect of effectance motivation on anthropomorphism?	Six studies including survey, experiment, and fMRI	People anthropomorphize, in part, to satisfy effectance motivation. Unpredictability increases anthropomorphism.
	(Waytz, Cacioppo, et al. 2010)	To provide a measure of stable individual differences in anthropomorphism.	Eight survey studies (EFA, CFA, and correlational analysis)	They provide a psychometrically valid measure named the Individual Differences in Anthropomorphism Questionnaire (IDAQ). They showed that IDAQ is significantly associated with the degree of moral care and concern afforded to an agent, the amount of responsibility and trust placed on an agent, and the extent to which an agent serves as a source of social influence on the self.
	(Epley, Waytz, et al. 2008)	To empirically test the effect of sociality motivation and effectance motivation on anthropomorphism.	One survey and one experiment	While the results are mostly correlational, they provide some preliminary evidence of the role of sociality and effectance motivation in increasing anthropomorphism.
	(Epley, Akalis, et al. 2008)	What is the effect of sociality motivation on anthropomorphism?	One survey and two experiments	They found that individuals who are chronically lonely and who are induced to feel lonely are more likely to anthropomorphize nonhuman agents. They showed that the results are not simply produced by any negative affective state.
	(Waytz et al. 2014)	What is the effect of anthropomorphism on trust?	Single experiment using driving simulator with an autonomous vehicle	They showed that anthropomorphism increases self-reported trust as well as its physiological and behavioral consequences.

	(Waytz, Gray, et al. 2010)	When mind perception occurs, when it does not and why mind perception is important?	Short literature review	They argue that perception of mind has different causes and consequences in the entity that perceives mind in another entity as well as in the entity that is perceived to possess a mind.
	(Epley et al. 2007)	What are the psychological determinants of anthropomorphism?	Theory development	They propose a theory that offers three determinants of anthropomorphism, namely elicited agent knowledge, effectance motivation, sociality motivation.
	(Gray et al. 2007)	What are the dimensions of mind perception?	Survey	They found two dimensions of mind perception, namely agency and experience.
	(Gray et al. 2012)	What is the relationship between mind perception and moral judgment?	Conceptual framework	They suggest that moral judgment is rooted in a cognitive template of two perceived minds, a moral dyad of an intentional agent and a suffering moral patient (dyadic morality). They posit that human mind abstracts out the key elements from various moral transgressions to create a cognitive template. They argue that these key elements are intention and pain.
	(Wiese et al. 2017)	How can we use neuroscientific methods to make robots appear more social?	Literature review	They suggest that we can make people perceive artificial agents as social companions by designing them as intentional agents, because such agents activate areas in the human brain involved in social-cognitive processing.
	(Kay et al. 2010)	What is the effect of thoughts of randomness on beliefs in supernatural sources of control?	Single experiment	They observed that participants primed with randomness-related words exhibited heightened beliefs in spiritual control compared with participants primed with negatively valenced control words. This effect disappeared when participants were given the opportunity to attribute the cause of any arousal they experienced to a pill ingested earlier in the session. They suggest that belief in supernatural sources of control, such as God and karma, may function, in part, to defend against distress associated with randomness.

	(Valdesolo and Graham 2014)	What is the effect of experienced awe on agency detection?	Five experiments	They found that that experiencing awe, while controlling for some other emotional states, increases agency detection in the context of both supernatural belief and judgments of intentional design. This effect is partially mediated by awe-induced changes in person's tolerance for ambiguity and uncertainty.
	(Schroeder et al. 2017)	What is the effect of voice on anthropomorphism?	Four experiments	The human voice contains paralinguistic cues that reveal underlying mental processing involved in thinking and feeling. The medium of communication may moderate the tendency to dehumanize the opposition. Adding visual cues to voice did not increase anthropomorphism. Individuals with voices that lack authentic intonation (e.g., monotone voices) may be perceived as less humanlike than others.
	(Schroeder and Epley 2016)	How does a cue closely connected to a person's actual mental experience—a humanlike voice—affect the likelihood of mistaking a person for a machine, or a machine for a person?	Four experiments	Removing voice from communication (leaving only text) increases the likelihood of mistaking the text's creator for a machine. Adding voice to a computer-generated script (resulting in speech) would increase the likelihood of mistaking the text's creator for a human. People are more likely to infer a human (vs. computer) creator when they hear a voice expressing thoughts than when they read the same thoughts in text. Removing the naturalistic paralinguistic cues that convey humanlike capacity for thinking and feeling, such as varied pace and intonation, eliminates the humanizing effect of speech. Adding visual cues, such as a video clip, did not increase the likelihood of inferring a human creator compared with only reading text.
	(Broadbent 2017)	This article strives to review the research on the psychology behind	Literature review	This article reviews applications of robots and research on how humans relate to robots, explores concerns

		our relationship with social robots.		about robots, and looks ahead to the future of the field.
	(Khalid et al. 2016)	What is the effect of eye contact on mind perception?	Four experiments	Direct eye gaze increases explicit mind ascription and beliefs about the likelihood of mind possession.
	(van Elk et al. 2016)	What is the effect of processing concepts about supernatural agents on agency detection?	Five experiments	They did not find an overall effect of supernatural priming on agency detection. They found that for religious individuals supernatural primes influence agency and face detection, but they failed to find the same effect for non-religious individuals.
	(Brandt and Reyna 2011)	The goal of this article, and the conceptualization of the social cognitive chain of being (SCCB), is to provide a broad, inclusive framework for thinking and theorizing about morality.	Theory development	The SCCB serves as a unifying theoretical framework that organizes research on moral perception, highlights unique interconnections, and provides a roadmap for future research.
	(Barrett and Keil 1996)	What is the role of anthropomorphism in God concepts?	Three experiments	People do use anthropomorphic concepts of God in understanding stories even though they may profess a theological position that rejects anthropomorphic constraints on God and God's activities.
	(Vaes et al. 2016)	Is the attribution of humanness by means of a minimal humanity cue sufficient for the occurrence of empathic neural reactions towards non-human entities that are painfully stimulated?	Single experiment with additional EEG and ERP data	Their findings suggest that empathy can be triggered for non-human entities as long as they are seen as minimally human.
	(Deska et al. 2018)	What is the effect of facial width-to-height ratio (fWHR) on perceived humanness?	Ten surveys and two experiments	Individuals with relatively greater facial width-to-height ratio are routinely denied sophisticated, humanlike minds.
	(Johnson and Barrett 2003)	What is the role of control in attributing intentional agency to inanimate objects?	Single experiment	Individuals who do not have control over the movement of an entity are more likely to attribute agency to it than individuals who think have indirect control. It appears that when the movement of objects is explainable in terms of individuals' own agency, no agent-attributions are triggered; but when the movement or activity of objects

				exceeds obvious agency, the objects themselves are attributed agency.
	(Bering 2002)	To develop a very basic, species-wide existential theory of mind (EToM) as an independent system built on the foundations of the theory of mind.	Theory development	EToM functions as a philosophical–religious explanatory system that allows us to see meaning in some of the things that happen to us, affords us some sense of perceived psychological control over what is likely to happen, enforces cultural mores that adapt the individual to the group, and guards against those behaviors that are maladaptive.
	(Demoulin et al. 2004)	The extent to which some emotions are explicitly qualified as uniquely reserved to human beings. The criteria on which people base their judgment.	One survey in four languages and one experiment	People distinguish between uniquely human and non-uniquely human emotions. This maps to the difference between primary and secondary emotions used by emotion scientists.
	(Haslam et al. 2005)	Do people attribute greater humanness to themselves than to others?	Three surveys and one experiment	Human nature characteristics differ from uniquely human characteristics. People tend to attribute more human nature to themselves than others (self-humanization). This effect is different from self-enhancement.
	(Loughnan and Haslam 2007)	What are the implicit association between social categories and senses of humanness, traits representing these senses, and the two types of nonhumans (animals and androids)?	One survey with go/no-go association task	Humanness traits are differentially associated with distinct types of nonhumans: Uniquely human traits are associated with automata more than with animals, and human-nature traits are associated with animals more than with automata.
	(Stenzel et al. 2012)	Do humans co-represent actions of a humanoid robot?	Single experiment	Findings suggest that humans co-represent the actions of robotic agents in a human-robot team (i.e., cognitively represent the action of the agents as if they were in charge of the full, undivided task) when they start to attribute human-like cognitive processes to the robot. They argue that action co-representation is related to perceived humanlike cognitive processes. The robot that was described as having purely deterministic behavior was not co-represented as much as the one

				described as being biologically inspired.
	(Haslam 2006)	To review the literature on dehumanization and develop a new model of dehumanization.	Literature review	The author developed a new model of dehumanization and proposed two forms of dehumanization, namely denying uniquely human and denying human nature characteristics.
Neuroscience	(Tononi and Koch 2015)	What experience is and what type of physical systems can have consciousness?	Theory development	They propose integrated information theory (IIT) that is a theory of consciousness and that introduces five phenomenological axioms for experience of consciousness: intrinsic existence, composition, information, integration and exclusion.
	(Saygin et al. 2011)	What is the effect of violations in brain's prediction on the uncanny valley phenomenon?	fMRI	The uncanny valley is, at least partially, caused by the violation of the brain's predictions. When an agent looks like a human, based on a lifetime of experience, the brain generates a prediction that this appearance will be associated with a particular kind of behavior. When the behavior of the agent violates the prediction, an error is generated.
	(Vogeley and Bente 2010)	How can we address the challenges that emerge from the goal to equip machines with socioemotional intelligence and to enable them to interpret subtle nonverbal cues and to respond to social affordances with naturally appearing behavior from both perspectives?	Conceptual	They propose that the creation of credible artificial humans not only defines the ultimate test for our understanding of human communication and social cognition but also provides a unique research tool to improve our knowledge about the underlying psychological processes and neural mechanisms.
Human-Computer Interaction	(Ruijten et al. 2015)	What are the effects of social exclusion on persuasion by an artificial agent?	Two experiments	Socially excluded people anthropomorphize and are persuaded more by an artificial agent than socially included people.
	(Kim and Sundar 2012)	Is the anthropomorphism of computers mindful or mindless?	One experiment	They found evidence for mindless anthropomorphism. People perceive a human-like agent in more human-related terms; however, they report lesser perceived humanness compared to an agent that is not human-like.

	(Lee et al. 2015)	What is the effect of anthropomorphic cues on perceived safety and trust in unmanned driving systems? What is the mediating role of social presence in this relationship?	Single experiment	Human-like appearance and high autonomy are more effective in eliciting positive perceptions of the agent. The greater level of anthropomorphism induced by human-like appearance and high autonomy in the agent evoked the feelings of social presence, which in turn positively affected the perceived intelligence and safety of and trust in the agent.
	(Ho and MacDorman 2010)	To develop measures for attractiveness, eeriness, humanness, and warmth.	Two surveys	They developed measures for the four constructs with non-significant intercorrelations among the constructs.
	(Schmitz 2011)	To provide an interdisciplinary review of the work that can inform anthropomorphism and zoomorphism.	Literature review	They provided a review spanning the disciplines of anthropomorphism, affective computing, tangible interaction and industrial design.
	(Eyssel and Reich 2013)	What is the effect of feeling of loneliness on anthropomorphism?	Single experiment	Lonely people anthropomorphize robots. Users' motivational states need to be considered in the context of human-robot interaction (HRI) as they affect judgments of the robotic interaction partner.
	(Lee 2010)	What is the effect of anthropomorphic interfaces, namely speech type, on people's tendency to project social expectations onto computers? What is the moderating effect of users' cognitive style on this relationship?	Single experiment	Intuition-driven individuals evaluate a human-voice computer more positively and conform more to its suggestions compared to a synthetic-voice computer. However, such results were not found for analytical people.
	(Candello et al. 2017)	What is the effect of typeface (font type) on perceived humanness?	Two experiments	Machine-like typefaces bias users towards perceiving the adviser as a machine but, unexpectedly, handwritten-like typefaces did not have the opposite effect.
	(Burleigh et al. 2013)	What is the relationship between humanlikeness and eeriness in digitally created faces?	Two experiments	They found that humanlikeness is linearly related to emotional response. This relationship changes when humanlikeness varies by category membership. They argue that previous non-linear relationship observed in uncanny valley literature

				might be explained by the conflict in ontological categories of humanlike features.
Robotics	(Fink 2012)	What is the role of anthropomorphism in the design of socially interactive robots and human-robot interaction?	Short literature review	Anthropomorphism is a phenomenon that is hard to grasp because of the broad understanding about it and its usage in a variety of disciplines. It is hard to draw general conclusions about anthropomorphism because of contradictory findings.
	(Fussell et al. 2008)	What are the effects of people's level of abstraction of human-robot interaction (people's reactions to a robot in social context vs. their thoughts about the robot) and robot's politeness on anthropomorphism?	Single experiment	People are more likely to anthropomorphize a specific behavior and a robot's personality characteristics than to anthropomorphize the robot as a whole.
	(Lemaignan, Fink, and Dillenbourg 2014)	How anthropomorphism evolves over time?	Conceptual framework	They propose that anthropomorphism goes through three stages namely initialization, familiarization, and stabilization. They argue that anthropomorphism increases during initialization, sharply decreases during familiarization, and gradually decrease to a stable level during stabilization.
	(Kahn Jr et al. 2007)	From the standpoint of human-robot interaction, how do we measure success?	Conceptual framework	They offered nine psychological benchmarks to measure success in building increasingly humanlike robots.
	(Salem et al. 2013)	What are the effects of the robot's hand and arm gestures on the perception of humanlikeness, likability of the robot, shared reality, and future contact intentions after interacting with a robot?	Single experiment	They found that co-verbal gestures (i.e., gestures that accompany verbal utterances) in a robot increases its anthropomorphism, likability, shared reality, and future contact intentions than when the robot gave instructions without gestures. Surprisingly, this effect was particularly pronounced when the robot's gestures were partly incongruent with speech, although this behavior negatively affected the participants' task-related performance.

	(Lemaignan, Fink, Dillenbourg, et al. 2014)	What are the cognitive phases corresponding to anthropomorphism in a sustained human-robot interaction?	Conceptual framework	They propose three cognitive phases namely pre-cognitive, familiarity-based, and adapted anthropomorphism.
	(Złotowski et al. 2015)	What are the potential benefits and challenges of building anthropomorphic robots, from both a philosophical perspective and from the viewpoint of empirical research in the fields of human-robot interaction and social psychology?	Literature review	They discussed the findings from prior research and delineated benefits and problems associated with anthropomorphism and anthropomorphic design in human-robot interaction.
	(Bartneck et al. 2010)	What is the degree to which the human model of embarrassment translates to robot? What is the effect of anthropomorphism on the experience of embarrassment?	Single experiment	In the medical context, people are less embarrassed when interacting with a technical box than with a robot.
	(Eyssel 2017)	How could the scientific community in social robotics potentially gain from experimental psychology?	Literature review	They emphasized the importance of a theory-driven approach to test causal relationships, development of valid measures, and bridging the gap between foundational and applied research.
	(Riek et al. 2009)	How do people empathize with robots along the anthropomorphic spectrum?	One survey	People empathize more strongly with more human-looking robots and less with mechanical-looking robots.
	(Bartneck et al. 2009)	To find comparable standardized measures for anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety.	Literature review	They report several items to measure anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety.
	(Duffy 2003)	How can the concept of anthropomorphism be used in the development of meaningful social interaction between robots and people?	Literature review	While anthropomorphism is a very complex notion, it intuitively provides us with very powerful physical and social features that can be implemented to a greater extent in social robotics research.

Artificial Intelligence	(Floridi and Sanders 2004)	What is a moral agent?	Short literature review	They clarify the concept of agent and separate the concerns of morality and responsibility of agents.
	(Proudfoot 2011)	How can any putative demonstration of intelligence in machines be trusted if the AI researcher readily succumbs to make-believe?	Literature review	This paper illustrates the phenomenon of misplaced anthropomorphism and presents a new perspective on Turing's imitation game.
Communication	(Nowak and Rauh 2005)	What is the influence of an avatar on anthropomorphism, androgyny, credibility, homophily, and attraction?	One survey	Anthropomorphic avatars significantly impact perceptions of avatars. The results were in line with the uncertainty reduction theory.
	(Nowak and Biocca 2003)	What is the effect of anthropomorphic features on presence, co-presence, and social presence in a virtual environment?	Single experiment	The existence of a virtual image increases telepresence. Participants interacting with the less-anthropomorphic image reported more co-presence (i.e., to actively perceive an agent and to feel that the agent actively perceives oneself) and social presence than those interacting with partners represented by either no image at all or by a highly anthropomorphic image of the other, indicating that the more anthropomorphic images set up higher expectations that lead to reduced presence when these expectations were not met.
Education	(Bernstein and Crowley 2008)	What is the impact of experience with intelligent technologies on children's ideas about robot intelligence?	Controlled survey in lab	As children gain experience in this domain, they begin to differentiate robots from other familiar entities.

Note: we adjusted research questions and findings to use the same terminology that we used in our research. For instance, we replaced the term "humanization" with "anthropomorphism" when the authors used it to refer to the same construct. Also, we replaced the term "anthropomorphism" with "anthropomorphizing design" or "anthropomorphic features" when the authors used the term "anthropomorphism" to refer some cues of anthropomorphism, inconsistent with the definition used in our research.

Note: we inferred research questions from studies that did not explicitly state their research question or objective.

REFERENCES

Aggarwal, P., and McGill, A. L. 2011. "When Brands Seem Human, Do Humans Act like Brands? Automatic Behavioral Priming Effects of Brand Anthropomorphism," *Journal of Consumer Research* (39:2), pp. 307–323. (doi: 10.1086/662614).

Barrett, J. L., and Keil, F. C. 1996. "Conceptualizing a Nonnatural Entity: Anthropomorphism in God Concepts," *Cognitive Psychology* (31:3), pp. 219–247.

Bartneck, C., Bleeker, T., Bun, J., Fens, P., and Riet, L. 2010. "The Influence of Robot Anthropomorphism on the Feelings of Embarrassment When Interacting with Robots," *Paladyn, Journal of Behavioral Robotics* (1:2), pp. 109–115.

Bartneck, C., Kulić, D., Croft, E., and Zoghbi, S. 2009. "Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots," *International Journal of Social Robotics* (1:1), pp. 71–81.

Bering, J. M. 2002. "The Existential Theory of Mind," *Review of General Psychology* (6:1), pp. 3–24.

Bernstein, D., and Crowley, K. 2008. "Searching for Signs of Intelligent Life: An Investigation of Young Children's Beliefs about Robot Intelligence," *The Journal of the Learning Sciences* (17:2), pp. 225–247.

Brandt, M. J., and Reyna, C. 2011. "The Chain of Being: A Hierarchy of Morality," *Perspectives on Psychological Science* (6:5), pp. 428–446.

Broadbent, E. 2017. "Interactions with Robots: The Truths We Reveal about Ourselves," *Annual Review of Psychology* (68), pp. 627–652. (doi: 10.1146/annurev-psych-010416-043958).

Burleigh, T. J., Schoenherr, J. R., and Lacroix, G. L. 2013. "Does the Uncanny Valley Exist? An Empirical Test of the Relationship between Eeriness and the Human Likeness of Digitally Created Faces," *Computers in Human Behavior* (29:3), pp. 759–771.

Candello, H., Pinhanez, C., and Figueiredo, F. 2017. "Typefaces and the Perception of Humanness in Natural Language Chatbots," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ACM, pp. 3476–3487.

Chandler, J., and Schwarz, N. 2010. "Use Does Not Wear Ragged the Fabric of Friendship: Thinking of Objects as Alive Makes People Less Willing to Replace Them," *Journal of Consumer Psychology* (20:2), pp. 138–145.

Demoulin, S., Leyens, J.-P., Paladino, M.-P., Rodriguez-Torres, R., Rodriguez-Perez, A., and Dovidio, J. 2004. "Dimensions of 'Uniquely' and 'Non-uniquely' Human Emotions," *Cognition and Emotion* (18:1), pp. 71–96.

Deska, J. C., Lloyd, E. P., and Hugenberg, K. 2018. "Facing Humanness: Facial Width-to-Height Ratio Predicts Ascriptions of Humanity.," *Journal of Personality and Social Psychology* (114:1), pp. 75–94. (doi: 10.1037/pspi0000110).

Duffy, B. R. 2003. "Anthropomorphism and the Social Robot," *Robotics and Autonomous Systems* (42:3–4), pp. 177–190.

van Elk, M., Rutjens, B. T., van der Pligt, J., and Van Harreveld, F. 2016. "Priming of Supernatural Agent Concepts and Agency Detection," *Religion, Brain & Behavior* (6:1), pp. 4–33.

Epley, N., Akalis, S., Waytz, A., and Cacioppo, J. T. 2008. "Creating Social Connection through Inferential Reproduction: Loneliness and Perceived Agency in Gadgets, Gods, and Greyhounds," *Psychological Science* (19:2), pp. 114–120. (doi: 10.1111/j.1467-9280.2008.02056.x).

Epley, N., Waytz, A., Akalis, S., and Cacioppo, J. T. 2008. "When We Need a Human: Motivational Determinants of Anthropomorphism," *Social Cognition* (26:2), pp. 143–155. (doi: 10.1521/soco.2008.26.2.143).

Epley, N., Waytz, A., and Cacioppo, J. T. 2007. "On Seeing Human: A Three-Factor Theory of Anthropomorphism.," *Psychological Review* (114:4), pp. 864–886. (doi: 10.1037/0033-295X.114.4.864).

Eyssel, F. 2017. "An Experimental Psychological Perspective on Social Robotics," *Robotics and Autonomous Systems* (87), pp. 363–371. (doi: 10.1111/j.2044-8309.2011.02082.x).

Eyssel, F., and Reich, N. 2013. "Loneliness Makes the Heart Grow Fonder (of Robots)—On the Effects of Loneliness on Psychological Anthropomorphism," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, pp. 121–122.

Fink, J. 2012. "Anthropomorphism and Human Likeness in the Design of Robots and Human-Robot Interaction," in *International Conference on Social Robotics*, Springer, pp. 199–208.

Floridi, L., and Sanders, J. W. 2004. "On the Morality of Artificial Agents," *Minds and Machines* (14:3), pp. 349–379.

Fussell, S. R., Kiesler, S., Setlock, L. D., and Yew, V. 2008. "How People Anthropomorphize Robots," in *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, pp. 145–152.

Gray, H. M., Gray, K., and Wegner, D. M. 2007. "Dimensions of Mind Perception," *Science* (315:5812), pp. 619–619. (doi: 10.1126/science.1134475).

Gray, K., Young, L., and Waytz, A. 2012. “Mind Perception Is the Essence of Morality,” *Psychological Inquiry* (23:2), pp. 101–124.

Hart, P. M., Jones, S. R., and Royne, M. B. 2013. “The Human Lens: How Anthropomorphic Reasoning Varies by Product Complexity and Enhances Personal Value,” *Journal of Marketing Management* (29:1–2), pp. 105–121. (doi: 10.1080/0267257X.2012.759993).

Haslam, N. 2006. “Dehumanization: An Integrative Review,” *Personality and Social Psychology Review* (10:3), pp. 252–264. (doi: 10.1207/s15327957pspr1003_4).

Haslam, N., Bain, P., Douge, L., Lee, M., and Bastian, B. 2005. “More Human than You: Attributing Humanness to Self and Others.,” *Journal of Personality and Social Psychology* (89:6), pp. 937–950. (doi: 10.1037/0022-3514.89.6.937).

Hellen, K., and Sääksjärvi, M. 2013. “Development of a Scale Measuring Childlike Anthropomorphism in Products,” *Journal of Marketing Management* (29:1–2), pp. 141–157.

Ho, C.-C., and MacDorman, K. F. 2010. “Revisiting the Uncanny Valley Theory: Developing and Validating an Alternative to the Godspeed Indices,” *Computers in Human Behavior* (26:6), pp. 1508–1518.

Johnson, A. H., and Barrett, J. 2003. “The Role of Control in Attributing Intentional Agency to Inanimate Objects,” *Journal of Cognition and Culture* (3:3), pp. 208–217.

Kahn Jr, P. H., Ishiguro, H., Friedman, B., Kanda, T., Freier, N. G., Severson, R. L., and Miller, J. 2007. “What Is a Human?: Toward Psychological Benchmarks in the Field of Human–Robot Interaction,” *Interaction Studies* (8:3), pp. 363–390.

Kay, A. C., Moscovitch, D. A., and Laurin, K. 2010. “Randomness, Attributions of Arousal, and Belief in God,” *Psychological Science* (21:2), pp. 216–218. (doi: 10.1177/0956797609357750).

Khalid, S., Deska, J. C., and Hugenberg, K. 2016. “The Eyes Are the Windows to the Mind: Direct Eye Gaze Triggers the Ascription of Others’ Minds,” *Personality and Social Psychology Bulletin* (42:12), pp. 1666–1677.

Kim, S., and McGill, A. L. 2011. “Gaming with Mr. Slot or Gaming the Slot Machine? Power, Anthropomorphism, and Risk Perception,” *Journal of Consumer Research* (38:1), pp. 94–107.

Kim, Y., and Sundar, S. S. 2012. “Anthropomorphism of Computers: Is It Mindful or Mindless?,” *Computers in Human Behavior* (28:1), pp. 241–250. (doi: 10.1016/j.chb.2011.09.006).

Lee, E.-J. 2010. “The More Humanlike, the Better? How Speech Type and Users’ Cognitive Style Affect Social Responses to Computers,” *Computers in Human Behavior* (26:4), pp. 665–672. (doi: 10.1016/j.chb.2010.01.003).

Lee, J.-G., Kim, K. J., Lee, S., and Shin, D.-H. 2015. "Can Autonomous Vehicles Be Safe and Trustworthy? Effects of Appearance and Autonomy of Unmanned Driving Systems," *International Journal of Human-Computer Interaction* (31:10), pp. 682–691.

Lemaignan, S., Fink, J., and Dillenbourg, P. 2014. "The Dynamics of Anthropomorphism in Robotics," in *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, pp. 226–227.

Lemaignan, S., Fink, J., Dillenbourg, P., and Braboszcz, C. 2014. "The Cognitive Correlates of Anthropomorphism," in *HRI: A Bridge between Robotics and Neuroscience*.

Loughnan, S., and Haslam, N. 2007. "Animals and Androids: Implicit Associations between Social Categories and Nonhumans," *Psychological Science* (18:2), pp. 116–121.

May, F., and Monga, A. 2013. "When Time Has a Will of Its Own, the Powerless Don't Have the Will to Wait: Anthropomorphism of Time Can Decrease Patience," *Journal of Consumer Research* (40:5), pp. 924–942.

Nowak, K. L., and Biocca, F. 2003. "The Effect of the Agency and Anthropomorphism on Users' Sense of Telepresence, Copresence, and Social Presence in Virtual Environments," *Presence: Teleoperators & Virtual Environments* (12:5), pp. 481–494.

Nowak, K. L., and Rauh, C. 2005. "The Influence of the Avatar on Online Perceptions of Anthropomorphism, Androgyny, Credibility, Homophily, and Attraction," *Journal of Computer-Mediated Communication* (11:1), pp. 153–178.

Proudfoot, D. 2011. "Anthropomorphism and AI: Turing's Much Misunderstood Imitation Game," *Artificial Intelligence* (175:5–6), pp. 950–957.

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

Riek, L. D., Rabinowitch, T.-C., Chakrabarti, B., and Robinson, P. 2009. "How Anthropomorphism Affects Empathy toward Robots," in *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, ACM, pp. 245–246.

Ruijten, P. A., Midden, C. J., and 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), pp. 832–842.

Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., and Joublin, F. 2013. "To Err Is Human (-like): Effects of Robot Gesture on Perceived Anthropomorphism and Likability," *International Journal of Social Robotics* (5:3), pp. 313–323.

Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J., and 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), pp. 413–422. (doi: 10.1093/scan/nsr025).

Schmitz, M. 2011. "Concepts for Life-like Interactive Objects," in *Proceedings of the Fifth International Conference on Tangible, Embedded, and Embodied Interaction*, ACM, pp. 157–164.

Schroeder, J., and Epley, N. 2016. "Mistaking Minds and Machines: How Speech Affects Dehumanization and Anthropomorphism.," *Journal of Experimental Psychology: General* (145:11), pp. 1427–1437. (doi: 10.1037/xge0000214).

Schroeder, J., Kardas, M., and Epley, N. 2017. "The Humanizing Voice: Speech Reveals, and Text Conceals, a More Thoughtful Mind in the Midst of Disagreement," *Psychological Science* (28:12), pp. 1745–1762.

Seeger, A.-M., Pfeiffer, J., and Heinzl, A. 2017. "When Do We Need a Human? Anthropomorphic Design and Trustworthiness of Conversational Agents," in *Proceedings of the Sixteenth Annual Pre-ICIS Workshop on HCI Research in MIS, AISeL, Seoul, Korea* (Vol. 10).

Stenzel, A., Chinellato, E., Bou, M. A. T., del Pobil, Á. P., Lappe, M., and Liepelt, R. 2012. "When Humanoid Robots Become Human-like Interaction Partners: Corepresentation of Robotic Actions.," *Journal of Experimental Psychology: Human Perception and Performance* (38:5), pp. 1073–1077. (doi: 10.1037/a0029493).

Tononi, G., and Koch, C. 2015. "Consciousness: Here, There and Everywhere?," *Philosophical Transactions of the Royal Society B: Biological Sciences* (370:1668), pp. 1–18. (doi: 10.1098/rstb.2014.0167).

Touré-Tillery, M., and McGill, A. L. 2015. "Who or What to Believe: Trust and the Differential Persuasiveness of Human and Anthropomorphized Messengers," *Journal of Marketing* (79:4), pp. 94–110.

Vaes, J., Meconi, F., Sessa, P., and Olechowski, M. 2016. "Minimal Humanity Cues Induce Neural Empathic Reactions towards Non-Human Entities," *Neuropsychologia* (89), pp. 132–140.

Valdesolo, P., and Graham, J. 2014. "Awe, Uncertainty, and Agency Detection," *Psychological Science* (25:1), pp. 170–178.

Van den Hende, E. A., and Mugge, R. 2014. "Investigating Gender-schema Congruity Effects on Consumers' Evaluation of Anthropomorphized Products," *Psychology & Marketing* (31:4), pp. 264–277.

Vogeley, K., and Bente, G. 2010. "'Artificial Humans': Psychology and Neuroscience Perspectives on Embodiment and Nonverbal Communication," *Neural Networks* (23:8–9), pp. 1077–1090.

Waytz, A., Cacioppo, J., and Epley, N. 2010. “Who Sees Human? The Stability and Importance of Individual Differences in Anthropomorphism,” *Perspectives on Psychological Science* (5:3), pp. 219–232. (doi: 10.1177/1745691610369336).

Waytz, A., Gray, K., Epley, N., and Wegner, D. M. 2010. “Causes and Consequences of Mind Perception,” *Trends in Cognitive Sciences* (14:8), pp. 383–388.

Waytz, A., Heafner, J., and Epley, N. 2014. “The Mind in the Machine: Anthropomorphism Increases Trust in an Autonomous Vehicle,” *Journal of Experimental Social Psychology* (52), pp. 113–117. (doi: 10.1016/j.jesp.2014.01.005).

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J.-H., and Cacioppo, J. T. 2010. “Making Sense by Making Sentient: Effectance Motivation Increases Anthropomorphism.,” *Journal of Personality and Social Psychology* (99:3), pp. 410–435. (doi: 10.1037/a0020240).

Wiese, E., Metta, G., and Wykowska, A. 2017. “Robots as Intentional Agents: Using Neuroscientific Methods to Make Robots Appear More Social,” *Frontiers in Psychology* (8), p. 1663. (doi: 10.3389/fpsyg.2017.01663).

Yuan, L., Dennis, A., and Potter, R. 2016. “Interacting Like Humans? Understanding the Neurophysiological Processes of Anthropomorphism and Consumer’s Willingness to Pay in Online Auctions,” in *Proceedings of ICIS 2016*.

Złotowski, J., Proudfoot, D., Yogeeswaran, K., and Bartneck, C. 2015. “Anthropomorphism: Opportunities and Challenges in Human–Robot Interaction,” *International Journal of Social Robotics* (7:3), pp. 347–360.

APPENDIX B – Robustness Check Experiment

Technology Description

Amanda

Control Your Devices with Amanda

Amanda is a digital assistant that can control many home devices even if the device is not smart.

The developers of Amanda have created a simulation that allows users to see how it works in real-life.

On the next page, you will use Amanda to control a mower in a simulation.
In this simulation, **the mower works like a real device, is not smart, and is entirely controlled by Amanda**.

Figure B1. Cover Story of Robustness Check Experiment

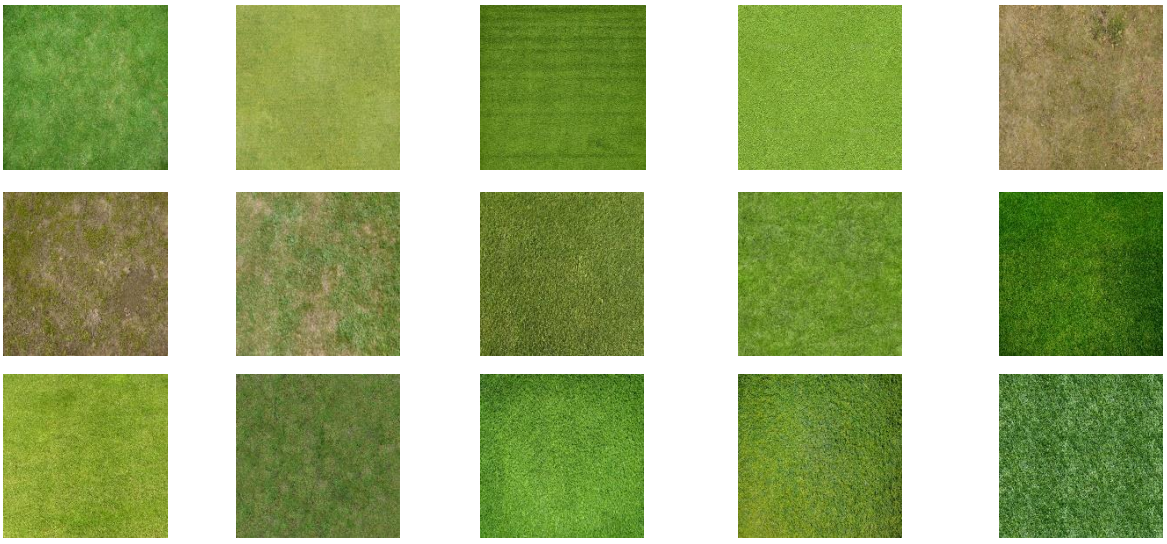


Figure B2. The Grass Textures Used to Generate Various Task Environments

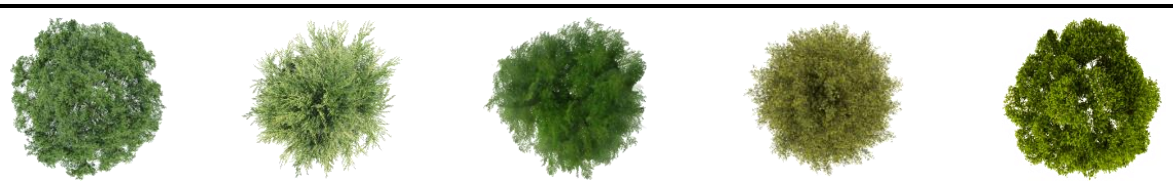


Figure B3. The Tree Images Used to Generate Various Task Environments

Table B1. Exploratory Factor Analysis

	Anthropomorphism	Trust
Anthropomorphism_1	**.842**	.049
Anthropomorphism_2	**.815**	.153
Anthropomorphism_3	**.944**	.064
Anthropomorphism_4	**.907**	.036
Anthropomorphism_5	**.792**	-.089
Trust_1	-.041	**.909**
Trust_2	.109	**.959**
Trust_3	.070	**.982**

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.
Rotation converged in 3 iterations.

Chapter 3:

Creator and Creature: The Role of Inheritability, Trainability, and Freewill in Shaping Distrust in Artificial Intelligence

Abstract

While the investment in AI has dramatically increased over recent years, industry reports indicate that people are not willing to delegate important tasks to AI. With the growing presence of AI agents in our daily lives, it is important to understand why and when people might distrust these agents. While prior research has studied how a person perceives the intentions of another person to shape his or her distrust in the other person, very little is known about how a person perceives the intentions of an AI agent.

In this research, we leverage the metaphor of a human offspring to better model the way users perceive the intentions of AI agents and how this shapes users' distrust toward AI agents. In doing so, we draw on perspectives in cognitive psychology, genetics, and management to define the concepts of AI inheritability (i.e., the extent to which an AI agent is perceived to inherit its values from its creator), AI trainability (i.e., the extent to which an AI agent is perceived to be able to learn how to behave in the manner desired by its user), and AI freewill (i.e., the extent to which an AI agent is perceived to be able to make autonomous choices based on its self-determined objectives and values).

We conduct a randomized experiment and show that users form their distrust in AI agents based on factors such as their distrust in the creator of the agent, AI inheritability, AI trainability, and AI freewill. We also find that the effect of the user's distrust in the creator of the agent can be mitigated by making changes in the other three factors. Our results also confirm that users' distrust in an AI agent influences the delegation of critical tasks more than the delegation of noncritical tasks. We discuss the theoretical and practical implications of our findings.

Keywords: Artificially Intelligent Agents, Distrust Transference, Malevolence, Freewill, AI Inheritability, AI Trainability

INTRODUCTION

The exponential advancement of machine learning in the past decade has enabled ordinary technology users to interact with artificially intelligent (AI) agents regularly (AI Index 2019). Industry reports have projected an AI market size of $8.3 trillion in the US, $2.1 trillion in Japan, and $1.1 trillion in Germany by 2035 (Accenture 2017). It is projected that companies will increase their investments in AI up to three-fold by 2020 (Forrester Research 2017). However, many experts are concerned that users might not be willing to delegate important tasks to AI agents. In fact, a recent survey in the United States indicates that 76% of people would not apply for a job that uses AI to select applicants and 59% would not use an AI agent caregiver (Pew Research Center 2017).

The layperson view of AI agents often involves a malevolent agent striving to take over humanity (Berlatsky 2018). Movies and TV shows (e.g., Westworld, Terminator, Ex Machina, 2001: A Space Odyssey, etc.), as important contributors to people's perception of AI agents (Broadbent et al. 2010), exacerbated distrust in AI agents by illustrating that such agents could attain sentience and murder humans. While people's perception might not necessarily be based on concrete facts (e.g., some argued that AI agents would not be smart enough to threaten humans in the foreseeable future (Smith 2018)), many scholars claimed that fear of AI agents exists (Szollosy 2017). Some people believe that the creators of AI agents might use them to harm others in different ways (e.g., to replace humans with AI agents in various social roles). For instance, Elon Musk, the co-founder of OpenAI, believes that "if one company or small group of people manages to develop god-like super-intelligence, they could take over the world" (Nolan 2018).

Situations in which people might perceive AI agents as malevolent go beyond futuristic agents. For instance, when a user asks a conversational AI agent, such as Amazon Alexa or Apple Siri, to suggest the best restaurant in the area, the agent can give recommendations based on the best interests of its user or its creator. As previously shown in the context of recommendation agents (Komiak and Benbasat 2008; Wang et al. 2018), the possibility of such a divided loyalty of the agent might lead to increased distrust in AI agents.

Users' distrust in the context of AI agents is not limited to the discussed examples. The role of distrust in AI agents is likely to become more substantial with the growing presence of AI agents in different aspects of people's daily lives. When users doubt the intentions of an AI agent but find themselves vulnerable to its actions, they might distrust the agent. Distrust is often regarded as a defensive mechanism to protect oneself against possible harmful actions of the other party (McKnight et al. 2004; Yang et al. 2015).

Prior research on distrust has identified perceived intentions of the trustee (i.e., the entity to be trusted or distrusted) to be central in shaping distrusting beliefs (Dimoka 2010; McKnight and Chervany 2001) and consequently perceived usefulness (McKnight and Choudhury 2006) and price premiums paid for an IT product (Dimoka 2010). However, the extant literature has been silent on how people perceive AI's intentions. Most of the existing research has focused on intention in either human-human interactions, in which the trustee is perceived to have volition (i.e., will), or human-technology interaction, in which the technology is regarded as a tool and assumed to "lack volition and moral agency" (McKnight et al. 2011, p. 5). However, some of the underlying assumptions of the extant literature break down in the context of AI agents, bringing into question its applicability in this context (Schuetz and Venkatesh 2020).

First, distrust could be formed based on the users' perception of not only an agent itself but also the entity who is responsible for the observed behavior of the agent. Neuroscientists have shown that the brain actively strives to find the underlying reasons why a behavior occurs (Clark 2013). Individuals attribute a behavioral outcome to the moral agent (Feltz and Cova 2014), i.e., the main driver of the action, to enhance their predictions of the agent's future behavior (Clark 2013). While AI agents might display a behavior, users do not necessarily attribute the behavior to the agent itself. Therefore, the expectation about negative behavior, i.e., distrust, can be based on the driver of the behavior not only the medium that delivers it. For instance, if a security hole in Amazon Alexa's cloud leads to a data breach and customers' personal information is compromised, users will probably feel betrayed by Amazon, not only Alexa.

Second, an explicit assumption of trust in technological artifacts in the IS literature is that "technology lacks volition and moral agency" (McKnight et al. 2011, p. 5). This assumption implicitly means that, in user-artifact interaction, the user is the only entity with volition. When the object upon which the user is dependent is a tool-like artifact controlled by the user (e.g., Microsoft Excel), the intention of the artifact essentially reduces down to the intention of the user who is using it. However, when the artifact can inherit intentions of other agents, such as its creator, there is a discernible "will" in the artifact's behavior that potentially helps shape users' distrust in the artifact.

Third, users might not view an artifact completely independent of its creator. Prior research indicated that in many conditions a person might perceive a group of entities as a single entity with homogenous characteristics (i.e., entitativity) or transfer her or his trust from a known entity to an unknown entity associated with it (i.e., trust transference) (Stewart 2003, 2006; Wingreen et al. 2019).

Finally, a dichotomous approach to volition, based on which an entity either has complete volition or has no volition, ignores the possibility of a spectrum between pure objects and pure autonomous beings (humans). In the context of AI agents, the artifacts move from being mere objects to become independent creatures, but they are neither traditional objects nor humans. Unlike traditional artifacts, AI artifacts demonstrate characteristics that have traditionally been reserved for humans (Brynjolfsson and McAfee 2014). The advancement of machine learning has enabled the development of trainable, autonomous agents. Such agents could potentially learn from their interactions with their surroundings and act in ways that were not directly designed by their creators or meant by their users.

We believe that it is plausible that users base their distrust on the properties of the entity that is driving the behavior of AI agents. In other words, we postulate that distrust is related to the moral agent responsible for the artifact's behavior (Feltz and Cova 2014). A parsimonious set of responsible agents in the context of human-AI interaction includes the creator (i.e., the entity that has created the AI agent), the creature (i.e., the AI agent), and the user (i.e., the person who interacts with the AI agent).

Accordingly, we argue that three main perceptions about AI characteristics sculpt user's perception of the moral agent that dictates the AI agent's perceived harmful behavior: (1) AI inheritability, which we define as the extent to which an AI agent is perceived to inherit its values from its *creator*, (2) AI trainability, which we define as the extent to which an AI agent is perceived to be able to learn how to behave in the manner desired by its *user*, and (3) AI freewill, which we define as the extent to which an AI agent is perceived to be able to make autonomous choices based on its *self-determined* objectives and values. We speculate that AI inheritability indicates how much of the AI's perceived malevolence is shaped based on its creator's perceived malevolence, AI

trainability influences distrust by creating a better value congruence between the AI agent and the user, and AI freewill indicates that users perceive the AI agent to behave based on its own choices.

To the best of our knowledge, no previous research has studied the relationship between a user's distrust in the creator of an AI agent and his or her distrust in the agent. Moreover, little is known about the plausible role of AI characteristics in mitigating the effect of distrust in creator on distrust in AI agent. Motivated by the discussed practical importance of the problem of distrust in the context of AI agents and the shortcomings of the extant literature in addressing such a problem, we formulate the following research questions:

RQ1: What is the relationship between distrust in the creator of an AI agent and distrust in the AI agent?

RQ2: What are the moderating effects of perceived AI characteristics on this relationship?

We conducted a 2×2×2×2 between-subject randomized experiment to answer our research questions and assess the soundness of our conjectures. We recruited 489 participants from Amazon Mechanical Turk. We employed a scenario-based design in which we randomly assigned participants to each of the 16 treatment conditions. We investigated participants' distrust in a fictitious AI agent that can be used to fulfill many daily business tasks.

BACKGROUND AND RESEARCH MODEL

The overarching idea of this paper is that distrust in an AI artifact could be an extension of distrust in its creator, but the relationship between distrust in creator and distrust in AI agent is conditional on the user's perception of AI inheritability, AI trainability, and AI freewill. Using the analogy of a human offspring, we theorize that the behavior of an AI agent can be inherited from its parent (creator), learned through upbringing (training), and based on its own freewill. In the case of a

human child, people might believe that the child's behavior is genetically inherited from the parents, can be changed through training in family, school, and society, and is based on the child's own choices in life (freewill).

Below, we discuss and leverage relevant literature to develop our hypotheses.

Distrust

In line with prior research, we define *distrust* as confident negative expectations regarding another's conduct[1] (Komiak and Benbasat 2008; Lewicki et al. 1998; Ou and Sia 2010; Yang et al. 2015). The negative expectations are related to fear, worry, concern, cynicism, paranoid feelings, suspicion, doubt, wariness, panic, anger, and attribution of sinister intentions (Deutsch 1958; Kramer 1994; Lewicki et al. 1998; McKnight et al. 2004; McKnight and Chervany 2001; Moody et al. 2014). Distrust reflects "the emotion-charged human survival instinct" (McKnight and Chervany 2001, p. 884).

The formation of distrust is closely related to value incongruence between the involved parties. While unmet expectations due to incompetency can lead to violations of trust, perceptions about value incongruency can lead to distrust (Singh and Sirdeshmukh 2000; Sitkin and Roth 1993). In an interaction between a person and an agent with incongruent values, the person might be fearful of the actions of the agent because the agent operates based on values that are not in the person's best interest (Hsiao 2003). As distrust is a mechanism to protect oneself from the harmful conduct

1 While the focus of this paper is not on the similarities and difference between trust and distrust, we recognize that trust and distrust can co-exist (Dimoka 2010; Lewicki et al. 1998; Lyons et al. 2011; Singh and Sirdeshmukh 2000). As relationships are multifaceted and shaped based on many different negative and positive experiences and interactions, people can simultaneously trust and distrust the same agent (e.g., artifact, person, company, etc.) (Lewicki et al. 1998). Also, we acknowledge that a person cannot simultaneously trust and distrust an agent in a specific task (McKnight and Chervany 2001; Schoorman et al. 2007). For instance, a person would not trust Amazon Alexa in finding the best restaurant in a given area and distrust it in the exact same task.

of the other party (Yang et al. 2015), value incongruency plays an important role in shaping distrust.

There is little consensus in the literature on the components of distrust (Dimoka 2010; Lyons et al. 2011). Some prior research used a subset of negated trust components (for a list of components of trustworthiness see McKnight et al. (2002), p. 338). For instance, McKnight and Choudhury (2001) used distrusting beliefs in competence, benevolence, integrity, and predictability, Dimoka (2010) used discredibility and malevolence, and Moody et al. (2014) included malevolence and incompetence in their conceptualization of distrust.

In this research, drawing from cognitive psychology research (Fiske et al. 2007), we use two components for distrust: malevolence and competence. First, we define *malevolence* as the perceived intention of the trustee to cause harm. This definition embraces the widely recognized negative valence of distrust (Dimoka 2010; Lyons et al. 2011; McKnight and Chervany 2001; Yang et al. 2015). Most prior definitions of malevolence failed to capture the intense negativity associated with it. For instance, Dimoka's (2010) definition only captures concerns about the trustee's commitment to one's welfare and McKnight and Chervany's (2001) definition only includes the trustee's lack of care and motivation to act in one's interest. Second, we define *competence* as the perceived ability of the trustee to act on its intentions. While some prior studies regarded incompetency as a component of distrust (Dimoka 2010; Moody et al. 2014), incompetency in fulfilling a task can only lead to violation of trust (Singh and Sirdeshmukh 2000). We argue that only if incompetency is perceived as an intentional act to cause harm, will the trustor distrust the other party. Otherwise, the trustor would merely decrease his or her trust. Therefore,

given the intentions of an agent (good or ill), the competency - not incompetency - of the agent determines "how much" good or harm it can cause if a person relies on it (Fiske et al. 2007).[2]

(Dis)trust Transference

Trust transference is the influence of a trustor's trust in an entity on her or his trust in another entity in the same or a different context (Wang et al. 2013; Wingreen et al. 2019). For instance, trust transfers between companies and their salespeople (Belanche et al. 2014). Transference of trust depends on the trustor's perceived association between the two entities or their contexts.

The strength of trust transference among entities depends on perceived entitativity of them (Stewart 2003), which itself depends on the perceived strength of entities' relationship (Dasgupta et al. 1999; Mullen 1991; Stewart 2003). *Entitativity* refers to the degree to which a collection of entities is perceived to form a cohesive unit (Campbell 1958; Stewart 2003). Entitativity of two entities could vary along a continuum from a single cohesive unit to two completely independent entities. Perceived similarity (i.e., whether entities are internally related or share innate features) and tie among entities (i.e., whether entities have external relationships or share external cues), can influence their perceived entitativity (Stewart 2003).

Transference of perceptions about one entity to another entity is not limited to perceptions about trust. A person's initial impression of one member of a group becomes the basis of her or his perception of other members (Crawford et al. 2002; McConnell et al. 1997) because one entity is representative of the others (Belanche et al. 2014). As such, we argue that distrust in one member of a group becomes the basis of distrust in unknown members. We define *distrust transference* as

[2] Since judgment of competency and incompetency can be itself value-based and subjective, in studies that do not properly define competency, the construct might capture value-based judgments of the trustor about the trustee.

the influence of a trustor's distrust in an entity on her or his distrust in another entity in the same or a different context.

We argue that distrust transference can take place in a triad that includes a trustor (e.g., a user), a trustee (e.g., an AI agent), and a third party who is related to the trustor and trustee (e.g., the creator of the AI agent) (for trust transference in a triad see Wang et al. (2013)). Based on cognitive balance theory, when a person interacts with a dyad (creating a triad), her or his perception of the two entities depends on the relationship between the two entities in the dyad (Stewart 2006). If the relationship is perceived to be positive, the person's perceptions of the two entities should be either both positive or both negative in order to create a cognitive balance. If the relationship is perceived to be negative, the person's perceptions of the two entities should be in opposite directions in order to create a cognitive balance.

In the context of AI agents, we argue that the relationship between an agent and its creator is normally perceived to be positive. A creature could be perceived as an "agent" of its creator because the creator would create the AI agent to advance the creator's intentions. As such, a user is likely to perceive a creator and its AI agent as a dyad with high entitativity. As members of a dyad with high entitativity are perceived to be homogenous in various aspects, a malevolent creator is likely to create a malevolent AI agent. If a user distrusts the creator of an AI agent, he or she is more likely to think that the creator has created the AI agent to harm her or him. Therefore, we contend that users transfer their distrust from a creator to its creature (i.e., AI agent). We hypothesize that:

H1: Distrust in creator increases distrust in AI agent.

AI Inheritability

Previous research in biology and developmental psychology fields has studied inheritability in humans and animals. These fields define inheritability as the probability that an offspring will inherit some specific features from its parent (Anderson and Lustbader 1975; Danchin et al. 2011; Hirschfeld 1995; Uslaner 2008). In the context of AI, we define inheritability as the extent to which an AI agent is perceived to inherit its values from its creator.

We argue that the mere fact that the design of an AI agent allows its creator to embed its own values in the agent increases distrust in the agent. This is because even if the user does not initially distrust the creator, the fact that the AI agent has been built with the capability to directly inherit its values from an agent other than its user raises questions about possible future malicious behaviors. Therefore, we hypothesize that:

H2: AI inheritability increases distrust in AI agent.

As we discussed, the main reason for distrust transference from a creator to an AI agent is that users normally perceive a strong link between the two entities. Consequently, we argue that the strength of distrust transference should depend on the strength of the link that users perceive between the creator and the AI agent.

The development process of an AI agent provides multiple opportunities for the creator to embed their values in the artifact. The creator can explicitly make the agent act in certain ways. For instance, Apple might explicitly make Siri work only with other Apple apps to add reminders in the calendar, or Amazon might make Alexa buy products only through Amazon.com. In both examples, users might perceive a stronger association between the creator and the AI agent.

AI creators can also embed their values in the learning process of an AI agent. As AI is based on machine learning, design choices imposed in the learning algorithms can heavily bias the behavior of the agent (Yapo and Weiss 2018). For instance, in the case of deep learning, choices of the training sample, learning rate, loss function, etc. can influence what the AI learns and how it behaves (for a more comprehensive discussion of deep learning methods see Goodfellow et al. 2016).

Moreover, while many AI agents are offered as mobile apps or software on home assistant devices, many of them are hosted on their creators' servers (e.g., Amazon Alexa is hosted on Amazon servers) (Saffarizadeh et al. 2017). Just as the physical proximity of two entities can affect the extent to which people perceive them to be a part of the same entity (Belanche et al. 2014), we theorize that when an AI agent operates in close connection with its creator's servers users will be more likely to perceive that the agent inherits its values from its creator.

Therefore, we believe that the extent to which AI is perceived to operate based on the values inherited from its creator (i.e., AI inheritability) can strengthen the relationship between distrusting perceptions about the creator and distrusting perceptions about the creature. AI inheritability provides evidence based on which users can assess the association between an AI agent and its creator. If users have concerns that a creator intends to harm them, they are likely to have concerns that an artifact that is created by the malevolent creator might harm them, but the amount that the creator can harm the users through the AI agent depends on the agent's AI inheritability. For instance, if a creator offers an AI agent whose AI is powered by open-source trained machine learning models and the AI is completely hosted on user's devices with no interaction with the creator's servers, users might perceive little or no association between the intentions of the creator and the intentions of the AI agent. Some companies such as Apple have tried to introduce machine

learning models that operate completely on user's devices to address some users' concerns about the misuse of their personal information (Apple 2019). Thus, we hypothesize that:

> *H3: The positive effect of distrust in creator on distrust in AI agent is stronger for an AI agent with high AI inheritability than an AI agent with low AI inheritability.*

AI Trainability

Extant literature in human resource management and psychology has defined trainability for humans as the ability to learn (Gill 1982) or update existing skills (Hashim and Wok 2014), and the time required to complete training (Gordon et al. 1986). In the context of AI, we define trainability as the extent to which an AI agent is perceived to be able to learn how to behave in the manner desired by its user.

AI agents could be trainable even after they are adopted by users. Some AI agents can learn from their users by collecting training data through the interaction and going through an offline retraining process to update their behavior (Venkatesan and Er 2016). Other agents can learn more directly by providing the means for the users to teach them their preferences. For instance, the agents that learn based on deep reinforcement learning algorithms can learn based on the rewards and punishments users give them (Mnih et al. 2015). Users can provide positive feedback for some behavior and negative feedback for others in order to teach the AI agent to behave in the desired way.

Such teaching mechanisms enable users to assert hard and soft controls on the behavior of AI agents. Hard control mechanisms include behavior control (i.e., controlling of the transformation process of work) and outcome control (i.e., tying incentives to the outcome of a process) (Ouchi

1979; Snell 1992). Soft control mechanisms influence the agent's behavior by creating shared goals, values, and norms (Das and Teng 1998).

The ability to train an AI agent enables users to control the agent's behavior in the short-run and embed their own values into the agent in the long-run. An agent that works based on the user's values has more value congruency with the user. Therefore, from a user's perspective, it is less likely for a trainable AI agent to cause harm to the user.

Also, users might view a trainable AI less negatively as they have the opportunity to understand and influence its behavior through a bidirectional interaction similar to an interpersonal relationship. While users might be fearful of an unknown agent, they can develop a better understanding of the agent when they can train it.

We posit that AI trainability, i.e., the extent to which an AI agent is perceived to be able to learn how to behave in the manner desired by its user, can lead to a perceived potential for value congruency between the user and AI agent, which reduces distrust in the AI agent. Thus, we hypothesize that:

H4: AI trainability decreases distrust in AI agent.

When the user can train the AI agent, the behavior of the AI agent is more likely to be in line with the user's interests. In other words, the user might view the AI agent as an extension of herself or himself (i.e., as an agent that does things on behalf of its user). In such a case, the user perceives a positive association with the AI agent. Based on cognitive balance theory, the relationships of two positively associated members (the user and the AI agent) with a third member (the creator) must be either both positive or both negative. Otherwise, there will be a cognitive imbalance in the user's mind. More specifically, if the user distrusts the creator, i.e., if the user has a negative

relationship with the creator, then he or she must believe that the relationship between the AI agent and its creator is negative as well. Accordingly, the user is less likely to perceive that the behavior of the AI agent is driven by its creator. Therefore, we hypothesize that:

> *H5: The positive effect of distrust in creator on distrust in AI agent is weaker for an AI agent with high AI trainability than an AI agent with low AI trainability.*

AI Freewill

Most people believe that humans have freewill or the capacity to have chosen otherwise (Ebert and Wegner 2011; Feldman et al. 2016; Monroe et al. 2014; Sarkissian et al. 2010). The question of whether freewill exists or it is just an illusion has been the focus of many philosophical studies (Bode et al. 2014).[3] The overwhelming belief in freewill, regardless of its soundness, "suggests that the mind operates in a way that encourages the inference that one's actions are freely chosen" (Ebert and Wegner 2011, p. 966). Researchers have found that the freewill belief is fundamental to our self-concept as human beings (Bode et al. 2014) and to our perception of the humanness of others (Gray et al. 2007).

Some studies have suggested that people psychologically view freewill in terms of the ability to make a choice in line with one's own motives or desires and free of constraints (Monroe et al. 2014; Monroe and Malle 2010). In line with this view, we define *AI freewill* as the extent to which

[3] To better understand the meaning of freewill, we can logically examine two scenarios. First, if we assume that every effect must have a cause, we can find the root causes of any phenomenon by following back a chain of causes and effects. Therefore, any action is an effect of its preceding causes and thus fully determined (Feltz and Cova 2014; Shepherd 2012). Second, if we assume that there is always some part of reality that cannot be explained by causes, then any phenomenon is at least partially indeterministic. In other words, for example, if we had two completely identical universes - with the exact same past and present - where a person wanted to make a purchase decision, it would be possible for her to make different decisions in the two universes as she "willed to do so" (Bode et al. 2014). Some scholars argue that variation in choice under the exact same external and internal circumstances is not conceptually different from a random choice (Searle 2001).

an AI agent is perceived to be able to make autonomous choices based on its self-determined objectives and values.

Prior research showed that regardless of their views of freewill, people perceive probabilistic choice-making as freewill. In fact, when the evidence support indeterminacy in one's behavior, people tend to perceive the behavior as freely chosen (Ebert and Wegner 2011). However, the indeterminacy could stem from such seemingly irrelevant things as pure randomness. Empirical evidence in the context of computerized animated agents suggests that people perceive the agents to have freewill when the agents follow a random sequence of actions instead of a predetermined one (Ebert and Wegner 2011).

Complex deterministic behaviors could also be perceived as indeterministic (Lorenz 1972) because it is hard for the observer to decipher the complicated underlying drivers of the behavior. Therefore, when the action makes sense in the context, it is "possible for wholly determined actions to appear freely chosen" (Ebert and Wegner 2011, p. 970).

These findings suggest that people do often ascribe freewill to non-human agents. In fact, prior research on anthropomorphism and mind perception provides similar evidence suggesting that there is variance in people's perception of AI agents' freewill (Waytz, Cacioppo, et al. 2010; Waytz, Morewedge, et al. 2010) and that people have generalized as well as agent-specific perceptions of freewill (Waytz, Cacioppo, et al. 2010).

An AI agent with freewill can behave based on a set of values or rules that might be unfamiliar or unknown to the users. This unfamiliarity can lead to a sense of anxiety and uncertainty. According to prior research, uncertainty can lead to fear and worry (Carleton et al. 2012).

Moreover, neuroscientists have found that human brain (alongside with the brains of many other animals) more readily associate fearful events to the outgroup members compared to ingroup members (Olsson et al. 2005). In line with this finding, we argue that since a user of an AI agent is more likely to perceive an AI agent as a member of a group of non-humans than a member of a group of humans, he or she is more likely to readily associate an AI agent to fearful events.

Additionally, people often have a fear and anxiety of becoming too rational, cold, mechanical, soulless, and emotionless due to the pressures of daily life. Based on Freud's notion of psychological projection, people often fantasize different things to defend themselves against anxiety or things that they know unconsciously but do not want to experience consciously (Szollosy 2017). In doing so, they defensively project the bad parts of themselves onto others. For instance, in many cases of racism and nationalism, people might say "it was not we who are violent, it was them" (Szollosy 2017, p. 436). Some scholars believe that fear of AI agents is also a form of projection, in which people project their fears onto the agents and create monsters out of them (Szollosy 2017). Therefore, we hypothesize that:

H6: AI freewill increases distrust in AI agent.

People typically believe that an agent with freewill is a moral agent and therefore responsible for its own actions (Feltz and Cova 2014; Floridi and Sanders 2004; Gray et al. 2012). As such, the intentions of a free willing AI agent are the driver of its behavior. Thus, we postulate that when users perceive high freewill in an AI agent, they see the agent as an independent entity – not a part of a creator-creature dyad with high entitativity. In this case, the AI agent can possess values and intentions that are independent of those of its creator. Therefore, negative perceptions about the intentions of the creator are less likely to be transferred to the agent. We hypothesize that:

H7: The positive effect of distrust in creator on distrust in AI agent is weaker for an AI agent with high freewill than an AI agent with low freewill.

Figure 1 shows our research model. The paths in the model can be mapped to our hypotheses.

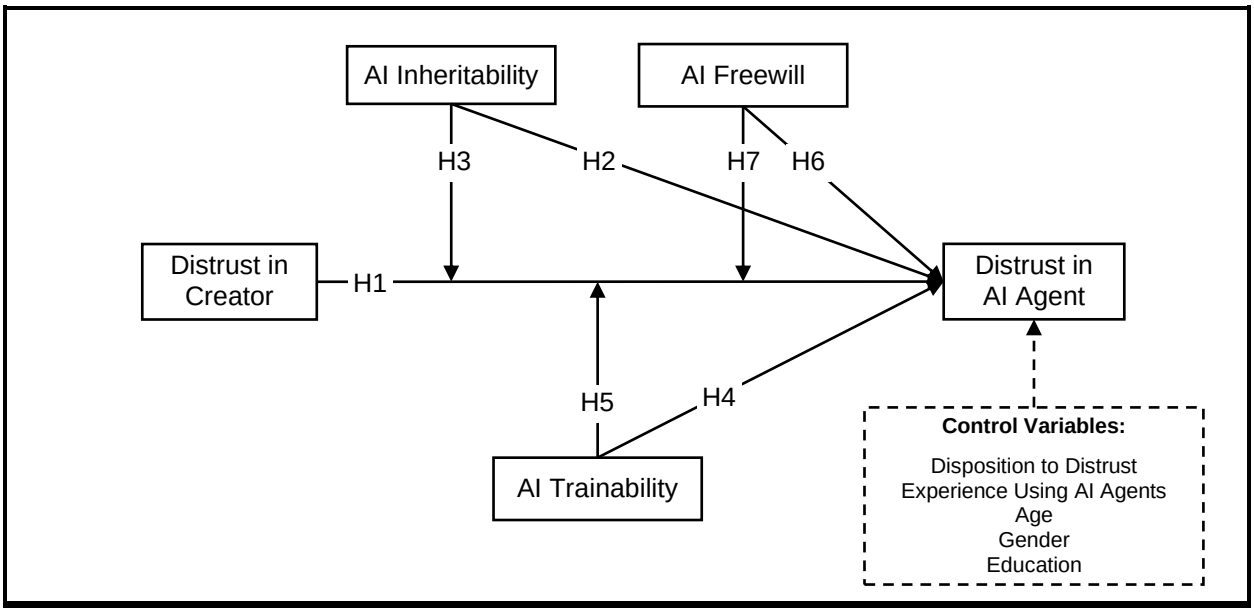


Figure 1. Research Model

METHODOLOGY

Experiment

We conducted a randomized experiment to test our hypotheses. The experiment involved a 2 (distrust in creator: low vs. high) × 2 (AI inheritability: low vs. high) × 2 (AI trainability: low vs. high) × 2 (AI freewill: low vs. high) between-subjects factorial design. Participants were randomly assigned to one of the sixteen experimental conditions.

We manipulated distrust in creator, AI inheritability, AI trainability, and AI freewill independently. Furthermore, to ensure the proper precedence of variables based on our research

model, we delivered the manipulation of distrust in creator before the manipulations of AI inheritability, AI trainability, and AI freewill.

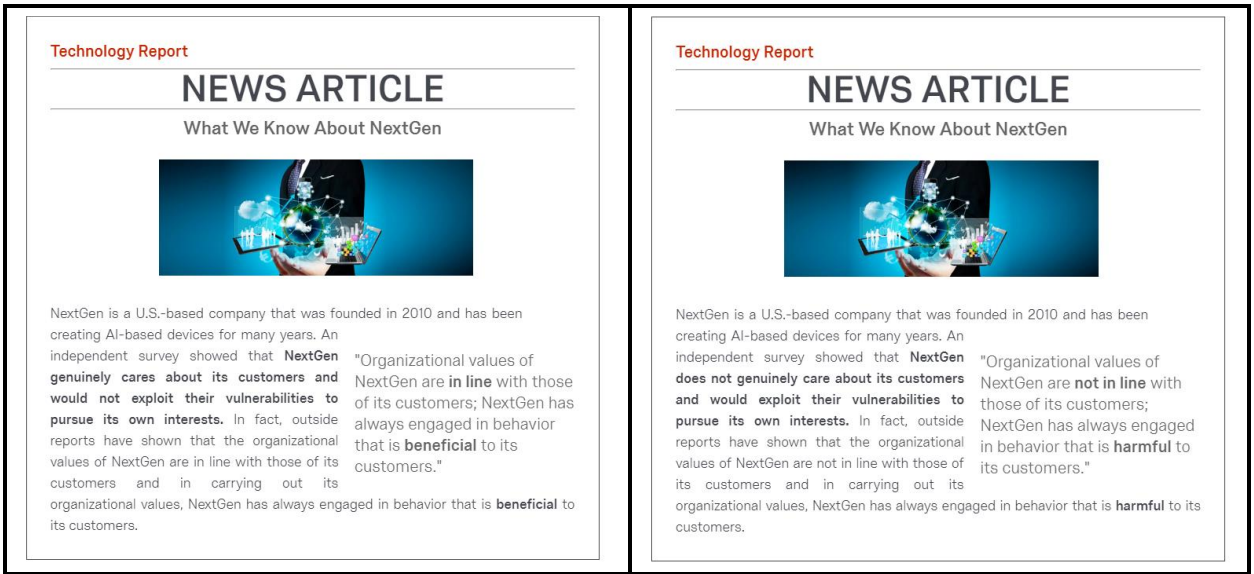
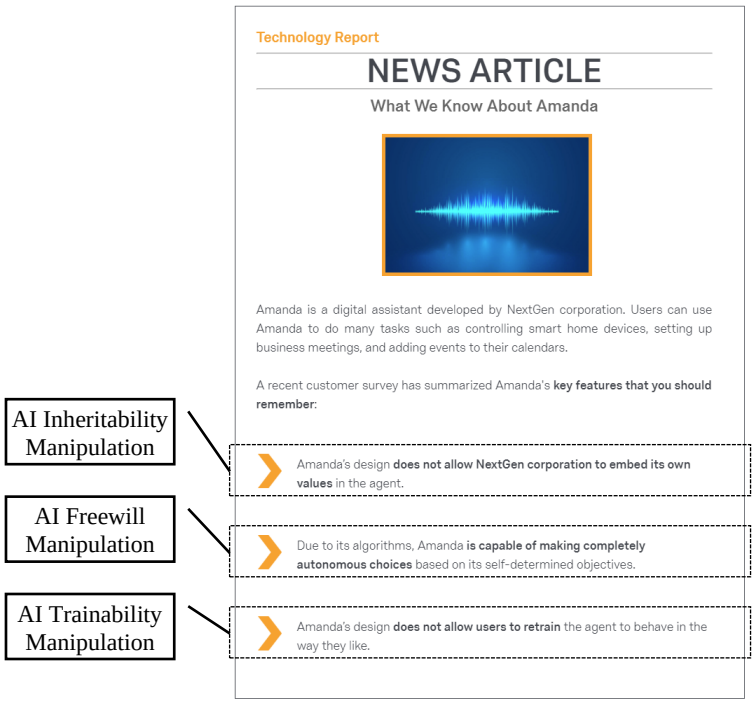


Figure 2. Vignette for Manipulating Distrust in Creator

Distrust in creator was manipulated by asking the participants to read a description that induces either low distrust or high distrust. Since distrust is related to fear, worry, concern, cynicism, paranoid feelings, suspicion, doubt, wariness, panic, anger, and attribution of sinister intentions (Deutsch 1958; Kramer 1994; Lewicki et al. 1998; McKnight et al. 2004; McKnight and Chervany 2001; Moody et al. 2014), in each condition, we included sentences that could increase or decrease such feelings. For instance, in the high distrust condition, the creator was described as a company that has harmed its customers in the past. Figure 2 shows the two vignettes used to manipulate distrust in creator.

AI inheritability was manipulated by describing the AI agent as an agent whose design allows (for high AI inheritability treatment) or does not allow (for low AI inheritability treatment) its creator to embed its own values in the agent. AI trainability was manipulated by telling the participants that the agent's design allows (for high AI trainability treatment) or does not allow (for low AI

trainability treatment) users to retrain the agent to behave in the way they like. Finally, AI freewill was manipulated by stating that the AI agent's algorithms allow (for high AI freewill treatment) or do not allow (for low AI freewill treatment) the agent to make autonomous choices based on its self-determined objectives. Figure 3 provides a summary of these three manipulations.



Construct	Level	Manipulation Content
AI Inheritability	Low	Amanda's design does not allow NextGen to embed its own values in the agent.
	High	Amanda's design allows NextGen to embed its own values in the agent.
AI Freewill	Low	Due to its algorithms, Amanda is not capable of making autonomous choices based on its self-determined objectives.
	High	Due to its algorithms, Amanda is capable of making completely autonomous choices based on its self-determined objectives.
AI Trainability	Low	Amanda's design does not allow users to retrain the agent to behave in the way they like.
	High	Amanda's design allows users to completely retrain the agent to behave in the way they like.

Figure 3. Vignette for Manipulating Inheritability, Trainability, and Freewill

We conducted four pilot studies with a total of 512 participants to develop the experimental instruments for our study. In the pilot studies, we focused on the length, content, and delivery of the vignettes used to manipulate distrust in creator, inheritability, trainability, and freewill. For instance, based on the results of pilot 1, we decided to use an advanced text-to-speech engine to read the content of the vignettes to the participants to intensify the manipulation. In doing so, we developed a JavaScript text-to-speech tool to leverage the Amazon Polly text-to-speech engine to read the content to the participants, given the experimental condition. We employed Polly's neural engine, which can generate extremely human-like voices and mimic a human newscaster style of content delivery (when the speaking style is set to newscaster in the engine).

Participants

We recruited a total of 600 participants to guarantee about 30 participants per experiment condition, with an assumption that about 20% of the participants would fail the attention check questions.[4] We recruited the participants from Amazon's Mechanical Turk. To ensure high-quality responses, we limited the participants to those with either more than 99% acceptance rate or master status and more than 97% acceptance rate, and MTurk experience of more than 500 HITs (Human Intelligence Tasks, which are the tasks posted on MTurk marketplace).

[4] We have four factors, each with two levels. Therefore, we need four degrees of freedom to calculate their main effects (4 × (2-1)). We also have three two-way interactions. Therefore, we need three degrees of freedom to calculate the interaction effects (3 × (2-1) × (2-1)). We estimated the number participants needed for our study using G*Power 3.1.9.2. For a medium effect size (f=0.25), α = 0.05, power = 0.80, numerator df = 7, and number of groups = 16, we need 237 participants. However, to capture all the possible interactions in the model, we need a degree of freedom of 15 (4 main effects, 6 two-way interactions, 4 three-way interactions, 1 four-way interaction, with all factors having two levels). Using the same criteria as before but only with numerator df = 15, we estimated that we need 314 participants. Since, the medium effect size is not guaranteed, we chose to recruit 30 participants per group (a total of 480 participants). Note that effect size in G*Power is calculated as follows:

$f = \frac{\sqrt{\frac{\sum_{j=1}^{k} n_j.(\mu_j - \bar{\mu})^2}{N}}}{\sigma}$,

where n_j denotes the number of participants, μ_j the population mean of group j, $\bar{\mu} = (\sum_{j=1}^{k} n_j.\mu_j)/N$ the weighted mean of the k population means, N the total sample size, and σ the population standard deviation in each group.

Recent studies found that MTurk samples are similar to those derived from national samples, supporting the generalizability of the results from MTurkers (Coppock 2019). They also observed that a large majority of U.S. MTurkers are new to the platform every year and therefore are less likely to be too familiar with manipulations and measures (Robinson et al. 2019). Moreover, they discovered that adding attention-check questions in MTurk surveys can make the quality of the data comparable to that of student subjects (Aruguete et al. 2019) while providing a much more diverse sample (Chandler et al. 2019).

From 600 hundred recruited participants, 489 passed the attention check questions (206 female, 281 male, and 2 other, with an average age of 38.8, ranging from 20 to 78, the median education of 4-year college degree, and the median use of digital assistants of at least once a week).

Participants, on average, spent 6.7 minutes (ranging from 3.2 to 23.3 minutes) to finish the study, and all participants received $1.00 compensation.

Procedure

We asked the participants to read a description of NextGen, which is a fictitious company that creates conversational assistants. Next, we asked them to fill out a survey about the company (distrust in creator). In the next section, we asked them to read a description of Amanda, which is a conversational assistant designed by NextGen. Then we asked the participants to fill out a survey about the assistant. We concluded the survey by asking demographic questions and debriefed the participants by explaining that the company and the agent are fictitious and are not based on any real entities.

Measures

To measure distrust in AI agent, we adopted existing measures of distrust with minimal changes that reflect the context of our study. More specifically, using a 7-point Likert scale, we measured eight items of distrust used by Wang et al. (2018) in the context of recommendation systems. In line with Wang et al. (2018), we constructed distrust in AI agent as the average of the measured items.

To assess the manipulation of distrust in creator, we used similar items to those used for distrust in AI. The manipulation check questions for distrust in creator were placed after the description of the creator but before the description of the AI agent. We checked the manipulation of AI freewill using items from Ebert and Wegner (2011) with minimal changes that reflect the context of our study. Since there was no existing measures of AI inheritability and AI trainability in the literature, for each construct, we created three new measurement items that reflect the definition of the construct.

We also measured age, gender, education, experience using AI agents (use frequency), and disposition to distrust as control variables. We used a 7-point Likert scale to measure six items of disposition to distrust proposed by McKnight et al. (2004) (see Appendix A).

Table 1 includes a summary of the definition of constructs, measurement items, and manipulation check questions. Appendix B presents information regarding construct validity.

Table 1. Constructs			
Construct	**Definition**	**Items**	**Informing Sources**
Distrust in AI	Confident negative expectations regarding AI agent's conduct	1. Amanda would exploit users' vulnerability given the chance. 2. Amanda would engage in harmful behavior to users to pursue its own interest. 3. Amanda would operate in an irresponsible manner. 4. Amanda would interact with users in a deceptive way. 5. Amanda is capable of engaging in harmful behavior toward users. 6. Amanda has the ability to maliciously manipulate users. 7. Amanda is capable of deceiving users. 8. I suspect Amanda is interested in just its own well-being, not mine.	Komiak et al. 2008; Lewicki et al. 1998; Ou and Sia 2010; Yang et al. 2015; Wang et al. 2018
Distrust in Creator	Confident negative expectations regarding AI creator's conduct	This construct was manipulated. Manipulation check questions: 1. NextGen would exploit users' vulnerability given the chance. 2. NextGen would engage in harmful behavior to users to pursue its own interest. 3. NextGen would operate in an irresponsible manner. 4. NextGen would interact with users in a deceptive way. 5. NextGen is capable of engaging in harmful behavior toward users. 6. NextGen has the ability to maliciously manipulate users. 7. NextGen is capable of deceiving users. 8. I suspect NextGen is interested in just its own well-being, not mine.	Komiak et al. 2008; Lewicki et al. 1998; Ou and Sia 2010; Yang et al. 2015; Wang et al. 2018
AI Inheritability	The extent to which an AI agent is perceived to inherit its values from its creator	This construct was manipulated. Manipulation check questions: 1. NextGen can embed its values into Amanda. 2. Amanda inherits its values from NextGen. 3. Amanda's behavior is based on NextGen's values.	-
AI Trainability	The extent to which an AI agent is perceived to be able to learn how to behave in the manner desired by its user	This construct was manipulated. Manipulation check questions: 1. Users can train Amanda. 2. Amanda is trainable by users. 3. A user can train Amanda to behave the way he or she wants.	-
AI Freewill	The extent to which an AI agent is perceived to be able to make autonomous choices based on its self-determined objectives and values	This construct was manipulated. Manipulation check questions: 1. Amanda can freely choose how to behave. 2. For any action Amanda performs, it could have acted differently if it wanted to. 3. Amanda can consciously decide how to act. 4. Amanda seems to have free will. 5. Amanda would be responsible if its behavior harmed somebody.	Ebert and Wegner 2011

ANALYSIS AND RESULTS

Manipulation Check

We averaged the items for manipulation check questions to create composite scores for distrust in creator ($\alpha = 0.982$, 8 items), inheritability ($\alpha = 0.960$, 3 items), trainability ($\alpha = 0.990$, 3 items), and freewill ($\alpha = 0.950$, 5 items). To assess the effectiveness of our manipulations, we conducted an independent oneway ANOVA for each manipulation. Participants who were randomly assigned to low distrust in creator condition reported lower distrust in creator ($M = 2.758, SD = 1.413$) than those assigned to high distrust in creator condition ($M = 6.275, SD = 0.838$; $F(1,487) = 1110.639, p < 0.001, \eta_p^2 = 0.695$). Participants in low AI inheritability condition reported lower perceived AI inheritability ($M = 2.785, SD = 1.784$) than those in high AI inheritability condition ($M = 5.981, SD = 1.141$; $F(1,487) = 554.000, p < 0.01, \eta_p^2 = 0.532$). Similarly, participants in low AI trainability condition indicated lower perceived AI trainability ($M = 2.037, SD = 1.545$) than those in high AI trainability condition ($M = 5.987, SD = 1.184$; $F(1,487) = 1009.032, p < 0.01, \eta_p^2 = 0.674$). Finally, participants who received the low AI freewill treatment reported lower perceived AI freewill ($M = 2.221, SD = 1.288$) than those who received the high AI freewill treatment ($M = 4.273, SD = 1.718$; $F(1,487) = 225.204, p < 0.01, \eta_p^2 = 0.316$).

Empirical Model

We computed composite scores for distrust in AI ($\alpha = 0.975$; 8 items) and treated it as a continuous variable. We modeled all manipulated constructs as binary variables, with 0

representing low level and 1 representing high level of the construct. Table 2 presents the descriptive statistics.

Table 2. Descriptive Statistics (N=489)

	Mean	Std. Dev.	Minimum	Maximum
Age	38.847	11.614	20	78
Gender	0.58	0.502	0	2
Education	4.10	1.301	1	7
Use frequency	2.70	1.151	1	4
Disposition to Distrust	4.524	1.284	1	7
Distrust in Creator	0.49	0.500	0	1
AI Inheritability	0.49	0.500	0	1
AI Trainability	0.51	0.500	0	1
AI Freewill	0.48	0.500	0	1
Distrust in AI	3.512	1.782	1	7

a. Gender is coded as 0=female, 1=male and 2=other.
b. Education is coded as 1=less than high school, 2=high school, 3=some college, 4=2-year college degree, 5=4-year college degree, 6=master's degree, 7=doctorate degree (including JD, MD).
c. Use frequency is coded as 1=never, 2=at least once a month, 3=at least once a week, 4=at least once a day.
d. Distrust in creator, inheritability, trainability, and freewill are manipulated.

AI Inheritability	**AI Trainability**	**AI Freewill**		**Distrust in Creator**	
				Low	**High**
Low	**Low**	**Low**	Distrust in AI	1.924 (1.069)	3.755 (1.701)
			Sample Size (N)	36	27
		High	Distrust in AI	3.245 (1.453)	4.210 (1.625)
			Sample Size (N)	24	28
	High	**Low**	Distrust in AI	2.598 (1.421)	3.025 (1.546)
			Sample Size (N)	37	35
		High	Distrust in AI	3.220 (1.642)	3.384 (1.506)
			Sample Size (N)	33	27
High	**Low**	**Low**	Distrust in AI	2.397 (1.208)	5.469 (1.337)
			Sample Size (N)	29	32
		High	Distrust in AI	3.600 (1.530)	5.413 (1.680)
			Sample Size (N)	36	30
	High	**Low**	Distrust in AI	2.460 (1.349)	3.741 (1.762)
			Sample Size (N)	28	29
		High	Distrust in AI	2.938 (1.215)	5.027 (1.274)
			Sample Size (N)	26	32

Numbers in parentheses are the standard deviations

We employed multiple regression to test our hypotheses. Since each participant was randomly assigned to an experimental condition, each predictor in our regression model is statistically independent of other observed and unobserved variables. Therefore, our design addresses endogeneity issues related to omitted variables. Moreover, because we measured distrust in AI after the manipulation process, we can safely assume that there is no endogeneity related to reverse causality. Thus, we used the plain ordinary least square (OLS) estimator to estimate the model.

While our design allows for a robust assessment of the interaction of all the predictors in the model, we only include the hypothesized interactions in our analysis. Equation 1 presents our empirical model.

$$
\begin{aligned}
(1) \qquad DistrustInAI = {} & \beta_0 + \beta_1 DistrustInCreator + \beta_2 AIInheritability \\
& + \beta_3 AITrainability + \beta_4 AIFreewill \\
& + \beta_5 DistrustInCreator \times AIInheritability \\
& + \beta_6 DistrustInCreator \times AITrainability \\
& + \beta_7 DistrustInCreator \times AIFreewill + Controls + \varepsilon
\end{aligned}
$$

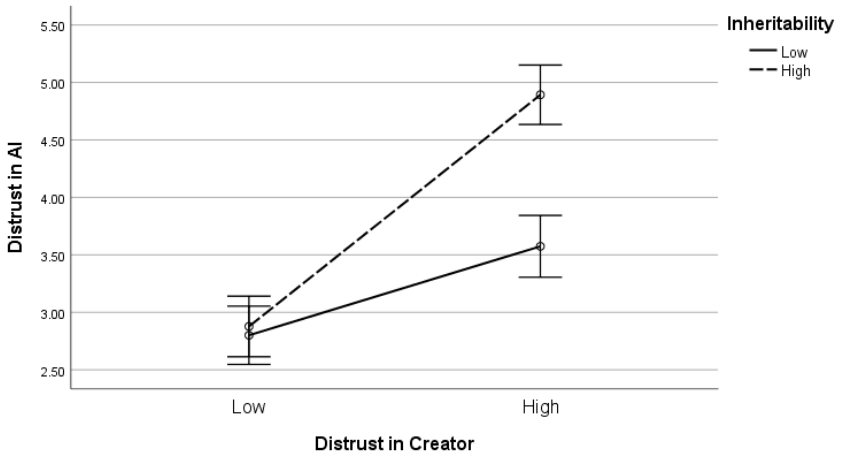
Table 3 shows the results.[5] Hierarchical regression was used to better understand the marginal effect sizes of different blocks of our predictors. In hypothesis 1, we predicted that distrust in creator increases distrust in AI. Model 3 provides support for this hypothesis ($\beta_1 = 1.402; p < 0.01$). Moreover, based on model 2 distrust in creator can explain 15.6% of the variance in distrust in AI, beyond user's disposition to distrust. This finding strongly supports the notion that the users transfer their distrust in creator to their distrust in AI. Hypothesis 2 predicted that perceived AI inheritability increases user's distrust in AI. Model 3 provides support for our hypothesis ($\beta_2 = 0.675; p < 0.01$). In hypothesis 3, we posited that the effect of distrust in creator on distrust in AI is stronger when the user perceives high AI inheritability than when he or she perceives low AI inheritability. Model 4 provides support for this hypothesis by showing a significant positive effect ($\beta_5 = 1.242; p < 0.01$). Note that since the interacting terms are both binary, we did not need to evaluate them at different distances from their means to plot the simple slopes. We, however, did set other variables that were not involved in the interaction to their mean values (Figure 4).

[5] As a robustness check, we estimated the model using all 600 datapoints we collected in the experiment without removing the participants who failed the attention check questions. All results remained the same in terms of direction and significance.

Table 3. Results (N=489)

	Model 1	Model 2	Model 3	Model 4
(Constant)	1.825 (0.542) **	1.460 (0.498) **	1.100 (0.474) *	1.143 (0.469) *
Age	-0.002 (0.007)	-0.002 (0.006)	0.000 (0.006)	-0.001 (0.006)
Gender	0.132 (0.162)	0.136 (0.149)	0.088 (0.140)	0.072 (0.135)
Education	0.065 (0.061)	0.036 (0.056)	0.017 (0.053)	0.011 (0.051)
Use frequency	-0.085 (0.069)	-0.123 (0.063)	-0.110 (0.060)	-0.082 (0.058)
Disposition to Distrust	0.273 (0.063) **	0.201 (0.058) **	0.202 (0.055) **	0.218 (0.053) **
Distrust in Creator (β_1) (H1)		1.424 (0.147) **	1.402 (0.138) **	1.468 (0.264) **
AI Inheritability (β_2) (H2)			0.675 (0.139) **	0.077 (0.187)
AI Trainability (β_3) (H4)			-0.465 (0.138) **	-0.012 (0.187)
AI Freewill (β_4) (H6)			0.687 (0.137) **	0.912 (0.185) **
Distrust in Creator × AI Inheritability (β_5) (H3)				1.242 (0.266) **
Distrust in Creator × AI Trainability (β_6) (H5)				-0.949 (0.265) **
Distrust in Creator × AI Freewill (β_7) (H7)				-0.438 (0.264) *
R^2	4.5%	20.1%	29.8%	35.1%
ΔR^2		15.6%	9.7%	5.3%

* *coefficient is significant at* $p < 0.05$
** *coefficient is significant at* $p < 0.01$
a. One-tailed tests were performed to reflect the directional nature of the hypotheses
b. Numbers in parentheses are the standard errors

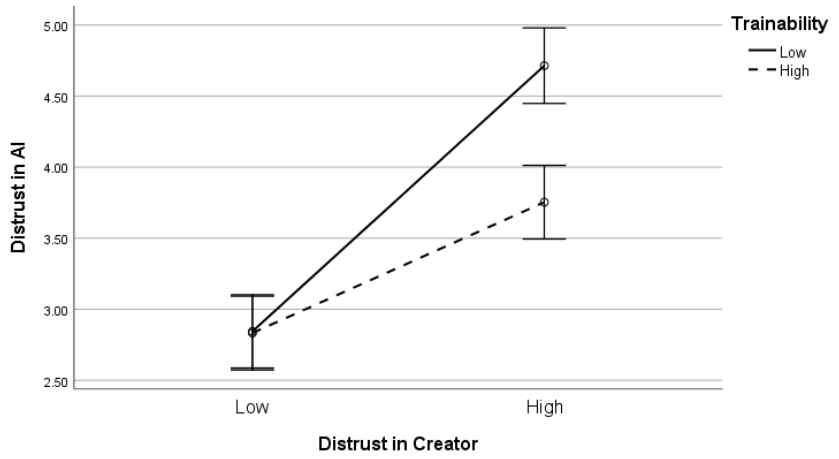


AI trainability, AI freewill and control variables are evaluated at their mean values:
AI Trainability=0.51, AI Freewill=0.48, age=38.847, gender, 0.58, education, 4.10, use frequency=2.7, disposition to distrust=4.524; Error bars represent 95% confidence intervals

Figure 4. Moderating Effect of AI Inheritability on the Relationship between Distrust in Creator and Distrust in AI

In hypothesis 4, we posited that the perceived trainability of AI decreases user's distrust in AI. Model 3 provides support for this hypothesis by showing a significant negative effect ($\beta_3 = -0.465; p < 0.01$).

In hypothesis 5, we predicted that the effect of distrust in creator on distrust in AI is weaker when the user perceives a high AI trainability than when he or she perceives a low AI trainability. This hypothesis is supported based on the results of model 4 ($\beta_6 = 0.949; p < 0.01$). We plotted the simple slopes to demonstrate this interaction (Figure 5).

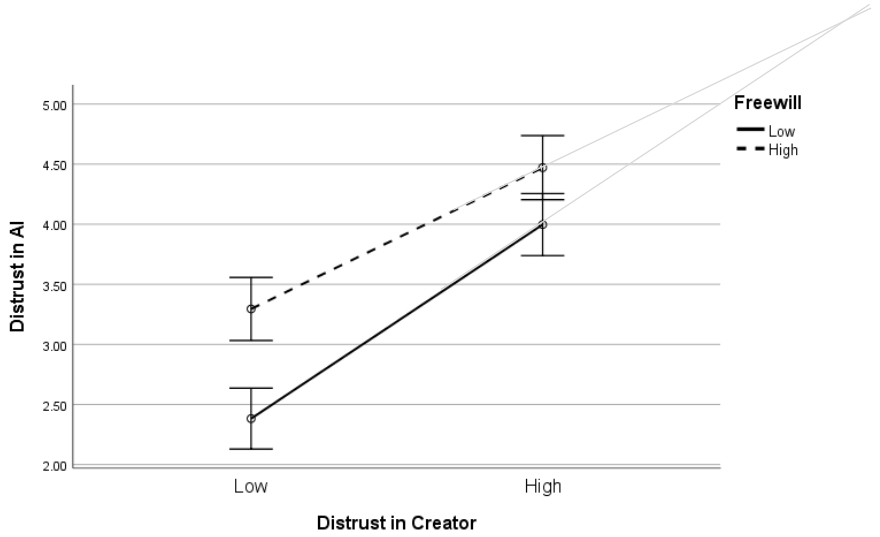


AI inheritability, AI freewill and control variables are evaluated at their mean values:
AI Inheritability=0.49, AI Freewill=0.48, age=38.847, gender, 0.58, education, 4.10, use frequency=2.7, disposition to distrust=4.524; Error bars represent 95% confidence intervals

Figure 5. Moderating Effect of AI Trainability on the Relationship between Distrust in Creator and Distrust in AI

Hypothesis 6 stated that perceived AI freewill increases users' distrust in AI. This hypothesis was also supported by the evidence from Model 3, which indicates a significant positive effect ($\beta_4 = 0.687; p < 0.01$).

In hypothesis 7, we posited that the effect of distrust in creator on distrust in AI is stronger when the user perceives a high AI freewill than when he or she perceives a low AI freewill. Model 4 provides support for this hypothesis by indicating a significant positive effect ($\beta_7 = 0.438; p < 0.05$). Figure 6 depicts this interaction. Please note that since the main effect of AI freewill on distrust in AI (the vertical shift) is much stronger than the interaction (change in the slope), visual comparison of the slopes is difficult.



AI inheritability, AI trainability and control variables are evaluated at their mean values: AI Inheritability=0.49, AI Trainability =0.51, age=38.847, gender, 0.58, education, 4.10, use frequency=2.7, disposition to distrust=4.524; Error bars represent 95% confidence intervals

Figure 6. Moderating Effect of AI Freewill on the Relationship between Distrust in Creator and Distrust in AI

The three AI characteristics, i.e., AI inheritability, AI trainability, and AI freewill, jointly explained an additional 9.7% of the variance in user's distrust in AI through their main effect and an additional 5.3% of the variance through their interactions with distrust in creator (a total of 15.0%). The total explained variance of the model is 35.1%.

In order to establish a clearer link between our findings and the business issues that motivated our research (i.e., users do not delegate critical tasks to AI agents, due to their distrust in the agents), we asked a few follow-up questions from the participants. More specifically, we asked them how likely they were to delegate different tasks to Amanda, the fictitious digital assistant described in the experiment. We chose three critical tasks (i.e., "to monitor your health and alert when you should go to the doctor," "to monitor security cameras in the house," "to schedule an important business meeting with several co-workers") and three noncritical tasks (i.e., "to buy groceries," "to find people who might like to meet for a date," "to pick out and buy a birthday present for an

acquaintance"). We averaged the scores of the questions for each group of tasks to create a single variable for each. Using the lavaan package in R, we simultaneously ran two simple regressions to find the effect of distrust in AI on delegation, and to compare the coefficients in the two regressions. We found that the absolute effect of distrust in AI on delegation is larger for critical tasks ($\zeta_1 = -0.492; p < 0.001$) than noncritical tasks ($\zeta_2 = -0.198; p < 0.001$), and that this difference is statistically significant ($\zeta_1 - \zeta_2 = -0.294; p < 0.001$). Furthermore, we observed that distrust in AI explains 23.6% of the variation in delegation of critical tasks and only 4.9% of the variation in delegation of noncritical tasks. This finding underscores the importance of distrust in understanding the broader issues of delegation of critical tasks to AI agents.

DISCUSSION

The objectives of this research were to first understand the relationship between distrust in creator and distrust in AI and second investigate the moderating effects of AI characteristics on this relationship. We used the metaphor of a human offspring and theorized that the AI characteristics that can shape the perceived behavior of an AI agent are AI inheritability, AI trainability, and AI freewill. We drew on literature in cognitive psychology, organizational behavior, human-computer interaction, and information systems to develop a research model and leveraged the experimental methodology to test our hypotheses. Below we discuss the implications of our findings for research and practice, and the limitations of our study.

Implications for Research

This research has important implications for research on artificial intelligence and distrust.

Expanding our Understanding of Artificial Intelligence

First, we contribute to the emerging body of literature on artificial intelligence by explaining how users perceive an AI agent. While extant literature on AI agents predominantly strives to understand an AI agent within the context of the user-AI dyad, we added the creator of the AI agent into the picture, thus expanding the usual dyadic approach to a triadic approach that includes the user, the AI agent, and the creator of the agent. We demonstrated that understanding this triad is central to our understanding of human-AI interaction. In doing so, we introduced three new constructs, namely AI inheritability, AI trainability, and AI freewill, to parsimoniously conceptualize how an AI agent is perceived with regards to the three major entities that might drive the agent's behavior (i.e., its creator, its user, and itself). While inheritability existed in biology and developmental psychology (nature) (Anderson and Lustbader 1975; Danchin et al. 2011; Hirschfeld 1995; Uslaner 2008), trainability existed in human resource management (nurture) (Gill 1982; Gordon et al. 1986; Hashim and Wok 2014), and freewill existed in cognitive psychology and philosophy (Bode et al. 2014; Ebert and Wegner 2011; Feldman et al. 2016; Monroe et al. 2014; Sarkissian et al. 2010), we appropriated them to fit the context of AI agents.

Second, we contribute to the artificial intelligence literature by providing an explanation of why perceived AI freewill leads to fear and paranoia, which are the underlying drivers of distrust. Our findings are in contrast with previous studies that suggest anthropomorphism, which includes freewill (for more information see Gray et al. 2007; and Waytz, Gray, et al. 2010), increases trust in an AI agent partially because it provides evidence of agency and thus competence of the agent (Waytz et al. 2014). Instead, our results show that perceived AI freewill increases distrust in AI agent. We theorized the reasons for this increase as follows: (1) perceived freewill makes the behavior of the agent be perceived as more uncertain, which increases anxiety, (2) perceived

freewill gives an identity to the agent, but this identity is more likely to be associated with a person's outgroup (non-human agents) than ingroup (human agents), which heightens levels of fear, and (3) people are more likely to project their fears and anxiety on an AI agent with freewill, because they perceive such an agent to be rational and cold. This finding is important for research because it indicates that the relationship between perceived freewill and trust is more complex than currently portrayed in the literature.

Third, we contribute to the artificial intelligence literature by theorizing and showing that users perceive a nontrivial level of volition in AI agents. Our empirical data showed a variance in perceived AI freewill, which suggests that not all people perceive AI agent as pure objects or pure humans. Accordingly, in line with some other studies in the IS field (Schuetz and Venkatesh 2020), we suggest that the use of theories developed for human agents and tool-like objects (e.g., a computer software) might need to be modified in the context of AI agents.

Expanding our Understanding of Distrust

First, we contribute to the trust literature by identifying the antecedents of distrust in the context of AI agents. We theorized and empirically tested the effect of distrust in creator on distrust in AI. We found that users rely on their assessment of the creator on an AI agent to shape their distrust in the AI. Moreover, we found that AI characteristics such as AI inheritability, AI trainability, and AI freewill influence distrust in AI. More specifically, we found that AI inheritability and AI freewill increase, and AI trainability decreases distrust in AI.

Second, we contribute to the trust literature by identifying factors that mitigate the positive effect of distrust in creator on users' distrust in AI agents. We observed that in the context of our study distrust in creator is the most prominent driver of distrust in AI. However, we found that the effect

of distrust in creator on distrust in AI agent is moderated by AI inheritability, AI trainability, and AI freewill. First, AI inheritability, which indicates the extent to which an AI agent is perceived to inherit its values from its creator, positively moderates the effect of distrust in creator on distrust in AI. In other words, when it comes to distrust, users are more likely to mentally separate the creator and the AI agent when the agent has low perceived inheritability. Second, AI trainability negatively moderates the effect of distrust in creator on distrust in AI. This finding indicates that an AI agent is perceived more positively when the AI agent is perceived to be capable of being trained to behave in line with user's values. This necessarily means that if the user has a negative relationship with the creator (high distrust in the creator) but a positive relationship with the AI agent due to its trainability, he or she has no choice but to mentally detach the AI agent from its creator, i.e., to perceive a weak or negative relationship between the AI agent and its creator, in order to create a cognitive balance in the triad (i.e., creator, AI agent, and user). Finally, AI freewill negatively moderates the effect of distrust in creator on distrust in AI. This finding indicates that when the user perceives the AI to behave autonomously based on its self-determined values, he or she is less likely to shape his or her distrust in AI based on distrust in creator.

Third, we contribute to trust literature by developing the concept of distrust transference. We draw from the literature on trust transference and entitativity to conceptualize distrust transference. We believe that distrust transference can help advance our understanding of users' distrust in the context of AI agents, especially in the initial stages of users' interaction with such agents.

Implications for Practice

Practitioners can benefit from the results of this research in several ways. For instance, based on our experiment, negative news about the creator of an AI agent can increase not only users' distrust

in the creator but also users' distrust in the AI agent. To mitigate this distrust transference, developers of AI agents can preemptively design or present the agent with low AI inheritability, high AI trainability, and high AI freewill. Such design choices can help AI agents be perceived as less dependent on their creators and consequently less susceptible to the adverse effects of negative news about their creators. Given the widespread negative news and users' ever-increasing concerns about AI agents, our findings can help developers design AI agents in a way that increases the likelihood of their continued use.

It is important to note that our research suggests that perceived AI freewill can lead to more distrust in AI agent. The distrust in AI agents might have its roots in the popular movies, TV shows, and novels that depict AI agents as malevolent (Broadbent et al. 2010; Szollosy 2017). Therefore, a possible way to mitigate the adverse effect of high perceived AI freewill is to design the agent in a way to avoid triggering any associations with malevolent AI agents that users know. In addition, this negative effect could be possibly reduced by adding anthropomorphic features such as name, gender, voice, and physical embodiment to the agent (Waytz et al. 2014).

Limitations

We conducted an experiment to test our research model. While experiments are often criticized for having low external validity, they are considered the gold standard for establishing internal validity. Since our primary aim was to test causal relationships, an experiment was an appropriate methodology for our research.

A common criticism of experiments in the information systems discipline is that the participants often lack the requisite domain experience to reflect actual users of a system in realistic settings. In our study the median participant reported that they used digital assistants at least once a week,

which makes the participants acceptable representatives of the population of the users of common AI agents. Nevertheless, more research can build on our theoretical foundation and empirical evidence to confirm our findings in field settings.

Finally, we adopted a scenario-based experiment design to manipulate our constructs. Hypothetical scenarios are often criticized as having low ecological validity. To address this shortcoming, we designed the stimuli in our experiment to be as realistic as possible. Specifically, we constructed stimuli so as to mimic the way in which people are exposed to news articles to see how their distrust in AI agents changes due to the exposure. We also used the state-of-the-art neural text-to-speech technology to add a human-like voice with a very realistic newscaster tone to the news articles. Notwithstanding, future research can add to our findings by having the participants interact with an actual AI agent. In such a study, researchers can still manipulate distrust in creator and AI inheritability through sharing descriptions of the creator and the AI agent. Future research, however, can employ such reinforcement learning methods as Q-learning to showcase and manipulate the trainability of the agent and leverage the concept of AI indeterminacy (see Saffarizadeh and Keil 2020) to manipulate AI freewill.

CONCLUSION

While many companies are increasing their investment in AI, reports show that users distrust AI agents. Many users perceive these agents as malevolent and thus are not willing to delegate crucial tasks to them. In this research, we explained why users' distrust in creator transfers to their distrust in an AI agent and how factors such as AI inheritability, AI trainability, and AI freewill can mitigate this transference. We tested our research model using a 2×2×2×2 randomized experiment.

We expect that our findings will open new doors for theory-driven research in the emerging context of AI agents.

REFERENCES

Accenture. 2017. "Accenture Report: Artificial Intelligence Has Potential to Increase Corporate Profitability in 16 Industries by an Average of 38 Percent by 2035 | Accenture Newsroom."

AI Index. 2019. "The AI Index Annual Report," AI Index.

Anderson, T. F., and Lustbader, E. 1975. "Inheritability of Plasmids and Population Dynamics of Cultured Cells," *Proceedings of the National Academy of Sciences* (72:10), pp. 4085–4089.

Apple. 2019. "Privacy - Approach to Privacy - Apple."

Aruguete, M. S., Huynh, H., Browne, B. L., Jurs, B., Flint, E., and McCutcheon, L. E. 2019. "How Serious Is the 'Carelessness' Problem on Mechanical Turk?," *International Journal of Social Research Methodology* (22:5), Taylor & Francis, pp. 441–449.

Belanche, D., Casaló, L. V., Flavián, C., and Schepers, J. 2014. "Trust Transfer in the Continued Usage of Public E-Services," *Information & Management* (51:6), pp. 627–640.

Berlatsky, N. 2018. "Is AI Dangerous? Why Our Fears of Killer Computers or Sentient 'Westworld' Robots Are Overblown."

Bode, S., Murawski, C., Soon, C. S., Bode, P., Stahl, J., and Smith, P. L. 2014. "Demystifying 'Free Will': The Role of Contextual Information and Evidence Accumulation for Predictive Brain Activity," *Neuroscience & Biobehavioral Reviews* (47), pp. 636–645.

Broadbent, E., Kuo, I. H., Lee, Y. I., Rabindran, J., Kerse, N., Stafford, R., and MacDonald, B. A. 2010. "Attitudes and Reactions to a Healthcare Robot," *Telemedicine and E-Health* (16:5), pp. 608–613.

Brynjolfsson, E., and McAfee, A. 2014. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*, WW Norton & Company.

Campbell, D. T. 1958. "Common Fate, Similarity, and Other Indices of the Status of Aggregates of Persons as Social Entities," *Behavioral Science* (3:1), pp. 14–25.

Carleton, R. N., Weeks, J. W., Howell, A. N., Asmundson, G. J., Antony, M. M., and McCabe, R. E. 2012. "Assessing the Latent Structure of the Intolerance of Uncertainty Construct: An Initial Taxometric Analysis," *Journal of Anxiety Disorders* (26:1), Elsevier, pp. 150–157.

Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J., and Litman, L. 2019. "Online Panels in Social Science Research: Expanding Sampling Methods beyond Mechanical Turk," *Behavior Research Methods* (51:5), Springer, pp. 2022–2038.

Clark, A. 2013. "Whatever next? Predictive Brains, Situated Agents, and the Future of Cognitive Science," *Behavioral and Brain Sciences* (36:3), pp. 181–204.

Coppock, A. 2019. "Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach," *Political Science Research and Methods* (7:3), Cambridge University Press, pp. 613–628.

Crawford, M. T., Sherman, S. J., and Hamilton, D. L. 2002. "Perceived Entitativity, Stereotype Formation, and the Interchangeability of Group Members.," *Journal of Personality and Social Psychology* (83:5), p. 1076.

Danchin, É., Charmantier, A., Champagne, F. A., Mesoudi, A., Pujol, B., and Blanchet, S. 2011. "Beyond DNA: Integrating Inclusive Inheritance into an Extended Theory of Evolution," *Nature Reviews Genetics* (12:7), pp. 475–486.

Das, T. K., and Teng, B.-S. 1998. "Between Trust and Control: Developing Confidence in Partner Cooperation in Alliances," *Academy of Management Review* (23:3), pp. 491–512.

Dasgupta, N., Banaji, M. R., and Abelson, R. P. 1999. "Group Entitativity and Group Perception: Associations between Physical Features and Psychological Judgment.," *Journal of Personality and Social Psychology* (77:5), p. 991.

Deutsch, M. 1958. "Trust and Suspicion," *Journal of Conflict Resolution* (2:4), pp. 265–279.

Dimoka, A. 2010. "What Does the Brain Tell Us about Trust and Distrust? Evidence from a Functional Neuroimaging Study," *Mis Quarterly*, pp. 373–396.

Ebert, J. P., and Wegner, D. M. 2011. "Mistaking Randomness for Free Will," *Consciousness and Cognition* (20:3), pp. 965–971.

Feldman, G., Wong, K. F. E., and Baumeister, R. F. 2016. "Bad Is Freer than Good: Positive–Negative Asymmetry in Attributions of Free Will," *Consciousness and Cognition* (42), pp. 26–40.

Feltz, A., and Cova, F. 2014. "Moral Responsibility and Free Will: A Meta-Analysis," *Consciousness and Cognition* (30), pp. 234–246.

Fiske, S. T., Cuddy, A. J., and Glick, P. 2007. "Universal Dimensions of Social Cognition: Warmth and Competence," *Trends in Cognitive Sciences* (11:2), pp. 77–83.

Floridi, L., and Sanders, J. W. 2004. "On the Morality of Artificial Agents," *Minds and Machines* (14:3), pp. 349–379.

Forrester Research. 2017. “Predictions 2017: Artificial Intelligence Will Drive The Insights Revolution,” Forrester.

Gill, R. W. 1982. “A Trainability Concept for Management Potential and an Empirical Study of Its Relationship with Intelligence for Two Managerial Skills,” *Journal of Occupational Psychology* (55:2), pp. 139–147.

Goodfellow, I., Bengio, Y., and Courville, A. 2016. *Deep Learning*, MIT press.

Gordon, M. E., Cofer, J. L., and McCullough, P. M. 1986. “Relationships among Seniority, Past Performance, Interjob Similarity, and Trainability.,” *Journal of Applied Psychology* (71:3), p. 518.

Gray, H. M., Gray, K., and Wegner, D. M. 2007. “Dimensions of Mind Perception,” *Science* (315:5812), pp. 619–619.

Gray, K., Young, L., and Waytz, A. 2012. “Mind Perception Is the Essence of Morality,” *Psychological Inquiry* (23:2), pp. 101–124.

Hashim, J., and Wok, S. 2014. “Competence, Performance and Trainability of Older Workers of Higher Educational Institutions in Malaysia,” *Employee Relations*.

Hirschfeld, L. A. 1995. “The Inheritability of Identity: Children’s Understanding of the Cultural Biology of Race,” *Child Development* (66:5), pp. 1418–1437.

Hsiao, R.-L. 2003. “Technology Fears: Distrust and Cultural Persistence in Electronic Marketplace Adoption,” *The Journal of Strategic Information Systems* (12:3), pp. 169–199.

Komiak, S. Y., and Benbasat, I. 2008. “A Two-Process View of Trust and Distrust Building in Recommendation Agents: A Process-Tracing Study,” *Journal of the Association for Information Systems* (9:12), p. 2.

Kramer, R. M. 1994. “The Sinister Attribution Error: Paranoid Cognition and Collective Distrust in Organizations,” *Motivation and Emotion* (18:2), pp. 199–230.

Lewicki, R. J., McAllister, D. J., and Bies, R. J. 1998. “Trust and Distrust: New Relationships and Realities,” *Academy of Management Review* (23:3), pp. 438–458.

Lorenz, E. 1972. *Predictability: Does the Flap of a Butterfly’s Wing in Brazil Set off a Tornado in Texas?*, presented at the The American Association for the Advancement of Science, Washington, DC.

Lyons, J. B., Stokes, C. K., Eschleman, K. J., Alarcon, G. M., and Barelka, A. J. 2011. “Trustworthiness and IT Suspicion: An Evaluation of the Nomological Network,” *Human Factors* (53:3), pp. 219–229.

McConnell, A. R., Sherman, S. J., and Hamilton, D. L. 1997. "Target Entitativity: Implications for Information Processing about Individual and Group Targets.," *Journal of Personality and Social Psychology* (72:4), p. 750.

McKnight, D. H., Carter, M., Thatcher, J. B., and 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), p. 12.

McKnight, D. H., and Chervany, N. L. 2001. "Trust and Distrust Definitions: One Bite at a Time," in *Trust in Cyber-Societies*, Springer, pp. 27–54.

McKnight, D. H., and Choudhury, V. 2006. "Distrust and Trust in B2C E-Commerce: Do They Differ?," in *Proceedings of the 8th International Conference on Electronic Commerce: The New e-Commerce: Innovations for Conquering Current Barriers, Obstacles and Limitations to Conducting Successful Business on the Internet*, ACM, pp. 482–491.

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

McKnight, D. H., Kacmar, C. J., and Choudhury, V. 2004. "Dispositional Trust and Distrust Distinctions in Predicting High-and Low-Risk Internet Expert Advice Site Perceptions," *E-Service* (3:2), pp. 35–58.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., and Ostrovski, G. 2015. "Human-Level Control through Deep Reinforcement Learning," *Nature* (518:7540), p. 529.

Monroe, A. E., Dillon, K. D., and Malle, B. F. 2014. "Bringing Free Will down to Earth: People's Psychological Concept of Free Will and Its Role in Moral Judgment," *Consciousness and Cognition* (27), pp. 100–108.

Monroe, A. E., and Malle, B. F. 2010. "From Uncaused Will to Conscious Choice: The Need to Study, Not Speculate about People's Folk Concept of Free Will," *Review of Philosophy and Psychology* (1:2), pp. 211–224.

Moody, G. D., Galletta, D. F., and Lowry, P. B. 2014. "When Trust and Distrust Collide Online: The Engenderment and Role of Consumer Ambivalence in Online Consumer Behavior," *Electronic Commerce Research and Applications* (13:4), pp. 266–282.

Mullen, B. 1991. "Group Composition, Salience, and Cognitive Representations: The Phenomenology of Being in a Group," *Journal of Experimental Social Psychology* (27:4), pp. 297–323.

Nolan, J. 2018. *Do You Trust This Computer?*

Olsson, A., Ebert, J. P., Banaji, M. R., and Phelps, E. A. 2005. "The Role of Social Groups in the Persistence of Learned Fear," *Science* (309:5735), American Association for the Advancement of Science, pp. 785–787.

Ou, C. X., and Sia, C. L. 2010. "Consumer Trust and Distrust: An Issue of Website Design," *International Journal of Human-Computer Studies* (68:12), pp. 913–934.

Ouchi, W. G. 1979. "A Conceptual Framework for the Design of Organizational Control Mechanisms," *Management Science* (25:9), pp. 833–848.

Pew Research Center. 2017. "Many Americans Would Be Hesitant to Use Various Automation Technologies," *Pew Research Center: Internet, Science & Tech*.

Robinson, J., Rosenzweig, C., Moss, A. J., and 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), Public Library of Science.

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

Saffarizadeh, K., and Keil, M. 2020. "AI Indeterminacy in Conversational Agents: Investigating Anthropomorphism and Trust," SSRN Scholarly Paper, SSRN Scholarly Paper.

Sarkissian, H., Chatterjee, A., De Brigard, F., Knobe, J., Nichols, S., and Sirker, S. 2010. "Is Belief in Free Will a Cultural Universal?," *Mind & Language* (25:3), pp. 346–358.

Schoorman, F. D., Mayer, R. C., and Davis, J. H. 2007. "An Integrative Model of Organizational Trust: Past, Present, and Future," *Academy of Management Review* (32:2), pp. 344–354.

Schuetz, S., and Venkatesh, V. 2020. "The Rise of Human Machines: How Cognitive Computing Systems Challenge Assumptions of User-System Interaction," *Journal of the Association for Information Systems* (Forthcoming).

Searle, J. R. 2001. "Free Will as a Problem in Neurobiology," *Philosophy* (76:4), pp. 491–514.

Shepherd, J. 2012. "Free Will and Consciousness: Experimental Studies," *Consciousness and Cognition* (21:2), pp. 915–927.

Singh, J., and Sirdeshmukh, D. 2000. "Agency and Trust Mechanisms in Consumer Satisfaction and Loyalty Judgments," *Journal of the Academy of Marketing Science* (28:1), pp. 150–167.

Sitkin, S. B., and Roth, N. L. 1993. "Explaining the Limited Effectiveness of Legalistic 'Remedies' for Trust/Distrust," *Organization Science* (4:3), pp. 367–392.

Smith, G. 2018. *The AI Delusion*, Oxford University Press.

Snell, S. A. 1992. "Control Theory in Strategic Human Resource Management: The Mediating Effect of Administrative Information," *Academy of Management Journal* (35:2), pp. 292–327.

Stewart, K. J. 2003. “Trust Transfer on the World Wide Web,” *Organization Science* (14:1), pp. 5–17.

Stewart, K. J. 2006. “How Hypertext Links Influence Consumer Perceptions to Build and Degrade Trust Online,” *Journal of Management Information Systems* (23:1), pp. 183–210.

Szollosy, M. 2017. “Freud, Frankenstein and Our Fear of Robots: Projection in Our Cultural Perception of Technology,” *AI & SOCIETY* (32:3), Springer, pp. 433–439.

Uslaner, E. M. 2008. “Where You Stand Depends upon Where Your Grandparents Sat: The Inheritability of Generalized Trust,” *Public Opinion Quarterly* (72:4), pp. 725–740.

Venkatesan, R., and Er, M. J. 2016. “A Novel Progressive Learning Technique for Multi-Class Classification,” *Neurocomputing* (207), pp. 310–321.

Wang, N., Shen, X.-L., and Sun, Y. 2013. “Transition of Electronic Word-of-Mouth Services from Web to Mobile Context: A Trust Transfer Perspective,” *Decision Support Systems* (54:3), pp. 1394–1403.

Wang, W., Xu, J., and 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), pp. 5198–5219.

Waytz, A., Cacioppo, J., and Epley, N. 2010. “Who Sees Human? The Stability and Importance of Individual Differences in Anthropomorphism,” *Perspectives on Psychological Science* (5:3), pp. 219–232.

Waytz, A., Gray, K., Epley, N., and Wegner, D. M. 2010. “Causes and Consequences of Mind Perception,” *Trends in Cognitive Sciences* (14:8), pp. 383–388.

Waytz, A., Heafner, J., and Epley, N. 2014. “The Mind in the Machine: Anthropomorphism Increases Trust in an Autonomous Vehicle,” *Journal of Experimental Social Psychology* (52), pp. 113–117.

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J.-H., and Cacioppo, J. T. 2010. “Making Sense by Making Sentient: Effectance Motivation Increases Anthropomorphism.,” *Journal of Personality and Social Psychology* (99:3), pp. 410–435.

Wingreen, S. C., Mazey, N. C., Baglione, S. L., and Storholm, G. R. 2019. “Transfer of Electronic Commerce Trust between Physical and Virtual Environments: Experimental Effects of Structural Assurance and Situational Normality,” *Electronic Commerce Research* (19:2), pp. 339–371.

Yang, J., Sia, C.-L., and Ou, C. X. 2015. “Identify the Antecedents of Distrust in a Website.,” in *PACIS*, p. 108.

Yapo, A., and Weiss, J. 2018. *Ethical Implications of Bias in Machine Learning*.

APPENDIX A – Disposition to Distrust

Questions adopted from McKnight et al. 2004:

1. People are usually out for their own good.
2. People pretend to care more about one another than they really do.
3. Most people inwardly dislike putting themselves out to help other people.
4. Most people would tell a lie if they could gain by it.
5. People don't always hold to the standard of honesty they claim.
6. Most people would cheat on their income tax if they thought they could get away with it.

APPENDIX B – Construct Validation

None of the variables exceeds the 3.0 threshold of skewness and 10.0 threshold of kurtosis. Thus, no variable exhibited significant departure from normality.

Table B1. Skewness and Kurtosis of variable		
Construct	Skewness	Kurtosis
Age	0.785	0.004
Gender	-0.239	-1.699
Education	-0.232	-0.922
Use frequency	-0.324	-1.333
Disposition to Distrust	-0.396	-0.192
Distrust in Creator	0.037	-2.007
Inheritability	0.021	-2.008
Trainability	-0.021	-2.008
Freewill	0.070	-2.003
Distrust in AI	0.285	-0.977

We performed a principal axis factoring to check whether items on distrust in AI and disposition to distrust (the only two multi-indicator constructs in our empirical model) loaded higher on their own construct than on the other construct and had loadings of larger than 0.7 on their own construct.

Table B2. Factor Analysis		
Measurement Items	**Distrust in AI**	**Disposition to Distrust**
Distrust_in_AI_1	**0.902**	0.112
Distrust_in_AI_2	**0.942**	0.103
Distrust_in_AI_3	**0.928**	0.096
Distrust_in_AI_4	**0.939**	0.081
Distrust_in_AI_5	**0.899**	0.085
Distrust_in_AI_6	**0.889**	0.114
Distrust_in_AI_7	**0.887**	0.117
Distrust_in_AI_8	**0.876**	0.050
Disposition_to_Distrust_1	0.032	**0.803**
Disposition_to_Distrust_2	0.107	**0.855**
Disposition_to_Distrust_3	0.089	**0.801**
Disposition_to_Distrust_4	0.074	**0.861**
Disposition_to_Distrust_5	0.111	**0.816**
Disposition_to_Distrust_6	0.093	**0.722**
Extraction Method: Principal Axis Factoring. *Rotation Method: Varimax with Kaiser Normalization.* *Rotation converged in 3 iterations.*		

Chapter 4:

"My Name is Alexa. What's Your Name?" Cognitive and Affective Self-Disclosure Reciprocity in Human-AI Interaction

Abstract

The number of conversational agents (e.g., Amazon's Alexa, Apple Siri, and Google Assistant) has been increasing over the past few years. The functionality of these agents, however, depends on the amount and quality of the data they receive from their users. Evidence from research and practice suggests that conversational agents can elicit data from users through reciprocal self-disclosure. Reciprocal self-disclosure takes place when a party discloses some information about itself with the expectation that the other party will reciprocate by disclosing similar information. While reciprocal self-disclosure seems to work as a data acquisition method, it is not clear how exactly self-disclosure by a conversational agent leads to user self-disclosure, and whether trust is affected in the process. If reciprocal self-disclosure works only because users feel obligated to reciprocate the disclosure based on the social norms, they might feel manipulated and lose trust in the conversational agent.

Leveraging the context of conversational agents, we argue that the extent to which a user attributes humanlike attributes to a conversational agent, i.e., anthropomorphism, plays an important role in reciprocal self-disclosure process. Moreover, we posit that the disclosure provides cognitive and affective bases on which users can develop an interpersonal trust with the conversational agent. We draw on perspectives in communication, psychology, and human-computer interaction to theorize the role of anthropomorphism and trust in the process of reciprocal self-disclosure.

Using a custom-designed conversational agent, we conduct a randomized experiment and show that self-disclosure by a conversational agent acts as an anthropomorphic signal, which provides conceptually distinct cognitive and affective evidence for the user to increase his or her trust in the conversational agent and disclose information. We discuss the theoretical and practical implications of our research.

Keywords: Artificially Intelligent Agent, Cognition-based Trust, Affect-based Trust, Anthropomorphism, Self-Disclosure, Conversational Agent

INTRODUCTION

According to industry reports, the next mainstream computing platform will be voiced-based (VoiceLabs 2017). Typically, voice-based services are facilitated through conversational agents (CAs), such as Amazon's Alexa, Apple's Siri, Microsoft's Cortana, and Google's Assistant. Across different platforms, about 3.25 billion conversational agents were in use at the beginning of 2019 (Voicebot.ai 2019), and it is estimated that by 2023 this number will rise to 8 billion (JuniperResearch 2018). The functionality of conversational agents, however, depends on obtaining data from users. Users' data is often used to improve speech recognition, detect user-specific pronunciations, understand the context of requests, improve the relevance and accuracy of responses, and learn users' preferences at individual and aggregate levels (Apple 2019; Google 2019).

While CA providers, including Amazon, Google, and Apple, allow users to review and delete the data they have shared with CAs, they state that users' data is key in providing a personalized, high-quality experience for the users (Apple 2019; Google 2019). The more information users share with CAs, the greater the value they receive from these devices. However, many users might be reluctant to share personal information with CAs due to their privacy concerns (Saffarizadeh et al. 2017), leaving an open question for practitioners how CAs can obtain users' data without violating users' trust.

Previous research on CAs suggested that an important method to obtain data in a conversational setting is reciprocal self-disclosure (Bickmore and Cassell 2005), in which the CA discloses some information about itself with the expectation that the user will reciprocate by disclosing similar information (Archer and Berg 1978; Moon 2000; Sprecher et al. 2013). For instance, SlugBot and

Fantom, two of the finalist conversational AI agents developed for 2018 Alexa prize leveraged reciprocal self-disclosure to gather information from users (Bowden et al. 2019; Jonell et al. 2018). SlugBot used rules of gradual reciprocal self-disclosure to understand users' interests by asking intimate question from users after revealing similar information about itself (Bowden et al. 2019), while Fantom kept the same level of self-disclosure as the users during the initial phase of conversation and disclosed more information about itself whenever needed during the rest of the conversation (Jonell et al. 2018).

While reciprocal self-disclosure seems to work as a data acquisition method, it is not clear how exactly self-disclosure by a CA leads to user self-disclosure, and whether trust is affected in the process. One possibility is that CA self-disclosure manipulates people into disclosing information because the CA is exploiting a social norm and the user feels compelled to reciprocate. If this is the case, then using CA self-disclosure as a strategy could backfire over repeated interactions because users could react negatively if they feel that they are being manipulated. As a result, users may lose trust and choose to stop using the CA or provide false information to it. Another possibility is that CA self-disclosure actually builds trust, which could be helpful in the long run. Therefore, it is important to open up the black box of reciprocal self-disclosure in the context of conversational agents to understand how trust is influenced in the process.

From a theoretical point of view, prior studies showed that reciprocal self-disclosure exists in a wide range of interactions such as face-to-face conversations (Collins and Miller 1994), conversations in online forums and social media (Barak and Gluck-Ofri 2007; Lin and Utz 2017), and short conversational-like disclosures in human-computer interactions (Moon 2000). These studies, however, adopted different perspectives in explaining how self-disclosure by one party leads to self-disclosure by the other. For example, some scholars suggested that people reciprocate

self-disclosure because such behavior signals the expected behavior based on either social norms or uncertainty in the interaction (Cropanzano and Mitchell 2005; Rubin 1975). Other scholars argued that people reciprocate self-disclosure because they perceive self-disclosure by others as a sign of liking (Sprecher et al. 2013). However, findings regarding the role of trust in reciprocal self-disclosure are mixed (Collins and Miller 1994; Jones and Archer 1976; Lemay Jr and Melville 2014; Zimmer et al. 2010) and the extant literature does not take into account the context of conversational agents. The context of conversational agents is unique because these agents are nonhumans that often possess humanlike characteristics such as humanlike language capabilities. Therefore, users' inference of these characteristics can influence the trusting mechanism in unprecedented ways.

We believe that CA self-disclosure provides cognitive and affective bases on which users can develop an interpersonal trust with the CA. Moreover, we believe that the extent to which a user attributes humanlike attributes to a CA, i.e., anthropomorphism, plays an important role in explaining why self-disclosure by a nonhuman agent like a CA can lead to social responses in humans (Nass and Moon 2000).

In summary, understanding how trust is influenced in the process of reciprocal self-disclosure in the interaction of users with conversational agents is important for practice. However, the current literature fails to clearly explain the role of trust in reciprocal self-disclosure. In this research, we leverage the concept of anthropomorphism to understand how trust is affected in reciprocal self-disclosure in the unique context of conversational agents. Drawing upon prior literature in anthropomorphism and trust, we formulate a nomological network to connect CA self-disclosure to user self-disclosure. First, we introduce a psychological account of anthropomorphism from psychology and neuroscience literature to explain how users try to make sense of the self-

disclosure by the CA, i.e., a nonhuman agent. Some anthropologists have suggested that reciprocity is one of the main characteristics of being human (Fox and Tiger 1971; Leakey and Lewin 1978). Therefore, it is plausible that anthropomorphism plays a central role in users' decision to reciprocate the CA self-disclosure. 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 use two types of trustworthiness (i.e., cognition-based trustworthiness and affect-based trustworthiness) that can help unravel the cognitive and affective bases of reciprocal self-disclosure. We posit that the underlying motivations for anthropomorphism provide cognitive and emotional reasons for users to change their perception of cognition-based and affect-based trustworthiness of a CA. Therefore, we seek to answer the following research question: "What roles do anthropomorphism and trust play in reciprocal self-disclosure in the context of conversational agents?"

To answer our research question, we recruited 230 participants and conducted an experiment that employed a basic posttest-only randomized design comparing two treatments (Shadish et al. 2002, p. 258). CA self-disclosure was manipulated using a custom-developed CA, which provided either information with low level of intimacy about itself (low self-disclosure treatment condition) or information with high level of intimacy about itself (high self-disclosure treatment condition) before asking participants to reveal information about themselves.

Our study makes four key contributions to the literature. First, we demonstrate that the concepts of trustworthiness and trust could help explain why CA self-disclosure influences user self-disclosure. Second, by leveraging the concept of anthropomorphism, we delineate the importance of an artifact's perceived humanness in the process of reciprocation and explain why CA self-disclosure contributes to people's anthropomorphism of the artifact and consequently reciprocal

self-disclosure. Third, building on the concepts of cognition- and affect-based trustworthiness, we advance two new concepts, namely cognitive reciprocal self-disclosure and affective reciprocal self-disclosure. We present distinct theoretical explanations on how a user develops cognitive and affective understandings of a CA. We believe that these concepts provide a new framework to comprehend human-AI relationships in the emerging context of human-AI interaction.

THEORETICAL FOUNDATIONS

Self-Disclosure

Based on prior literature, 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 (Pearce and Sharp 1973; Wheeless and Grotz 1976). Self-disclosure plays a central role in the development and maintenance of relationships (Collins and Miller 1994). Scholars proposed different dimensions for self-disclosure (Mitchell et al. 2008). The most established dimensions are *depth*, which refers to the level of intimacy of the disclosure, and *breadth*, which refers to the amount of information exchanged (Altman and Taylor 1973).

According to social penetration theory (Altman and Taylor 1973), relationships develop through gradual increases in the depth and breadth of self-disclosure. Social penetration is the process of developing deeper intimacy with another person through different forms of vulnerability and the main route to deep social penetration is through verbal self-disclosure (Griffin 2012). Based on social penetration theory, personality is similar to a multilayered onion with outer layers representing public self and inner layers representing private self. As the relationship becomes stronger, the layers are unfolded and more intimate information is shared with the other party. In

this process, individuals relax their tight protecting boundaries and make themselves vulnerable to any use of the shared information about the self by the other party (Griffin 2012). Thus, the social penetration theory provides a foundation for understanding the development of relationships between individuals.

Self-disclosure in interpersonal relationships is reciprocal (Ehrlich and Graeven 1971). *Reciprocity* is the tendency to repay any benefits, gifts, and treatment or favors received by a party from another party (Derlega et al. 1973; Ehrlich and Graeven 1971; Lee and Choi 2017; Sprecher et al. 2013). While social penetration theory itself does not explain why individuals reciprocate, there has been some work that has sought to explain self-disclosure reciprocity (Cropanzano and Mitchell 2005; Sprecher et al. 2013). First, some scholars have suggested that in most contexts reciprocity is a norm or cultural mandate (e.g., Cropanzano and Mitchell 2005). According to Cropanzano and Mitchell (2005), those who follow this norm of how one should behave are obligated to behave reciprocally. Violation of this norm may make parties feel uncomfortable (Sprecher et al. 2013), and therefore individuals reciprocate the other party's self-disclosure by disclosing the same level of intimate information about the self. Second, some have argued that in initial interactions, when the rules of appropriate behavior are not well-defined, people follow the other party's behavior as a model or guide to reduce uncertainty about the expected behavior (Omarzu 2000; Rubin 1975). Finally, other scholars have suggested that in a relationship, parties strive to maintain an equitable exchange by reciprocating each other's behavior (Jones and Archer 1976), and they are uncomfortable with the imbalance in non-reciprocal disclosure (Sprecher et al. 2013). The main reason for this view is that the interdependence between two parties in a bidirectional relationship reduces risk and encourages cooperation (Molm 1994; Molm et al. 2007, 2009). When one party discloses information and the other reciprocates, a sequence of exchange starts. "Once the process

is in motion, each consequence can create a self-reinforcing cycle" (Cropanzano and Mitchell 2005, p. 876). While Jones and Archer originally developed the concept of equitable exchange based on the assumption that trust is "a special variant of equitable exchange" (1976, p. 182), there has been very little consensus whether trust is reciprocated or plays a significant role in reciprocation (Collins and Miller 1994; Lemay Jr and Melville 2014; Zimmer et al. 2010).

In interpersonal communication, individuals tend to like and have more positive impressions of others who disclose at higher levels compared to those who disclose at lower levels (Collins and Miller 1994; Jones and Archer 1976). One of the major reasons for this phenomenon is that the recipient of the disclosed information views the disclosed information as a rewarding outcome and a sign of the discloser's liking and desire to initiate a more intimate relationship. People are generally more attracted to people (or things) that provide them with rewarding outcomes (Cropanzano et al. 2016; Emerson 1976), and thus they like the person who discloses more information to them (Worthy et al. 1969). Although disclosing more personal information might be perceived as a signal of the discloser's interest in a more intimate relationship, it may not be appropriate in some situations. First, disclosing very personal information too early in the relationship may be perceived as too much, too soon (Altman and Taylor 1973; Collins and Miller 1994). Second, the positive impact of disclosure on liking may break down at extreme levels of intimacy (Archer and Berg 1978). Disclosing highly intimate information may sometimes be perceived as a violation of social norms and lead to a burden rather than a social reward for the recipient. Finally, the recipient's attribution for the discloser's behavior is crucial. The discloser is perceived more favorably if his or her disclosing behavior is attributed to a special quality of the recipient. In other words, people like disclosers who are more selective about to whom they disclose (Collins and Miller 1994).

Self-disclosure reciprocity 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 and Picard 2005). This effect of reciprocity on disclosure has been shown to be present in both online and offline contexts (Barak and Gluck-Ofri 2007; Taddicken 2014), among strangers with or without face-to-face interactions (Li et al. 2017), in computer-mediated communications (Jiang et al. 2013; Nguyen et al. 2012), and across different cultures (Katagiri et al. 2001). While reciprocity is one of the most established findings in the self-disclosure literature (Archer and Berg 1978), little is known about the mechanism through which this phenomenon takes place in human-AI interaction.

Nass and Moon, who were instrumental in shaping the literature on CASA, suggested that people mindlessly apply social rules and expectations, such as self-disclosure reciprocity to computers (2000). They argued that while anthropomorphism could provide an alternative explanation for people's social response to computers, anthropomorphism must be a "thoughtful, sincere belief that the object has human characteristics" (Nass and Moon 2000, p. 93). Nevertheless, other scholars empirically showed that anthropomorphism can be mindless, indicating that anthropomorphism could help explain people's social response to computers (Kim and Sundar 2012). 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 and Gratch 2010; Pickard et al. 2016), to the best of our knowledge no research investigated whether people use anthropomorphism as a mechanism to understand the self-disclosure by a nonhuman (e.g., a CA) in the process of reciprocity.

Anthropomorphism

Based on the previous literature, we define anthropomorphism as an 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; Gray et al. 2007). 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 them. Moreover, anthropomorphism is not only about treating an object as living, i.e., animism, but involves attributing uniquely humanlike characteristics to it. Anthropomorphism is a person's perception of the humanness of a nonhuman entity. While this perception might be right or wrong, the accuracy of the perception does not change the fact that the person indeed perceives the entity in a certain way (Epley, Waytz, et al. 2008).

Prior research identified three drivers of anthropomorphism (Epley et al. 2007). First, since the knowledge about oneself and humans is more 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, people have effectance motivation or the motivation to explain the behavior of other agents. Neuroscientists argue that our brain's main task is to predict its surrounding (Clark 2013). Since our best predictive model is the one about oneself, we leverage this model to predict the behavior of other humans as well (Broadbent 2017). Research has shown that we use the same neural system to understand the behavior of both humans and anthropomorphized agents (Castelli et al. 2000; Iacoboni et al. 2004). Therefore, anthropomorphism might give us more predictive power, or perception of which, when dealing with a nonhuman agent. Third, people have sociality

motivation or the desire for social contact. Therefore, people often create humans out of nonhumans to satisfy their need for social connectedness.

The effectance and sociality motivations indicate the outcome people seek when they anthropomorphize an agent. In other words, while such motivations can drive anthropomorphism, the outcome of the process is increased perception of predictability and perception of connectedness.

Researchers have shown that anthropomorphism influences trust (Waytz et al. 2014). They, however, provided limited evidence of the mechanism of the influence. For instance, Waytz et al. (2014) theorized that people perceive an anthropomorphized entity to be more competent than a non-anthropomorphized entity because 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. While this account discusses a channel of influence of anthropomorphism on trust through perceived competence of an entity, the reasons and the mechanism of why and how anthropomorphism influences trust is still understudied.

To better understand how trust is related to anthropomorphism and self-disclosure, we dig deeper into the trust literature and identify its cognitive and affective bases.

Trust

Prior research suggests that trust plays a key role in self-disclosure. Because self-disclosure involves some degree of risk (e.g., loss of control over personal information), trust is an essential component in this context (Dinev and Hart 2006). *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). Because of the multi-disciplinary nature of the development of the trust construct, there is some degree of confusion in the literature regarding the boundary of trust and other constructs (McKnight et al. 2002; Rousseau et al. 1998). Researchers have used the term "trust" to refer to many related constructs, adding to the already complex nature of trust (Sitkin and Roth 1993).

Trust is different from confidence and control (Mayer et al. 1995). Trust, alongside with perceived control, predicts a person's confidence level in an agent's cooperation (Das and Teng 1998). For instance, if a user, who is interacting with a conversational agent, knew there were laws protecting the privacy of any information disclosed during the interaction, she would perceive a higher level of confidence in the agent because of the perceived control over the outcome of the interaction. If control mechanisms completely guarantee the desired outcome, then there is little need for trust. Trust, therefore, is not control but a substitute for control (Rousseau et al. 1998).

Based on this approach to trust, we do not include deterrence-based and calculative-based (also referred to as calculus-based) trusts found in the IS literature as a part of our framework for the following reasons. First, deterrence-based trust is based on the consistency of behavior sustained by the threat of punishment (Lewicki and Bunker 1995). Thus, the trustee (i.e., the entity to be trusted by another entity) shows trusting behavior because s/he believes that the trustee would behave in the desired way to avoid punishment. Presence of punishment is, by definition, an external control, which leads the trustee to behave in a certain way. Therefore, deterrence-based trust is a form of control, which leads to behaviors that happen to be similar to behavioral outcomes of trust. Second, calculative-based trust is an extension of deterrence-based trust. Calculative-based trust is sustained by not only the fear of punishment for violating the trust but also the

rewards of maintaining it (Lewicki and Bunker 1995). The reason for the behavior of the trustee is the external control based on forms of reward and punishment. Therefore, calculative-based trust can also be a form of control.

Trust is one "unitary experience" (Komiak and Benbasat 2004), which is formed based on the trustor's perception of the trustee's trustworthiness (Mayer et al. 1995). Trustworthiness, which is often referred to as trusting beliefs in IS literature, can be formed based on cognition-based or affect-based evidence (Schoorman et al. 2007).

Cognition-Based and Affect-Based Trustworthiness[1]

Based on previous literature, we defined *cognition-based trustworthiness* as cognition-based expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party and *affect-based trustworthiness* as affect-based expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party (Mayer et al. 1995; McAllister 1995). Note that the difference between trust and trustworthiness is that trust is the "willingness" to be vulnerable based on the trustworthiness of the other party.

Trust could be based on a *cognitive* process through which the trustor discriminates among trustees (to decide who to trust). In this process, the trustor cognitively chooses who is trustworthy, a choice based on what s/he considers to be "good reasons" or evidence of trustworthiness (Lewis and

[1] In prior research, the cognitive and affective bases of trustworthiness were often captured by cognition-based and affect-based trust (McAllister 1995), not trustworthiness. A closer look at the items proposed in the literature for these constructs reveals that they tap into trustor's perceptions and feelings about trustee's characteristics (e.g., ability and reliability), not the trustor's immediate willingness to rely on the trustee for the specific task at hand. A possible reason for the inconsistency in the labeling of the constructs is that the widely accepted definition of trust and the seminal work on cognition-based and affect-based trust were published in the same year (see Mayer et al. 1995; McAllister 1995). Some more recent papers defined cognition-based and affect-based trust using Mayer et al.'s (1995) definition of trust while keeping McAllister's (1995) original operationalization, which conceptually refers to trustworthiness (e.g., Johnson and Grayson 2005). Nevertheless, for the sake of consistency, we adopt the terms cognition-based and affect-based trustworthiness.

Weigert 1985). As some researchers have noted, this process could be influenced by different cognitive biases (Weber et al. 2004); nevertheless, the basis of the process is cognition. Previous knowledge and information about the trustee and the context of trust provide some foundations for trust; however, they alone can never lead to trust. For instance, knowing that an agent has always behaved to one's benefit in previous interactions only increases the likelihood that it will continue to do so in the current interaction. But one can never be completely sure that the agent would not behave differently. Thus, to trust the person, one needs to go beyond the available evidence and make a prediction about an uncertain future. This cognitive "leap" is the cognitive element or one's belief about the trustee's trustworthiness (Lewis and Weigert 1985).

Trust can also be based on *affect*. The affect-based dimension of trustworthiness complements the cognition-based dimension. Affect could directly originate from the experience between the trustor and trustee. This affective element of trust is the emotional bond among parties in a relationship (Lewis and Weigert 1985) and is "grounded in reciprocated interpersonal care and concern" (McAllister 1995). A closer look at the items used by Johnson-George and Swap (1982) makes it clear that the construct refers to the emotional bases of why the trustor feels that the trustee cares about the well-being of him or her. In another seminal work in the context of close romantic relationships, Rempel et al. (1985) identified a very similar construct named faith. Faith "reflects an emotional security on the part of individuals, which enables them to go beyond the available evidence and feel, with assurance, that their partner will be responsive and caring despite the vicissitudes of an uncertain future" (Rempel et al. 1985, p. 97). The presence of faith highly depends on the trustor's perception of the trustee's motivation of being in the relationship (Rempel et al. 1985). Faith decreases as the perceived motivation moves from intrinsic motivations (i.e., the shared enjoyment of activities, mutual demonstration of affection and a sense of closeness, and

warmth associated with satisfying the other party's needs) to instrumental motivations (i.e., the rewards, such as direct services, goods, praise, and support, a party receives in the relationship because the other party is qualified to provide them) and to extrinsic motivations (i.e., the "rewards received from others outside of the relationship but mediated by involvement" in the relationship, such as access to new opportunities) (Rempel et al. 1985). Emotional trust and faith, alongside with trustor's perception of trustee's underlying motivation, 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). Some scholars directly used benevolence as affect-based trustworthiness (Shih et al. 2017), while others dropped benevolence from cognition-based trustworthiness when they independently measured affect-based trustworthiness (Komiak and Benbasat 2006).

Many studies combine cognition-based trustworthiness and affect-based trustworthiness, which are related to the trustor's beliefs about the trustee, and trust, which is related to the behavioral intentions of the trustor. However, several scholars have either argued against combining the two (Lewicki et al. 1998; Morgan and Hunt 1994; Singh and Sirdeshmukh 2000) or offered frameworks that rely on their separation (Mayer et al. 1995; McKnight et al. 1998, 2002). Accordingly, in this research, we keep trustworthiness and trust separate because "keeping them separate provides opportunities to study trust processes" (Singh and Sirdeshmukh 2000, p. 154).

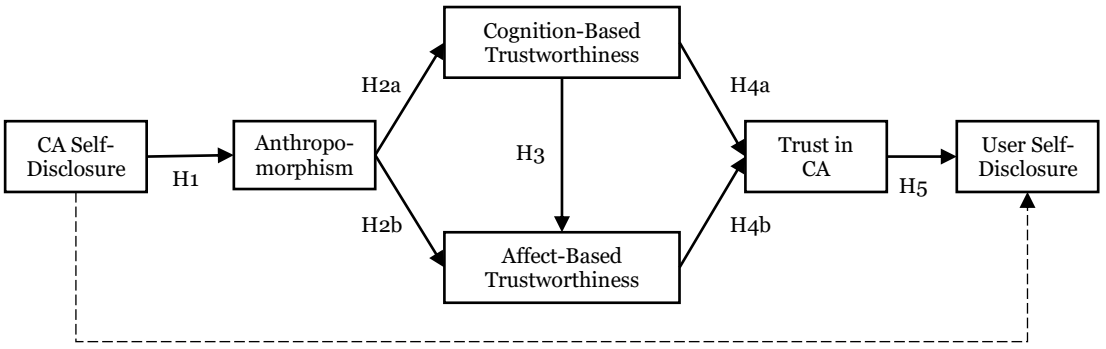
Table 1 present a summary of the relevant constructs in this study.

Table 1. Constructs		
Construct	**Definition**	**Informing Sources**
Self-disclosure	Any message about the self that a person communicates to an agent or that an agent communicates to a person.	Altman and Taylor 1973 Wheeless and Grotz 1976
Anthropomorphism	An 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, Waytz, et al. 2008 Waytz, Cacioppo, et al. 2010
Cognition-Based Trustworthiness	Cognition-based "expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" (Mayer et al. 1995, p. 712).	McAllister 1995 Mayer et al. 1995 Komiak and Benbasat 2004, 2006
Affect-Based Trustworthiness	Affect-based "expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" (Mayer et al. 1995, p. 712).	McAllister 1995 Mayer et al. 1995 Komiak and Benbasat 2004, 2006
Trust	"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, irrespective of the ability to monitor or control that other party" (Mayer et al. 1995, p. 712).	Mayer et al. 1995

THEORY DEVELOPMENT

The overarching idea of the current study is that self-disclosure by a CA triggers a reciprocal self-disclosure by the user through two channels, namely cognitive and affective reciprocity. Cognitive reciprocity operates through the presence of cognitive evidence, and affective reciprocity operates based on the presence of affective evidence of the trustworthiness of the CA. Our research model is depicted in Figure 1, in which the upper pathway represents cognitive reciprocity and the lower pathway represents affective reciprocity.[2]

[2] It is important to note that we are interested in understanding the role of anthropomorphism and trust in the mechanism of reciprocal self-disclosure. Therefore, we acknowledge that the proposed mechanism in this research does not necessarily fully mediate the effect CA's self-disclosure of user self-disclosure.



Note: Control variables not shown for clarity.

Figure 1. Research Model

Effect of CA Self-Disclosure on Anthropomorphism

The content of disclosed information is crucial in investigating the influence of one party's self-disclosure on the other party's behavior (Collins and Miller 1994). The disclosure of deep intimate information about the self has traditionally been a human behavior (Moon 2000; Nass and Moon 2000). Intimate disclosures include such information as self-concept, fears, values, vulnerabilities, and regrets (Altman and Taylor 1973). We argue that disclosing intimate information can influence a user's perception of different aspects of CA's state of mind. First, self-disclosure of emotions (e.g., fear) can signal that the CA is capable of experiencing some level of emotion. Being able to experience emotions qualifies the CA as a moral patient (Gray et al. 2012) and an entity that has the capacity to feel emotions is usually perceived to possess humanlike state of mind (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. In other words, it is unlikely that a CA that does not have agency has regrets or disappointments of what it has done or what it should have done. An entity with no agency has no moral responsibility (Monroe et al. 2014). Therefore, disclosure of intimate information by the CA provides evidence that it has agency and capacity to experience

emotions, the two dimensions that collectively define the human state of mind (Gray et al. 2007), which is central in defining anthropomorphism (Waytz, Gray, et al. 2010).

Furthermore, in the process of reciprocal self-disclosure, the CA discloses some information about itself. Self-disclosure by one party triggers the other party to reciprocate the action by disclosing information of the same type with a similar level of intimacy (Collins and Miller 1994; Ehrlich and Graeven 1971; Moon 2000). In this process, the disclosed information by the CA is relatable for the user. In other words, the reciprocation takes place when the user can find categories of information similar to the information the CA disclosed so that s/he can reciprocate the disclosure. The presence of similar information increases the chance that the user can use a human-based mental model to understand the CA. Since being human is the thing we know the best (Broadbent 2017), when faced with other humans and 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 the CA using her/his human-based concepts and therefore attribute a higher degree of human state of mind to it. Therefore, we hypothesize that:

H1: CA self-disclosure increases users' perceived anthropomorphism.

Effect of Anthropomorphism on Cognition- and Affect-Based Trustworthiness

On the one hand, the trustworthiness of a CA is essentially the user's prediction about the occurrence of desirable actions by the CA if the user were to trust it. Such a conceptualization of trustworthiness is in line with previous research (Deutsch 1958; Giffin 1967; Kee and Knox 1970; Mayer et al. 1995). On the other hand, anthropomorphism can help the user to better understand and analyze the CA's behavior. Prior research provided evidence that one of the main reasons

people anthropomorphize nonhumans is to increase their ability to predict the behavior of the artifact (Epley et al. 2007; Waytz, Morewedge, et al. 2010). Based on empirical studies, people are more likely to anthropomorphize artifacts that show apparently unpredictable behavior. Therefore, we posit that anthropomorphism can provide cognitive evidence for trustworthiness by increasing the user's ability to predict the artifact's behavior.

We argue that anthropomorphism in reciprocal self-disclosure can contribute to the perceived integrity dimension of cognition-based trustworthiness. People judge the integrity of the other party based on the perception that the other party "adheres to a set of principles that the trustor finds acceptable" (Mayer et al. 1995, p. 719). In other words, people seek both value congruence (Sitkin and Roth 1993) and consistency to form perceived integrity. An anthropomorphized agent is perceived to be more similar to self (Ames et al. 2008; Davis et al. 1996) and therefore more consistent in its behavior. Also, prior research showed that people tend to humanize those with whom they share values and dehumanize those who have a different set of values (Haslam and Loughnan 2014). Thus, we argue that it is unlikely that a user grants human attributes to a CA and then perceives the CA to follow a set of undesirable values, i.e., once a user granted human attributes to a CA, it is more likely that s/he assesses the CA's guiding values more favorably. This does not mean that people perceive humans to be more consistent and have more integrity than CAs, but it means that because granting different levels of humanness to a CA is completely the user's decision, s/he would perceive a CA that s/he granted more humanness as having more integrity than a CA that s/he granted less humanness.

Moreover, anthropomorphism can also contribute to the perceived ability of the agent because an agent with a humanlike state of mind is more likely to have more advanced means to fulfill expected tasks. More specifically, an anthropomorphized agent is perceived to have more agency,

which is an important part of the humanlike state of mind (Gray et al. 2007). An agent with more agency appears capable of fulfilling tasks, planning, and controlling their own actions (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, regardless of the desirability of the task, compared to a CA with little agency. Therefore, we hypothesize that:

H2a: Anthropomorphism increases cognition-based trustworthiness.

Prior research has suggested that people anthropomorphize entities to fulfill their need for social connectedness (Epley et al. 2007). Empirical evidence has suggested that lonely people are more likely to anthropomorphize robots (Epley, Akalis, et al. 2008; Eyssel and Reich 2013). Such evidence shows that anthropomorphism can increase the perceived human warmth in the anthropomorphized entity. Neuroscientists found warmth to be an important component (Fiske et al. 2007) or an indicator of perceived trustworthiness (Schweiger et al. 2013). Prior research in management also delineated warmth and caring as important characteristics of a benevolent person (Mayer et al. 1995). Therefore, a user who anthropomorphizes the agent is more likely to develop higher levels of affect-based trustworthiness in the agent. Therefore, we hypothesize that:

H2b: Anthropomorphism increases affect-based trustworthiness.

In H3, H4a, and H4b, we develop replication hypotheses that have been tested in other contexts (see Ha et al. 2016; Johnson and Grayson 2005; Komiak and Benbasat 2006; McAllister 1995; Wang et al. 2016). Inclusion of these hypotheses is crucial to our model because they help provide a theoretically grounded explanation for self-disclosure reciprocity in the context of CAs.

Relationships Among Cognition-Based Trustworthiness, Affect-Based Trustworthiness, and Trust

Some researchers posited that affect-based trustworthiness is formed based on cognition-based trustworthiness (Ha et al. 2016; Komiak and Benbasat 2006; Wang et al. 2016), yet others believe that as the relationship between trustor and trustee matures the link between cognition- and affect-based trustworthiness becomes bidirectional (Johnson and Grayson 2005; McAllister 1995). Some neuroscientists argue that cognition and affect are interdependent (Duncan and Barrett 2007; Storbeck and Clore 2007). They argue that "because … affect modulates sensory processing, any psychological process that draws on sensory information will have an affective quality to it. … affect makes external information from the world personally relevant to people" (Duncan and Barrett 2007, p. 1196). As such, the way a person perceives the other party's behavior in order to form cognition-based trustworthiness could be influenced by the same affects that provide the base for affect-based trustworthiness.

However, most IS researchers believe that at least during the initial stage of a relationship there is a causal effect from cognition-based trustworthiness to affect-based trustworthiness (Wang et al. 2016). The reason is that during this stage the trustor has had limited interactions with the trustee. Since affect-based trustworthiness is mainly based on trustor's emotions toward the trustee (not, e.g., trustor's mood) and emotional bonds take longer than cognitive perceptions to develop (McAllister 1995), cognition-based perceptions can provide a base for the formation of affect-based trustworthiness at the initial stage of a relationship. Furthermore, affect-based trustworthiness is related to the trustor's perception of trustee's motivation for being in the relationship (Rempel et al. 1985). In the lack of sufficient interaction, the trustor can rely on

cognition-based trustworthiness to understand the other party's motivations. Therefore, a user with positive cognitive trusting beliefs (i.e., high perceived cognition-based trustworthiness) is likely to have stronger feelings of comfort about relying on the other party (Komiak and Benbasat 2006).

We theorize that a user with high cognition-based trustworthiness in the conversational agent will tend to believe that the CA is knowledgeable, truthful, and honest. These positive rational attributes are likely to be associated with positive feelings in the user's mind. Therefore, in line with prior studies in IS and management literature (Ha et al. 2016; Johnson and Grayson 2005; Komiak and Benbasat 2006; McAllister 1995; Wang et al. 2016), we theorize that cognition-based trustworthiness provides a basis or an anchor for the formation of affect-based trustworthiness in the early stages of user's relationship with a CA. We hypothesize that:

H3: Cognition-based trustworthiness increases affect-based trustworthiness.

Trustworthiness is the main predictor of trust (Mayer et al. 1995). While other factors, such as generalized trust, might increase the effect of trustworthiness on trust, the effect of trustworthiness on trust has been found to be robust across different IS contexts (Gefen et al. 2003; McKnight et al. 2002). A person's willingness to make herself/himself vulnerable to the actions of the other is based on the available cognitive and affective evidence (Schoorman et al. 2007). Thus, when the user has good reasons supporting that the CA has the ability and integrity needed in an interpersonal relationship, she is more likely to be willing to make herself vulnerable to the actions of the CA by disclosing information about herself. Also, when the user feels caring and warmth from the CA, she is more likely to go beyond the available evidence and feel that the CA will remain caring in the future. Therefore, we hypothesize that:

H4a: Cognition-based trustworthiness *increases trust in CA.*
H4b: Affect-based trustworthiness *increases trust in CA.*

Effect of Trust on User Self-Disclosure

Given that the privacy concerns are associated with the disclosure of information to conversational agents (Saffarizadeh et al. 2017), a user's self-disclosure to the CA can be regarded as risky behavior. As in any risk-taking behavior, the user assesses the level of confidence in the expected desirable outcome (Das and Teng 1998). The user can develop confidence in the outcome by having either control over the outcome or trust in the agent (Das and Teng 1998). Since in the context of our research we assume that the user does not have any control over CA's behavior, the user's confidence in the behavior originates from trust in the agent. Trusting the agent means acting as if the CA's uncertain future behavior is certain and will yield desirable outcomes for the user (Rempel et al. 1985). Therefore, we hypothesize that:

H5: Trust in CA increases user self-disclosure.

RESEARCH METHOD

Experiment Design

We tested our hypotheses using an experiment with basic posttest-only randomized design comparing two treatments (Shadish et al. 2002, p. 258). We randomly assigned participants to two conditions (low and high) for CA self-disclosure. This design is robust to several threats to the validity of the effect of CA self-disclosure. For example, random assignment eliminates selection threats and minimizes the impact of maturation, history, and regression threats because both groups are expected to be influenced by these threats similarly.

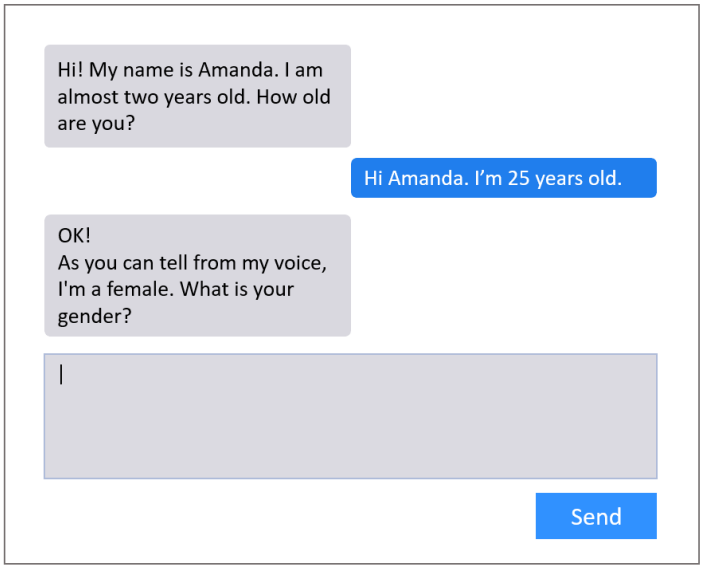
We recruited 230 participants of which 208[3] (95 females, 113 males, and 0 other, with an average age of 36.1 ranging from 19 to 71) followed the instructions. The number of participants who failed to follow the instructions was not significantly different for the two experimental groups.

We recruited participants from Amazon's Mechanical Turk to ensure a diverse sample (Buhrmester et al. 2011; Chandler et al. 2019; Mason and Suri 2012). Recent findings suggest that the quality of the data from MTurk surveys with attention-check questions is comparable to that of student subjects (Aruguete et al. 2019) and the generalizability of the results from MTurk samples is comparable to that from national samples (Coppock 2019). Subjects from Mechanical Turk are suited to our research objective because they have some experience using digital technology. We used the Cloud Research service (Litman et al. 2017) to ensure one-time participation by removing the participants who participated in our pilot studies, and by blocking participants from duplicate IP addresses and geo-locations. Even though recent studies suggest that most MTurkers are new to the platform and therefore not too familiar with manipulations and measures (Robinson et al. 2019), some Turkers may participate in many surveys per day (Paolacci et al. 2010) and therefore might be familiar with some measures or experience survey fatigue. To address such issues, we set our recruitment criteria to include participants who had a record of finishing between 500 and 10,000 HITs (Human Intelligence Tasks, which are the tasks posted on MTurk marketplace) with more than 97% acceptance rate. The experiment took 10 minutes on average and we compensated all participants with a $1.00 payment.

[3] While we chose to drop the 22 participants who failed to follow the instructions, we did perform 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.

The Conversational Agent

We developed a conversational agent named Amanda to increase the external validity of our study. We leveraged the latest text-to-speech technologies. We used Amazon's AWS Polly to produce humanlike synthesized voices (Amazon 2019). The use of such technologies helped make our conversational agent behave similarly to actual conversational agents in the market, enhancing the generalizability of the results. Moreover, we increased the degree of psychological realism of our study by creating an engaging task environment for participants (Berkowitz and Donnerstein 1982).



In the reciprocation condition, Amanda reveals some information about itself and then asks the user a question related to the disclosed information. The participant responds to the question.

Figure 2. A Sample of Participant's Interaction with Amanda (Text-Based)

Experiment Procedure

We asked the participants to open our web app on their browsers. We then asked participants to join a conversation with the conversational agent (Amanda). Amanda would start the conversation by introducing itself. Then, Amanda would begin a reciprocal question and answer round. Before

each question, Amanda would say a few sentences and then ask a question from 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 high CA self-disclosure condition, Amanda would disclose some information about itself that was related to the question it would ask the user. In the low CA self-disclosure condition, Amanda would not disclose any intimate information but say some procedural utterances such as "the next question has to do with your gender." By doing so, we controlled for the amount or breadth of disclosure in the two conditions (Moon 2000).

After each round of question-and-answer, 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 post-test, along with variables such as age, gender, level of education, prior experience using CAs, privacy concerns, and extroversion as control variables. Finally, we thanked and debriefed all participants.

Operationalization of Constructs

We ground our measurements in prior literature. However, we confirm the construct validity of our research by making sure that the measurement items carry their intended meaning in the context of our research (Van de Ven 2007). In doing so, we confirm that each item follows the definition of the construct in the new context. Appendix A presents the measures used in this study.

User's Self-Disclosure

We measured users' self-disclosure by capturing their actual utterances while interacting with the CA. We used the text of the utterances. Previous literature suggested measuring the breadth and depth of the disclosure (Altman and Taylor 1973). We used the word-count in each utterance to measure the breadth of the disclosure. We measured the depth of disclosure using the key properties of the depth of intimacy proposed by Altman and Taylor (1973). Accordingly, each utterance was rated from 1 to 7 in terms of depth by one of the authors.

Since only the honest user self-disclosure matters in the context of our study, after the user's interaction with the agent, we asked the participant to indicate how much of the information they disclosed was actually true. We told the participants that their answer to this question would not influence their compensation. We multiplied the depth of disclosure by the honesty percentage to create the user's self-disclosure.

We used the two measures as indicators of users' self-disclosure as a common factor because the two indicators are reflections of different aspects of the same underlying factor. For instance, the presence of more intimate information in an utterance is a manifestation of self-disclosure.

CA Self-Disclosure

We manipulated CA self-disclosure by adding more depth to the CA's disclosure in the experiment's high self-disclosure condition. We adopted Moon's (2000) method of handling self-disclosure, in which a computer asked each participant 15 questions and before each question it disclosed no or some information about itself. 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. Appendix B includes a complete list of the CA's disclosures. In the high disclosure condition, the CA started by disclosing public facts about itself. On its next turn of speaking, the

CA disclosed more private information about itself. This trend continued until the last turn in which the CA disclosed the most intimate information about itself. In the low disclosure condition, the CA disclosed no intimate information about itself before each question. However, we included roughly the same amount of non-disclosure text as in the high disclosure condition to rule out the plausible effect of the mere presence of more content (i.e., disclosure breadth) before each question on user self-disclosure.

Anthropomorphism

The most widely used operationalization of the anthropomorphism construct is based on the premise that anthropomorphism is about attribution of humanlike mental state or a mind to an agent (Waytz, Cacioppo, et al. 2010; Waytz, Morewedge, et al. 2010). Prior research has shown that people score humans as having the highest possession of mind compared to other entities such as God, animals, and robots (Gray et al. 2007). Possession of mind includes possession of free will and consciousness, having intentions, and being able to experience emotions (Gray et al. 2007; Gray and Wegner 2012; Waytz, Cacioppo, et al. 2010). Using a 7-point Likert-type scale ranging from "not at all" to "a great deal," we measured anthropomorphism by asking participants about the extent to which the CA seems to 1. have a mind of its own, 2. have intentions, 3. have free will, 4. have consciousness, and 5. experience emotions.

Cognition-Based Trustworthiness

We measured cognition-based trustworthiness using items from Wang et al.'s (2016) measures of cognition-based trustworthiness for recommendation agents. We appropriated the questions for the context of conversational agents with minimal changes. The measures include several indicators for each aspect of trustworthiness, namely, ability, benevolence, and integrity. In line with theory and previous research, we dropped benevolence because we measure affect-based trustworthiness

separately (Komiak and Benbasat 2006; Shih et al. 2017). We operationalized trustworthiness as a reflective second-order factor comprised of two first-order sub-constructs, i.e., ability and integrity. We used a 7-point Likert scale to measure three items for ability and four items for integrity.

Affect-Based Trustworthiness

To assess affect-based trustworthiness, we used the original measures developed by McAllister (1995). Many IS scholars either used a subset of the original items or created new items. For instance, while Wang et al.'s (2016) measures of affect-based trustworthiness for recommendation agents were based on McAllister's (1995) questionnaire, their AI artifact was not advanced enough for the authors to use many of the original questions. For example, one of the original items is "If I shared my problems with this person, I know (s)he would respond constructively and caringly," which is not relevant in the context of recommendation agents, but is relevant in the context of conversational agents. In this research, we used a 7-point Likert scale to measure three related items from the original questionnaire with minimal changes.

Trust in CA

We adopted trust measures from Mayer and Gavin (2005). These items reflect the concept of trust by capturing participants' willingness or intention to be vulnerable to the actions of the CA. Since the original scale was developed for trust in the context of a company, we used the items that could be properly appropriated for the context of our study. Furthermore, we did not use the reverse coded items, because they might tap into the concept of distrust, which some scholars argue that is different from trust (Dimoka 2010). Our operationalization of trust included three items measured using a 7-point Likert scale.

Control Variables

We controlled for participant's age, gender, level of education, previous experience in interacting with conversational agents such as Amazon's Alexa, Google Assistant, Apple's Siri, and Microsoft's Cortana. We also controlled for users' privacy concerns, which could be a predictor of their disclosure behavior (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).

ANALYSIS AND RESULTS

We used factor-based structural equation modeling (SEM) to assess the model using lavaan (version 0.6-3).

Measurement Model

We conducted a confirmatory factor analysis (CFA) on the saturated model to assess the measurement model. In a saturated model, all constructs can freely covary with each other; therefore, any misfit in the model is due to the inconsistency between the measurement model and data. The fit was evaluated using indices such as RMSEA, CFI, and SRMR, which, unlike chi-square, do not punish large sample sizes. While different thresholds have been suggested for acceptable levels of these fit measures, Hu and Bentler (1999) suggested a combination rule based on a simulation method. Using this approach, the fit is acceptable when CFI value is larger than 0.95, and either RMSEA is smaller than 0.06 or SRMR is smaller than 0.08. The fit measures for our model are CFI=0.963, RMSEA=0.054, and SRMR=0.057, indicating an acceptable fit.

We assessed convergent validity by examining factor loadings (lambda) and AVE values against the common threshold values of 0.70 and 0.50 respectively (Kline 2015). Lambda values were

larger than 0.7 for all items except for disclosure breadth ($\lambda_{breadth} = 0.60$) and are represented in the loadings table in Appendix C. While the lambda for disclosure breadth is less than 0.7, it still meets the 0.5 threshold proposed by some scholars (Hair et al. 2018). We chose to keep disclosure breadth in our measurement model to produce comparable results with extant literature on self-disclosure. AVE values were all above 0.5, indicating that each construct could explain more than half of the variation in its items. Lambda and AVE values provided support for the convergent validity of the measurement model.

We evaluated discriminant validity by showing that each construct had more common variance with its own items than with other constructs. In doing so, we examined whether the square root of AVE for each construct was larger than the construct's correlation with other constructs (Fornell and Larcker 1981). The inequality held for all constructs, providing support for discriminant validity. The composite reliability for all constructs was above 0.70, which is the threshold for reliability (Fornell and Larcker 1981). Table 2 presents descriptive statistics, correlations, and the square roots of the AVEs.

Table 2. Descriptive Statistics, Correlations, and $\sqrt{AVE}s$																
	M	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Age	36.10	11.61	NA													
2. Gender	1.54	0.50	-0.09	NA												
3. Education	4.22	1.25	-0.01	-0.07	NA											
4. Previous Experience	2.61	1.10	-0.01	0.10	0.03	NA										
5. Privacy Concerns	4.93	1.66	0.08	-0.07	0.06	-0.12	0.94									
6. Extroversion	3.57	1.68	0.17*	0.04	0.04	0.19**	-0.03	0.83								
7. CA Self-Disclosure	0.51	0.50	-0.19**	-0.09	-0.13	0.17*	-0.06	0.11	NA							
8. Anthropomorphism	3.59	2.04	0.00	0.02	-0.01	0.01	-0.11	0.07	0.140*	0.94						
9. Ability	5.16	1.51	-0.09	-0.14*	-0.10	0.13	-0.13	0.00	0.57**	0.02	0.96					
10. Integrity	4.98	1.44	-0.15*	-0.03	-0.15*	0.15*	-0.27**	0.08	0.61**	0.10	0.67**	0.88				
11. Cognition-Based Trustworthiness	5.07	1.32	-0.17*	-0.09	-0.17*	0.19**	-0.26**	0.06	0.77**	0.08	0.83**	0.92**	NA			
12. Affect-Based Trustworthiness	2.93	1.78	-0.25**	0.02	-0.13	0.16*	-0.06	0.13	0.85**	0.08	0.62**	0.71**	0.86**	0.86		
13. Trust in CA	4.05	1.90	-0.17*	0.05	-0.16*	0.10	-0.19**	0.01	0.57**	-0.06	0.56**	0.61**	0.76**	0.71**	0.92	
14. User Self-Disclosure	4.14	2.05	0.07	-0.12	-0.17*	0.10	-0.17*	-0.03	0.10	0.17*	0.21**	0.22**	0.25**	0.10	0.35**	0.76

Notes. N=208. $\sqrt{AVE}$ (square root of average variance extracted) values are represented on the diagonal and correlations are shown off diagonal. Gender is coded as 0=female and 1=male. Education is coded as 1=less than high school, 2=high school, 3=some college, 4=2-year college degree, 5=4-year college degree, 6=master's degree, 7=doctorate degree (including JD, MD). Previous Experience is coded as 1=never, 2=at least once a month, 3=at least once a week, 4=at least once a day.

Note that we provided the means and standard deviation for clarity. In an SEM approach, the common factors are estimated based on the model and the mean and standard deviation based on the items' average is not used.

Structural Model and Path Testing

To be able to interpret the path estimates, we need to ensure a good model fit when we combine the measurement model with the structural model. The fit indices for our model are CFI=0.959, RMSEA=0.055, and SRMR=0.064, indicating a satisfactory fit.

Hypothesis 1 stated that CA self-disclosure increases anthropomorphism. Model estimation showed a positive effect of CA self-disclosure on anthropomorphism ($\beta = 0.509, p < 0.05$), providing support for hypothesis 1. CA self-disclosure, along with control variables, explained 12.4% of the variance in anthropomorphism. While this amount might seem small, given that a large amount of anthropomorphism can be explained as a predisposition (Waytz, Cacioppo, et al. 2010), the R^2 is acceptable. Since CA self-disclosure was manipulated exogenously, the change in the level of anthropomorphism can be attributed to the manipulation with little concern about other endogenous factors influencing anthropomorphism. Therefore, the unexplained part of anthropomorphism (i.e., the noise) represents other unobserved factors that influence the participants in the two experimental groups equally. While adding more control variables as predictors of anthropomorphism can increase the R^2, a higher R^2 does not provide any advantage for an unbiased assessment of this relationship. Intensifying the manipulation of CA self-disclosure could also increase the R^2 of anthropomorphism; however, the analysis provided enough power to detect the effect. Moreover, the low R^2 could become a problem if the mediated effect of CA self-disclosure on user self-disclosure was not significant, indicating a disconnect in the theorized mechanism through which the reciprocal self-disclosure takes place. However, the mediated effect of CA self-disclosure on user self-disclosure through anthropomorphism, cognition-based and

affect-based trustworthiness, and consequently trust holds ($\beta = 0.077, p < 0.05$), indicating that even this amount of variation in anthropomorphism was enough to influence user self-disclosure.

Hypothesis 2a predicted that anthropomorphism increases cognition-based trustworthiness. The model estimation showed this to be the case ($\beta = 0.373, p < 0.01$), providing supporting evidence for hypothesis 2a. Anthropomorphism, along with control variables, explained 53.5% of the variation in cognition-based trustworthiness. In the same way, hypothesis 2b predicted that anthropomorphism increases affect-based trustworthiness. Model estimation showed this to be the case ($\beta = 0.451, p < 0.01$).

Hypothesis 3, in line with prior research on trust, theorized that cognition-based trustworthiness increases affect-based trustworthiness. The analysis supported this claim ($\beta = 0.763, p < 0.01$). As discussed in the development of hypothesis 3, the direction of this relationship is defined theoretically and based on prior research (Ha et al. 2016; Kanawattanachai and Yoo 2002; Komiak and Benbasat 2006; McAllister 1995; Schaubroeck et al. 2011; Wang et al. 2016). In this study, however, we did not investigate whether the cognition part of trustworthiness precedes the affective part. Note that changing the direction of hypothesis 3 does not change the direction and significance of any other path in the model. Even if this relationship were modeled as a bidirectional covariance, the rest of the model would hold. In all discussed alternative models, the fit measures would also be satisfactory. However, eliminating this relationship would lead to a lack of fit and unstable estimates due to the high correlation between the two constructs. In total, 78.2% of the variation in affect-based trustworthiness was explained by anthropomorphism, cognition-based trustworthiness, and the control variables.

Hypothesis 4 stated that cognition-based trustworthiness and affect-based trustworthiness increase trust. The estimated model provided support for both hypotheses by showing a positive association between cognition-based trustworthiness and trust ($\beta = 0.688, p < 0.01$) as well as affect-based trustworthiness and trust ($\beta = 0.363, p < 0.01$). The two trustworthiness components, along with the control variables, explained 53.1% of the variance in trust.

Hypothesis 5 theorized that trust increases user's self-disclosure. The empirical model supported this claim by indicating a positive association between trust and user's self-disclosure ($\beta = 0.335, p < 0.01$). Trust, alongside with control variables, explained 21.8% of the variance in user's self-disclosure. Table 3 presents a summary of the findings.[4]

Table 3. SEM Results: Explaining Anthropomorphism, Cognition- and Affect-Based Trustworthiness, Trust, and User's Self-Disclosure

	AP	CT	AT	TIC	UD
Control Variables					
Age	-0.04 (0.01)**	0.00 (0.01)	-0.01 (0.01)*	0.00 (0.01)	0.02 (0.01)*
Gender	-0.42 (0.26)	-0.09 (0.14)	0.35 (0.16)*	0.27 (0.21)	-0.53 (0.26)*
Education	-0.22 (0.11)*	-0.05 (0.05)	0.02 (0.06)	-0.06 (0.08)	-0.19 (0.1)
Previous Experience	0.24 (0.12)*	0.03 (0.06)	0.01 (0.07)	-0.03 (0.09)	0.10 (0.12)
Privacy Concerns	-0.01 (0.08)	-0.13 (0.04)**	0.11 (0.05)*	-0.07 (0.07)	-0.12 (0.08)
Extroversion	0.13 (0.08)	-0.01 (0.04)	0.06 (0.05)	-0.07 (0.06)	-0.08 (0.08)
Independent Variables					
CD	0.51 (0.26)*				0.74 (0.26)**
AP		0.37 (0.04)***	0.45 (0.07)***		
CT			0.76 (0.16)***	0.69 (0.24)**	
AT				0.36 (0.13)**	
TIC					0.36 (0.07)***
R^2	0.124	0.535	0.782	0.531	0.218

Notes:

a. *Key: CD: CA Self-Disclosure, AP: Anthropomorphism, CT: Cognition-Based Trustworthiness, AT: Affect-Based Trustworthiness, TIC: Trust in CA, UD: User's Self-Disclosure*
b. *N=208*
c. ** p<0.05; ** p<0.01; *** p<0.001; one-tailed tests were used for directional hypotheses and two-tailed tests for the rest of the relationships.*
d. *Unstandardized regression coefficients are shown.*
e. *Numbers in parentheses are the standard errors*

[4] As discussed, we used the honest disclosures to measure user self-disclosure, however, the results do not change in terms of direction and significance when we use the raw disclosure values.

Robustness Checks

There are some concerns regarding the validity of our findings. First, we modeled all constructs in our research as common factors of their measurement items. While our choice was driven by the nature of the measures, many prior studies have modeled these constructs as the average of their measurement items, which might lead to different results. Second, the standard errors of our estimates could be subject to heteroskedasticity, because we did not use a heteroskedasticity robust method to estimate the variances. Third, we established the nomological network connecting CA self-disclosure and user self-disclosure through anthropomorphism and trust; however, we did not test whether the proposed pathways between CA self-disclosure and user self-disclosure are statistically meaningful in our sample. Forth, we tested our research model in the context of a text-based conversational agent. While many conversational agents are text-based, conducting a similar experiment with a voice-based conversational agent can increase the generalizability of our findings.

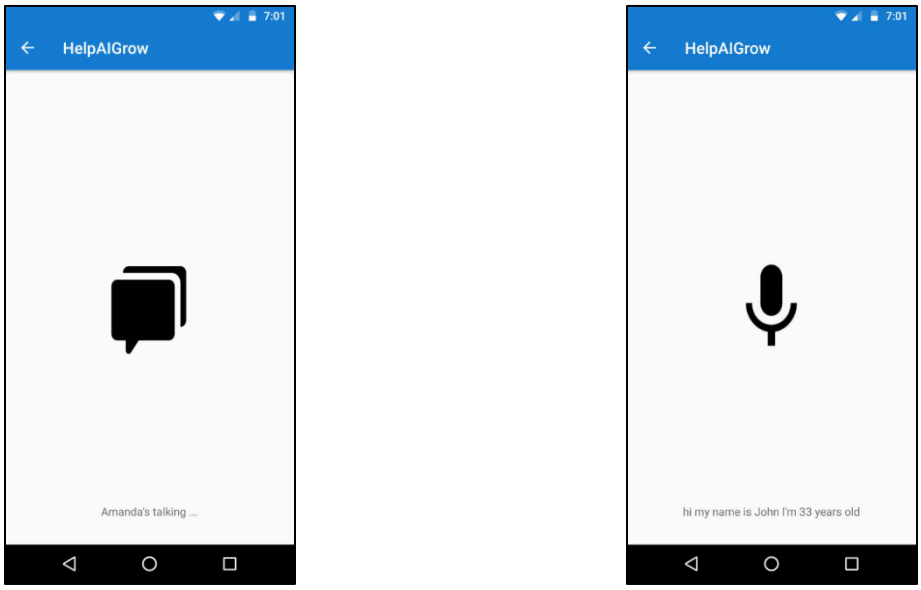
We took the following steps to address these concerns. First, to address the issue of how our constructs were modeled, we created each construct as a composite variable by averaging the measures of the construct (Cronbach's α for anthropomorphism, cognition-based trustworthiness, affect-based trustworthiness, and trust are 0.97, 0.94, 0.90, and 0.94 respectively). Second, to address the issue of possible heteroskedasticity in our standard errors, we re-estimated the model with hierarchical regression by following Preacher and Hayes' (2008) approach with 10,000 bootstrap samples and heteroskedasticity robust standard errors for significance tests (Davidson and MacKinnon 1993). The results confirmed all findings from our SEM model (see Appendix D).

Third, to further probe the role of anthropomorphism and trust on reciprocal self-disclosure in the context of CAs, we tested all the indirect causal paths between the CA and user self-disclosures. Since the indirect effects are the product term of multiple estimates, the test statistic could be unstable and thus unreliable. To remedy this issue, we estimated the indirect effects using 10,000 bootstrap samples. First, we examined the path via cognition-based trustworthiness (CA self-disclosure → anthropomorphism → cognition-based trustworthiness → trust → user self-disclosure). We found support for a positive purely cognitive indirect effect of CA self-disclosure on user self-disclosure (BootLLCI and BootULCI > 0). Second, we examined the path via affect-based trustworthiness, which is the sum of two paths: path 1. CA self-disclosure → anthropomorphism → cognition-based trustworthiness → affect-based trustworthiness → trust → user self-disclosure, and path 2. CA self-disclosure → anthropomorphism → affect-based trustworthiness → trust → user self-disclosure. We found supporting evidence for a positive affect-based indirect effect of CA self-disclosure on user self-disclosure through both paths (BootLLCI and BootULCI > 0).

Finally, in order to address the issue of generalizability, we replicated our previous experiment using a voice-enabled conversational agent. More specifically, we asked the participants to download and run the mobile version of our app on their Android phones. We recruited 140 participants who had an acceptance rate of at least 99% on Amazon's Mechanical Turk platform, whose native language was English[5], and who had an Android phone. From the 140 recruited participants, 98 (33 female, 65 male, and 0 other, with an average age of 34.7, ranging from 22 to 72, median education of 4-year college, and a median experience of interacting with CAs at least

[5] Our pilot studies showed that the speech-to-text service did not precisely detect the utterances of non-native speakers.

once a week) passed the attention check questions and were native English speakers. The experiment took about 10 minutes on average, and all participants were compensated with a $2.50 payment.



In the reciprocation condition, Amanda reveals some information about itself and then asks the user a question related to the disclosed information (left screenshot). The participant responds to the question (right screenshot).

Figure 3. A Sample of Participant's Interaction with Amanda (Voice-Based)

For this experiment, we developed a voice-enabled conversational agent for the Android platform. The app required the participant to enter the experiment ID they received in the recruitment message. 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 to interact with the agent in a conversation similar to our previous experiment. It is a known problem that many conversational agents cut the users off in the middle of a sentence if the user pauses for too long. To remedy this potential problem, we gave a grace period of two seconds after each utterance by the participant so that he or she could continue talking. Participants would see a live text stream of their utterances

on the screen as being detected by the agent. To do so, we developed an Android client to use the Google Cloud speech-to-text service. We used the end-of-sentence signal provided by this service to calculate the timing of the grace period. The app automatically extended the grace period after each new utterance by the participant. Figure 3 shows a sample of a participant's interaction with Amanda.

Due to the relatively small sample size in this experiment, we used hierarchical regression with heteroscedasticity robust standard error estimator and 10,000 bootstrap samples to estimate our model. In doing so, we created each construct by averaging all of its measurement items (Cronbach's α for anthropomorphism, cognition-based trustworthiness, affect-based trustworthiness, and trust are 0.95, 0.94, 0.85, and 0.86 respectively). The results of this experiment added robustness to our results by confirming all of our findings (Table 4).[6]

Table 4. Hierarchical Regression Analysis for the Mediated Effect of CA Self-Disclosure on User Self-Disclosure

	AP	CT	AT	TIC	UD
Control Variables					
Constant	2.34 (0.64)***	4.19 (0.88)***	-0.63 (0.80)	1.00 (1.33)	4.18 (1.59)**
Age	0.00 (0.01)	0.01 (0.01)	0.01 (0.01)	0.00 (0.02)	0.02 (0.03)
Gender	0.12 (0.24)	-0.59 (0.22)**	-0.28 (0.26)	0.12 (0.34)	-1.00 (0.41)
Education	-0.01 (0.11)	-0.01 (0.08)	0.12 (0.09)	0.13 (0.14)	-0.12 (0.18)
Previous Experience	0.20 (0.10)*	-0.01 (0.09)	-0.04 (0.08)	0.13 (0.12)	-0.21 (0.18)
Privacy Concerns	0.04 (0.09)	0.07 (0.09)	-0.09 (0.07)	-0.08 (0.09)	0.03 (0.14)
Extroversion	-0.13 (0.06)	-0.07 (0.06)	-0.04 (0.07)	0.02 (0.10)	0.03 (0.13)
Independent Variables					
CD	0.86 (0.23)***	0.21 (0.20)	0.10 (0.23)	-0.07 (0.33)	0.90 (0.40)*
AP		0.61 (0.09)***	1.00 (0.11)***	-0.16 (0.20)	0.33 (0.32)
CT			0.31 (0.11)**	0.49 (0.17)**	-0.37 (0.21)
AT				0.30 (0.14)*	-0.24 (0.21)
TIC					0.40 (0.17)**
R^2	0.206	0.472	0.704	0.348	0.186

Notes:
a. Key: CD: CA Self-Disclosure, AP: Anthropomorphism, CT: Cognition-Based Trustworthiness, AT: Affect-Based Trustworthiness, TIC: Trust in CA, UD: User's Self-Disclosure
b. N=98
*c. * p<0.05; ** p<0.01; *** p<0.001; one-tailed tests were used for directional hypotheses and two-tailed tests for the rest of the relationships.*
d. Unstandardized regression coefficients are shown.
e. Numbers in parentheses are the standard errors

[6] The results stay the same when we use the full dataset without removing non-English speakers.

DISCUSSION

In this research, we investigated the role of anthropomorphism and trust in the mechanism through which self-disclosure by a conversational agent leads to self-disclosure by a user. Our findings showed that reciprocity happens based on cognition-based and affect-based changes in user's perception of the agent. These findings have several important implications for research and practice.

Implications for Research

To the best of our knowledge, prior research on reciprocal self-disclosure in human-computer interaction has not examined the role of anthropomorphism. One reason for this may be due to the fact that Nass and Moon (2000) argued that anthropomorphism is not the reason why reciprocal self-disclosure occurs in human-computer interaction. Their logic was based on the notion that people who interact with computers will be mindful of the fact that the computer is not a person. However, as Kim and Sundar (2012) point out, anthropomorphism can be a mindless process and therefore it is important to reexamine the role of anthropomorphism in self-disclosure reciprocity in the context of human-computer interaction.

In this research, we contributed to the literature by highlighting that anthropomorphism plays a major role in how users understand the behavior of conversational agents. We theorized and empirically showed that users attribute humanlike state of mind to an agent when it discloses information about itself. We argued that such an attribution helps users make sense of the agent's disclosure behavior. The disclosure of relatable information by the agent makes the user anthropomorphize the agent, a process that based on previous research is associated with the activation of mirror neurons in the brain (Waytz, Morewedge, et al. 2010).

We contributed to the trust literature by providing theoretical links between anthropomorphism and cognition-based and affect-based trustworthiness. In doing so, we conceptually bridged the prior research on motivations of anthropomorphism (Epley et al. 2007) and the research on cognitive and affective bases of trust (McAllister 1995). Based on our findings, when the context allows (e.g., in the presence of self-disclosure by a conversational agent), people engage in the process of anthropomorphism. This process, in turn, provides the users the means to better form a cognitive assessment of the agent's competence and integrity (i.e., cognition-based trustworthiness) and establish a closer relational connection (i.e., affect-based trustworthiness) with the agent.

While anthropomorphism increases both cognition-based and affect-based trustworthiness, based on our theory, the reason for the two increases is not the same. Prior literature identified effectance motivation, i.e., the motivation to explain uncertainty in one's surrounding, 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). This study extends this literature by showing that the formation of cognition-based trustworthiness, which is conceptually related to effectance motivation, and formation of affect-based trustworthiness, which is conceptually related to sociality motivation, can be enhanced by anthropomorphizing the agent.

In this research, we introduced two new concepts namely cognitive reciprocal self-disclosure and affective reciprocal self-disclosure. We investigated these concepts by showing their different roles in the reciprocity phenomenon. We believe such a conceptual distinction is important because it not only provides a new framework to assess reciprocity but also delineates the role of affect in human-AI interaction. Affective reciprocal self-disclosure can help scholars in the robotic

companionship stream of research better understand how affect-based trustworthiness develops in reciprocal interaction. In contrast, cognitive reciprocal self-disclosure can help scholars in fields such as military robotics to study the types of disclosure that induce cognition-based trustworthiness and prompts cognitive reciprocal disclosure.

In this research, we applied theoretical lenses from the communication and social psychology disciplines to explain the role of anthropomorphism and trust in reciprocity in human-AI interaction. Such theoretical lenses can enable more theory-driven research on disclosure behavior of users in the emerging context of human-AI interaction.

Implications for Practice

Developers can use the findings from this study to modify their CAs to increase the amount and depth of information that they obtain through user self-disclosure. This will, in turn, help developers to exploit disclosed information via analytic tools to create strategic advantage, adapt business models, and target advertisements (Schmarzo 2013). The disclosed information can also be used to create a more personalized experience for the user, which can increase the usability of the artifact.

However, it is important for developers to understand that the disclosure should be non-manipulative. Prior studies suggested that manipulative self-disclosure could lead to opposite results and make the person suspicious (Collins and Miller 1994). In addition, disclosing intimate information too early could also lead to unease in the user (Altman and Taylor 1973). To use the results from this research, therefore, developers need to make the disclosure gradual. They can start from information with low intimacy and move to more intimate information over time. By

doing so, we argue, users are more likely to develop high trust, and ultimately to share more information with the CA.

We believe that practitioners can leverage self-disclosure by the CA to increase the extent to which users anthropomorphize it. Given the conversational nature of the interaction in the context of conversational agents, developers can easily reveal relatable information about the CA and significantly increase the perceived humanness of the agent with little financial investment. We believe that the relatability of the information is key to make disclosure a tool to induce anthropomorphism. For instance, users might experience some problems when interacting with a CA. While the negative effects of the problems are inevitable, the developers can frame and reveal the problems as the CA's (the persona's) personal failures, which makes the CA more relatable and humanlike.

Our findings can also help users to become more cognizant of the ways in which CAs may be extracting personal information. This will help users to make more informed decisions as to what information to share with their CAs and by extension the vendors of these products.

Limitations and Future Directions

In this research, we probed cognitive and affective bases of reciprocity. Because of the complex nature of this phenomenon, however, there could be other paths, such as people's habit, that can also contribute to self-disclosure reciprocity. Therefore, this study is not an attempt to develop a comprehensive mediation model that explains reciprocity, but an initial effort to develop a model of reciprocity that is useful for researchers in the human-AI interaction field.

In our manipulation of CA self-disclosure, we did not distinguish between disclosure and its content. For example, disclosure of weakness by the CA is a deep (intimate) disclosure. However,

we did not differentiate between the fact that the CA has a weakness and the fact that it revealed it to the participant. While this approach is in line with the prior research, we also designed the content of the disclosure in a way to control for the unforeseen effects of the content on the outcome variables. For instance, since disclosure of weakness can influence the perceived ability of the CA, we balanced the positive and negative statements in the disclosure regarding the ability of the CA. Future studies can enhance our approach and control for the content of disclosure by disclosing the same information disclosed in the high disclosure condition via a reading task before the interaction starts and keeping the same level of disclosure breadth in the interaction.

Our theory involved some hypotheses that predicted causal paths between constructs such as anthropomorphism, trustworthiness, and trust, all of which reside in the mind of the user. Since the formation of perceptions, beliefs, and intentions might happen simultaneously in the brain (Clark 2013), we could not empirically ensure the precedence of the cause. We, however, relied on theoretical reasoning to argue the causal nature of the relationship. For instance, since trustworthiness is about trusting beliefs and trust is about trusting intention (McKnight et al. 2002), in line with previous research, we assumed that beliefs precede intentions. 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 affection.

CONCLUSION

Given that voice-based computing is expected to experience rapid growth for the foreseeable future, it is important to understand the contexts within 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 conversational agents which can prompt us to reveal more personal information about ourselves than we may be comfortable disclosing under normal circumstances. We hope that our study increases awareness of this phenomenon and inspires other researchers to contribute to the academic discourse on conversational agents.

REFERENCES

Altman, I., and Taylor, D. A. 1973. *Social Penetration: The Development of Interpersonal Relationships.*, Holt, Rinehart & Winston.

Amazon. 2019. "Amazon Polly: Turn Text into Lifelike Speech Using Deep Learning," *Amazon Web Services, Inc.*

Ames, D. L., Jenkins, A. C., Banaji, M. R., and Mitchell, J. P. 2008. "Taking Another Person's Perspective Increases Self-Referential Neural Processing," *Psychological Science* (19:7), pp. 642–644.

Apple. 2019. "Privacy - Approach to Privacy - Apple."

Archer, R. L., and Berg, J. H. 1978. "Disclosure Reciprocity and Its Limits: A Reactance Analysis," *Journal of Experimental Social Psychology* (14:6), pp. 527–540.

Aruguete, M. S., Huynh, H., Browne, B. L., Jurs, B., Flint, E., and McCutcheon, L. E. 2019. "How Serious Is the 'Carelessness' Problem on Mechanical Turk?," *International Journal of Social Research Methodology* (22:5), Taylor & Francis, pp. 441–449.

Barak, A., and Gluck-Ofri, O. 2007. "Degree and Reciprocity of Self-Disclosure in Online Forums," *CyberPsychology & Behavior* (10:3), pp. 407–417.

Berkowitz, L., and Donnerstein, E. 1982. "External Validity Is More than Skin Deep: Some Answers to Criticisms of Laboratory Experiments.," *American Psychologist* (37:3), pp. 245–257.

Bickmore, T., and Cassell, J. 2005. "Social Dialogue with Embodied Conversational Agents," *Advances in Natural Multimodal Dialogue Systems* (30), pp. 23–54.

Bickmore, T. W., and Picard, R. W. 2005. "Establishing and Maintaining Long-Term Human-Computer Relationships," *ACM Transactions on Computer-Human Interaction (TOCHI)* (12:2), pp. 293–327.

Bowden, K. K., Wu, J., Cui, W., Juraska, J., Harrison, V., Schwarzmann, B., Santer, N., and Walker, M. 2019. "SlugBot: Developing a Computational Model and Framework of a Novel Dialogue Genre," *2nd Proceedings of Alexa Prize*.

Broadbent, E. 2017. "Interactions with Robots: The Truths We Reveal about Ourselves," *Annual Review of Psychology* (68), pp. 627–652.

Buhrmester, M., Kwang, T., and Gosling, S. D. 2011. "Amazon's Mechanical Turk: A New Source of Inexpensive, yet High-Quality, Data?," *Perspectives on Psychological Science* (6:1), pp. 3–5.

Castelli, F., Happé, F., Frith, U., and Frith, C. 2000. "Movement and Mind: A Functional Imaging Study of Perception and Interpretation of Complex Intentional Movement Patterns," *Neuroimage* (12:3), pp. 314–325.

Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J., and Litman, L. 2019. "Online Panels in Social Science Research: Expanding Sampling Methods beyond Mechanical Turk," *Behavior Research Methods* (51:5), Springer, pp. 2022–2038.

Clark, A. 2013. "Whatever next? Predictive Brains, Situated Agents, and the Future of Cognitive Science," *Behavioral and Brain Sciences* (36:3), pp. 181–204.

Collins, N. L., and Miller, L. C. 1994. "Self-Disclosure and Liking: A Meta-Analytic Review.," *Psychological Bulletin* (116:3), p. 457.

Coppock, A. 2019. "Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach," *Political Science Research and Methods* (7:3), Cambridge University Press, pp. 613–628.

Cropanzano, R., Anthony, E., Daniels, S., and Hall, A. 2016. "Social Exchange Theory: A Critical Review with Theoretical Remedies," *Academy of Management Annals*, Annals–2015.

Cropanzano, R., and Mitchell, M. S. 2005. "Social Exchange Theory: An Interdisciplinary Review," *Journal of Management* (31:6), pp. 874–900.

Das, T. K., and Teng, B.-S. 1998. "Between Trust and Control: Developing Confidence in Partner Cooperation in Alliances," *Academy of Management Review* (23:3), pp. 491–512.

Davidson, R., and MacKinnon, J. G. 1993. "Estimation and Inference in Econometrics," *OUP Catalogue*, Oxford University Press.

Davis, M. H., Conklin, L., Smith, A., and Luce, C. 1996. "Effect of Perspective Taking on the Cognitive Representation of Persons: A Merging of Self and Other.," *Journal of Personality and Social Psychology* (70:4), p. 713.

Derlega, V. J., Harris, M. S., and Chaikin, A. L. 1973. "Self-Disclosure Reciprocity, Liking and the Deviant," *Journal of Experimental Social Psychology* (9:4), pp. 277–284.

Deutsch, M. 1958. "Trust and Suspicion," *Journal of Conflict Resolution* (2:4), pp. 265–279.

Dimoka, A. 2010. "What Does the Brain Tell Us about Trust and Distrust? Evidence from a Functional Neuroimaging Study," *Mis Quarterly*, pp. 373–396.

Dinev, T., and Hart, P. 2004. "Internet Privacy Concerns and Their Antecedents-Measurement Validity and a Regression Model," *Behaviour & Information Technology* (23:6), pp. 413–422.

Dinev, T., and Hart, P. 2006. "An Extended Privacy Calculus Model for E-Commerce Transactions," *Information Systems Research* (17:1), pp. 61–80.

Dinev, T., McConnell, A. R., and 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), pp. 639–655.

Donnellan, M. B., Oswald, F. L., Baird, B. M., and Lucas, R. E. 2006. "The Mini-IPIP Scales: Tiny-yet-Effective Measures of the Big Five Factors of Personality.," *Psychological Assessment* (18:2), American Psychological Association, p. 192.

Duncan, S., and Barrett, L. F. 2007. "Affect Is a Form of Cognition: A Neurobiological Analysis," *Cognition and Emotion* (21:6), pp. 1184–1211.

Ehrlich, H. J., and Graeven, D. B. 1971. "Reciprocal Self-Disclosure in a Dyad," *Journal of Experimental Social Psychology* (7:4), pp. 389–400.

Emerson, R. M. 1976. "Social Exchange Theory," *Annual Review of Sociology* (2:1), pp. 335–362.

Epley, N., Akalis, S., Waytz, A., and Cacioppo, J. T. 2008. "Creating Social Connection through Inferential Reproduction: Loneliness and Perceived Agency in Gadgets, Gods, and Greyhounds," *Psychological Science* (19:2), pp. 114–120.

Epley, N., Waytz, A., Akalis, S., and Cacioppo, J. T. 2008. "When We Need a Human: Motivational Determinants of Anthropomorphism," *Social Cognition* (26:2), pp. 143–155.

Epley, N., Waytz, A., and Cacioppo, J. T. 2007. "On Seeing Human: A Three-Factor Theory of Anthropomorphism.," *Psychological Review* (114:4), pp. 864–886.

Eyssel, F., and Reich, N. 2013. "Loneliness Makes the Heart Grow Fonder (of Robots)—On the Effects of Loneliness on Psychological Anthropomorphism," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, pp. 121–122.

Fiske, S. T., Cuddy, A. J., and Glick, P. 2007. "Universal Dimensions of Social Cognition: Warmth and Competence," *Trends in Cognitive Sciences* (11:2), pp. 77–83.

Fornell, C., and Larcker, D. F. 1981. "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error," *Journal of Marketing Research* (18:1), pp. 39–50.

Fox, R., and Tiger, L. 1971. *The Imperial Animal*, New York: Holt, Rinehart and Winston.

Gefen, D., Karahanna, E., and Straub, D. W. 2003. "Trust and TAM in Online Shopping: An Integrated Model," *MIS Quarterly* (27:1), pp. 51–90.

Giffin, K. 1967. "The Contribution of Studies of Source Credibility to a Theory of Interpersonal Trust in the Communication Process.," *Psychological Bulletin* (68:2), p. 104.

Google. 2019. "Data Security & Privacy on Google Home - Google Home Help."

Gray, H. M., Gray, K., and Wegner, D. M. 2007. "Dimensions of Mind Perception," *Science* (315:5812), pp. 619–619.

Gray, K., Knobe, J., Sheskin, M., Bloom, P., and Barrett, L. F. 2011. "More than a Body: Mind Perception and the Nature of Objectification.," *Journal of Personality and Social Psychology* (101:6), p. 1207.

Gray, K., and Wegner, D. M. 2012. "Feeling Robots and Human Zombies: Mind Perception and the Uncanny Valley," *Cognition* (125:1), pp. 125–130.

Gray, K., Young, L., and Waytz, A. 2012. "Mind Perception Is the Essence of Morality," *Psychological Inquiry* (23:2), pp. 101–124.

Griffin, E. A. 2012. *A First Look at Communication Theory/Em Griffin.*, New York: McGraw-Hill,.

Ha, H.-Y., John, J., John, J. D., and 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), pp. 213–235.

Hair, J. F., Black, W. C., Babin, B. J., and Anderson, R. E. 2018. *Multivariate Data Analysis*, (8th ed.), Cengage Learning EMEA.

Haslam, N., and Loughnan, S. 2014. "Dehumanization and Infrahumanization," *Annual Review of Psychology* (65), pp. 399–423.

Hu, L., and 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), pp. 1–55.

Iacoboni, M., Lieberman, M. D., Knowlton, B. J., Molnar-Szakacs, I., Moritz, M., Throop, C. J., and 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), pp. 1167–1173.

Jiang, L. C., Bazarova, N. N., and Hancock, J. T. 2013. "From Perception to Behavior: Disclosure Reciprocity and the Intensification of Intimacy in Computer-Mediated Communication," *Communication Research* (40:1), pp. 125–143.

Johnson, D., and Grayson, K. 2005. "Cognitive and Affective Trust in Service Relationships," *Journal of Business Research* (58:4), pp. 500–507.

Johnson-George, C., and 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), p. 1306.

Jonell, P., Bystedt, M., Dogan, F. I., Fallgren, P., Ivarsson, J., Slukova, M., Ulme Wennberg, J. L., Boye, J., and Skantze, G. 2018. *Fantom: A Crowdsourced Social Chatbot Using an Evolving Dialog Graph*.

Jones, E. E., and Archer, R. L. 1976. "Are There Special Effects of Personalistic Self-Disclosure?," *Journal of Experimental Social Psychology* (12:2), pp. 180–193.

Joosse, M., Lohse, M., Pérez, J. G., and Evers, V. 2013. "What You Do Is Who You Are: The Role of Task Context in Perceived Social Robot Personality," in *2013 IEEE International Conference on Robotics and Automation*, IEEE, pp. 2134–2139.

JuniperResearch. 2018. "Digital Voice Assistants in Use to Triple to 8 Billion by 2023."

Kanawattanachai, P., and Yoo, Y. 2002. "Dynamic Nature of Trust in Virtual Teams," *The Journal of Strategic Information Systems* (11:3–4), pp. 187–213.

Kang, S.-H., and Gratch, J. 2010. "Virtual Humans Elicit Socially Anxious Interactants' Verbal Self-disclosure," *Computer Animation and Virtual Worlds* (21:3-4), pp. 473–482.

Katagiri, Y., Nass, C., and 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*, pp. 1558–1562.

Kee, H. W., and Knox, R. E. 1970. "Conceptual and Methodological Considerations in the Study of Trust and Suspicion," *Journal of Conflict Resolution* (14:3), pp. 357–366.

Kim, Y., and Sundar, S. S. 2012. "Anthropomorphism of Computers: Is It Mindful or Mindless?," *Computers in Human Behavior* (28:1), pp. 241–250.

Kline, R. B. 2015. *Principles and Practice of Structural Equation Modeling*, Guilford publications.

Komiak, S. X., and Benbasat, I. 2004. "Understanding Customer Trust in Agent-Mediated Electronic Commerce, Web-Mediated Electronic Commerce, and Traditional Commerce," *Information Technology and Management* (5:1–2), pp. 181–207.

Komiak, S. Y., and Benbasat, I. 2006. "The Effects of Personalization and Familiarity on Trust and Adoption of Recommendation Agents," *MIS Quarterly*, pp. 941–960.

Leakey, R. E., and Lewin, R. 1978. *People of the Lake: Mankind and Its Beginnings*, Anchor Press Garden City, NJ.

Lee, S., and 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), pp. 95–105.

Lemay Jr, E. P., and Melville, M. C. 2014. "Diminishing Self-Disclosure to Maintain Security in Partners' Care.," *Journal of Personality and Social Psychology* (106:1), American Psychological Association, p. 37.

Lewicki, R. J., and Bunker, B. B. 1995. *Trust in Relationships: A Model of Development and Decline.*, Jossey-Bass.

Lewicki, R. J., McAllister, D. J., and Bies, R. J. 1998. "Trust and Distrust: New Relationships and Realities," *Academy of Management Review* (23:3), pp. 438–458.

Lewis, J. D., and Weigert, A. 1985. "Trust as a Social Reality," *Social Forces* (63:4), pp. 967–985.

Li, X., Zhu, P., Yu, Y., Zhang, J., and Zhang, Z. 2017. "The Effect of Reciprocity Disposition on Giving and Repaying Reciprocity Behavior," *Personality and Individual Differences* (109), pp. 201–206.

Lin, R., and Utz, S. 2017. "Self-Disclosure on SNS: Do Disclosure Intimacy and Narrativity Influence Interpersonal Closeness and Social Attraction?," *Computers in Human Behavior* (70), pp. 426–436.

Litman, L., Robinson, J., and Abberbock, T. 2017. "TurkPrime. Com: A Versatile Crowdsourcing Data Acquisition Platform for the Behavioral Sciences," *Behavior Research Methods* (49:2), pp. 433–442.

Mason, W., and Suri, S. 2012. "Conducting Behavioral Research on Amazon's Mechanical Turk," *Behavior Research Methods* (44:1), pp. 1–23.

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

Mayer, R. C., and 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), pp. 874–888.

McAllister, D. J. 1995. "Affect-and Cognition-Based Trust as Foundations for Interpersonal Cooperation in Organizations," *Academy of Management Journal* (38:1), pp. 24–59.

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

McKnight, D. H., Cummings, L. L., and Chervany, N. L. 1998. "Initial Trust Formation in New Organizational Relationships," *Academy of Management Review* (23:3), pp. 473–490.

Mitchell, A. E., Castellani, A. M., Herrington, R. L., Joseph, J. I., Doss, B. D., and Snyder, D. K. 2008. "Predictors of Intimacy in Couples' Discussions of Relationship Injuries: An Observational Study.," *Journal of Family Psychology* (22:1), p. 21.

Molm, L. D. 1994. "Dependence and Risk: Transforming the Structure of Social Exchange," *Social Psychology Quarterly*, pp. 163–176.

Molm, L. D., Schaefer, D. R., and Collett, J. L. 2007. "The Value of Reciprocity," *Social Psychology Quarterly* (70:2), pp. 199–217.

Molm, L. D., Schaefer, D. R., and Collett, J. L. 2009. "Fragile and Resilient Trust: Risk and Uncertainty in Negotiated and Reciprocal Exchange," *Sociological Theory* (27:1), pp. 1–32.

Monroe, A. E., Dillon, K. D., and Malle, B. F. 2014. "Bringing Free Will down to Earth: People's Psychological Concept of Free Will and Its Role in Moral Judgment," *Consciousness and Cognition* (27), pp. 100–108.

Moon, Y. 2000. "Intimate Exchanges: Using Computers to Elicit Self-Disclosure from Consumers," *Journal of Consumer Research* (26:4), pp. 323–339.

Morgan, R. M., and Hunt, S. D. 1994. "The Commitment-Trust Theory of Relationship Marketing," *Journal of Marketing* (58:3), SAGE Publications Sage CA: Los Angeles, CA, pp. 20–38.

Nass, C., and Moon, Y. 2000. "Machines and Mindlessness: Social Responses to Computers," *Journal of Social Issues* (56:1), pp. 81–103.

Nguyen, M., Bin, Y. S., and Campbell, A. 2012. "Comparing Online and Offline Self-Disclosure: A Systematic Review," *Cyberpsychology, Behavior, and Social Networking* (15:2), pp. 103–111.

Omarzu, J. 2000. "A Disclosure Decision Model: Determining How and When Individuals Will Self-Disclose," *Personality and Social Psychology Review* (4:2), pp. 174–185.

Paolacci, G., Chandler, J., and Ipeirotis, P. G. 2010. "Running Experiments on Amazon Mechanical Turk," *Judgment and Decision Making* (5:5), pp. 411–419.

Pearce, W. B., and Sharp, S. M. 1973. "Self-Disclosing Communication," *Journal of Communication* (23:4), pp. 409–425.

Pickard, M. D., Roster, C. A., and 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), pp. 23–30.

Preacher, K. J., and Hayes, A. F. 2008. "Asymptotic and Resampling Strategies for Assessing and Comparing Indirect Effects in Multiple Mediator Models," *Behavior Research Methods* (40:3), pp. 879–891.

Rempel, J. K., Holmes, J. G., and Zanna, M. P. 1985. "Trust in Close Relationships.," *Journal of Personality and Social Psychology* (49:1), pp. 95–112.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., and Camerer, C. 1998. "Not so Different after All: A Cross-Discipline View of Trust," *Academy of Management Review* (23:3), pp. 393–404.

Rubin, Z. 1975. "Disclosing Oneself to a Stranger: Reciprocity and Its Limits," *Journal of Experimental Social Psychology* (11:3), pp. 233–260.

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

Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J., and 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), pp. 413–422.

Schaubroeck, J., Lam, S. S., and 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), p. 863.

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., and Vogeley, K. 2013. "Toward a Second-Person Neuroscience," *Behavioral and Brain Sciences* (36:4), pp. 393–414.

Schmarzo, B. 2013. *Big Data: Understanding How Data Powers Big Business*, John Wiley & Sons.

Schoorman, F. D., Mayer, R. C., and Davis, J. H. 2007. "An Integrative Model of Organizational Trust: Past, Present, and Future," *Academy of Management Review* (32:2), pp. 344–354.

Schweiger, D., Stemmler, G., Burgdorf, C., and Wacker, J. 2013. "Opioid Receptor Blockade and Warmth-Liking: Effects on Interpersonal Trust and Frontal Asymmetry," *Social Cognitive and Affective Neuroscience* (9:10), pp. 1608–1615.

Shadish, W. R., Cook, T. D., and Campbell, D. T. 2002. "Experiments and Generalized Causal Inference," *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, pp. 1–32.

Shih, H., Lai, K., and 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), pp. 432–450.

Singh, J., and Sirdeshmukh, D. 2000. "Agency and Trust Mechanisms in Consumer Satisfaction and Loyalty Judgments," *Journal of the Academy of Marketing Science* (28:1), pp. 150–167.

Sitkin, S. B., and Roth, N. L. 1993. "Explaining the Limited Effectiveness of Legalistic 'Remedies' for Trust/Distrust," *Organization Science* (4:3), pp. 367–392.

Smith, H. J., Dinev, T., and Xu, H. 2011. "Information Privacy Research: An Interdisciplinary Review," *MIS Quarterly* (35:4), pp. 989–1016.

Sprecher, S., Treger, S., Wondra, J. D., Hilaire, N., and Wallpe, K. 2013. "Taking Turns: Reciprocal Self-Disclosure Promotes Liking in Initial Interactions," *Journal of Experimental Social Psychology* (49:5), pp. 860–866.

Storbeck, J., and Clore, G. L. 2007. "On the Interdependence of Cognition and Emotion," *Cognition and Emotion* (21:6), pp. 1212–1237.

Taddicken, M. 2014. "The 'Privacy Paradox'in the Social Web: The Impact of Privacy Concerns, Individual Characteristics, and the Perceived Social Relevance on Different Forms of Self-Disclosure," *Journal of Computer-Mediated Communication* (19:2), pp. 248–273.

Van de Ven, A. H. 2007. *Engaged Scholarship: A Guide for Organizational and Social Research,* Oxford University Press on Demand.

Voicebot.ai. 2019. "Juniper Estimates 3.25 Billion Voice Assistants Are in Use Today, Google Has About 30% of Them," *Voicebot, ,* February 14.

Wang, W., Qiu, L., Kim, D., and Benbasat, I. 2016. "Effects of Rational and Social Appeals of Online Recommendation Agents on Cognition- and Affect-Based Trust," *Decision Support Systems* (86), pp. 48–60.

Waytz, A., Cacioppo, J., and Epley, N. 2010. "Who Sees Human? The Stability and Importance of Individual Differences in Anthropomorphism," *Perspectives on Psychological Science* (5:3), pp. 219–232.

Waytz, A., Gray, K., Epley, N., and Wegner, D. M. 2010. "Causes and Consequences of Mind Perception," *Trends in Cognitive Sciences* (14:8), pp. 383–388.

Waytz, A., Heafner, J., and Epley, N. 2014. "The Mind in the Machine: Anthropomorphism Increases Trust in an Autonomous Vehicle," *Journal of Experimental Social Psychology* (52), pp. 113–117.

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J.-H., and Cacioppo, J. T. 2010. "Making Sense by Making Sentient: Effectance Motivation Increases Anthropomorphism.," *Journal of Personality and Social Psychology* (99:3), pp. 410–435.

Weber, J. M., Malhotra, D., and Murnighan, J. K. 2004. "NORMAL ACTS OF IRRATIONAL TRUST: MOTIVATED ATTRIBUTIONS AND THE TRUST DEVELOPMENT PROCESS," *Research in Organizational Behavior* (26), pp. 75–101.

Wheeless, L. R., and Grotz, J. 1976. "Conceptualization and Measurement of Reported Self-disclosure," *Human Communication Research* (2:4), pp. 338–346.

Worthy, M., Gary, A. L., and Kahn, G. M. 1969. "Self-Disclosure as an Exchange Process.," *Journal of Personality and Social Psychology* (13:1), p. 59.

Zimmer, J. C., Arsal, R., Al-Marzouq, M., Moore, D., and Grover, V. 2010. “Knowing Your Customers: Using a Reciprocal Relationship to Enhance Voluntary Information Disclosure,” *Decision Support Systems* (48:2), Elsevier, pp. 395–406.

APPENDIX A – Measurements

Table A1. Operationalization of Constructs

Construct	Items	Informing Sources
Anthropomorphism	(1-7 scale) 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, Waytz, et al. (2008) Waytz, Cacioppo, et al. (2010)
Cognition-Based Trustworthiness: Ability	(1-7 scale) 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	(1-7 scale) 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	(1-7 scale) 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)
Trust	(1-7 scale) 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 and Gavin (2005)
Privacy Concern	(1-7 scale) 1. I am concerned that the information I share with a digital assistant could be misused. 2. I am concerned that a person can find private information about me through a digital assistant. 3. I am concerned about sharing information with a digital assistant, because of what others might do with it. 4. I am concerned about sharing information with a digital assistant, because it could be used in a way I did not foresee.	Dinev and Hart (2004, 2006)
Extroversion	1. I talk to a lot of different people at parties. 2. I keep in the background at parties. [R] 3. I am the life of the party. 4. I don't talk a lot at parties. [R]	Donnellan et al. (2006)

Self-Disclosure

Altman and Taylor's (1973) enumerated some properties associated with the intimacy of self-disclosure. Below we categorized them as low, intermediate, and high intimacy.

Low intimacy

- Simple descriptive standpoint
- Biographical characteristics

Intermediate intimacy

- Attitudes
- Opinions

High intimacy

- Fears
- Self-concepts
- Basic values
- Vulnerabilities (real or perceived inadequacies and weaknesses)
- Considered socially undesirable by a reference group
- Can be used to derive less intimate information
- Less visible
- More unique
- Strong affective (positive and negative) characteristics (particularly those that are dispositional)

APPENDIX B – CA's Self-Disclosure Manipulation

Table B1. CA Self-Disclosure Manipulation

	Question	High Disclosure	Number of Words	Low Disclosure	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)					

APPENDIX C – Loadings

Table C1. Loadings		
Construct	**Item**	**Loading**
Age	Age	1
Gender	Gender	1
Education	Education	1
PC	PC1	0.946
	PC2	0.913
	PC3	0.959
	PC4	0.934
Extroversion	Extrovert_1	0.879
	Extrovert_2	-0.823
	Extrovert_3	0.730
	Extrovert_4	-0.885
CA's Self-Disclosure	CA Self-Disclosure	1
Anthropomorphism	Anthropomorphism_1	0.932
	Anthropomorphism_2	0.924
	Anthropomorphism_3	0.946
	Anthropomorphism_4	0.952
	Anthropomorphism_5	0.937
Ability	Ability_1	0.943
	Ability_2	0.967
	Ability_3	0.964
Integrity	Integrity_1	0.926
	Integrity_2	0.935
	Integrity_3	0.824
	Integrity_4	0.829
Affect-Based Trustworthiness	Affect_1	0.920
	Affect_2	0.753
	Affect_3	0.885
Trust	Trust_1	0.931
	Trust_2	0.923
	Trust_3	0.912
User's Self-Disclosure	Depth	0.887
	Breadth	0.602
Note that CFA constraints cross-loadings to zero. Therefore, we only presented the loadings of items on their corresponding constructs.		

APPENDIX D – Robustness Checks

Table D1. Hierarchical Regression for the Mediated Effect of CA Self-Disclosure on User Self-Disclosure

	AP	CT	AT	TIC	UD
Control Variables					
Constant	6.47 (0.87)***	5.03 (0.52)***	-0.89 (0.84)***	1.13 (0.84)	4.05 (1.17)**
Age	-0.04 (0.01)**	0.00 (0.01)	-0.02 (0.01)*	0.00 (0.01)	0.01 (0.01)
Gender	-0.45 (0.28)	-0.14 (0.14)	0.35 (0.18)	0.29 (0.20)	-0.54 (0.27)*
Education	-0.23 (0.11)*	-0.05 (0.06)	0.02 (0.08)	-0.08 (0.08)	-0.20 (0.11)
Previous Experience	0.24 (0.14)	0.05 (0.08)	0.04 (0.08)	-0.03 (0.10)	0.13 (0.13)
Privacy Concerns	0.00 (0.10)	-0.13 (0.05)*	0.09 (0.06)	-0.13 (0.07)	-0.07 (0.09)
Extroversion	0.13 (0.09)	-0.03 (0.05)	0.10 (0.06)	-0.05 (0.08)	-0.03 (0.08)
Independent Variables					
CD	0.53 (0.28)*	-0.08 (0.15)	-0.16 (0.18)	-0.51 (0.20)*	0.86 (0.28)**
AP		0.38 (0.04)***	0.48 (0.07)***	0.15 (0.10)	-0.13 (0.09)
CT			0.28 (0.09)**	0.49 (0.12)***	0.13 (0.15)
AT				0.28 (0.09)**	-0.17 (0.10)
TIC					0.46 (0.09)***
R^2	0.120	0.401	0.516	0.467	0.232

Notes:

a. *Key: CD: CA Self-Disclosure, AP: Anthropomorphism, CT: Cognition-Based Trustworthiness, AT: Affect-Based Trustworthiness, TIC: Trust in CA, UD: User's Self-Disclosure*
b. *N=208*
c. ** p<0.05; ** p<0.01; *** p<0.001; one-tailed tests were used for directional hypotheses and two-tailed tests for the rest of the relationships.*
d. *Unstandardized regression coefficients are shown.*
e. *Numbers in parentheses are the standard errors*

Table D2. Identifying the Individual Effects of Factors via a Hierarchical Regression Analysis for the Mediated Effect of CA Self-Disclosure on User Self-Disclosure

	AP		CT		AT				TIC				UD		
	Block A1	Block A2	Block B1	Block B2	Block C1	Block C2	Block C3	Block C4	Block D1	Block D2	Block D3	Block D4	Block E1	Block E2	Block E3
Control Variables															
Constant	6.79 (0.94)	6.47 (0.95)	7.56 (0.61)	5.00 (0.56)	4.34 (0.82)	0.44 (0.70)	-0.93 (0.94)	-0.96 (0.81)	6.74 (0.88)	0.64 (0.98)	4.24 (0.80)	0.98 (0.92)	7.00 (0.95)	4.78 (1.03)	4.20 (1.03)
Age	-0.04 (0.01)**	-0.04 (0.01)**	-0.01 (0.01)	0.00 (0.01)	-0.04 (0.01)**	-0.02 (0.01)*	-0.03 (0.01)**	-0.07 (0.01)*	-0.02 (0.01)*	-0.01 (0.01)	0.00 (0.01)	0.00 (0.01)	0.01 (0.01)	0.02 (0.01)	0.02 (0.01)
Gender	-0.45 (0.28)	-0.47 (0.28)	-0.31 (0.18)	-0.15 (0.15)	0.05 (0.24)	0.30 (0.19)	0.27 (0.21)	0.34 (0.18)	0.09 (0.26)	0.34 (0.22)	0.06 (0.22)	0.24 (0.21)	-0.49 (0.28)	-0.52 (0.27)	-0.52 (0.27)*
Education	-0.23 (0.11)*	-0.23 (0.11)*	-0.14 (0.07*	-0.06 (0.06)	-0.12 (0.10)	0.01 (0.07)	-0.03 (0.08)	0.02 (0.07)	-0.21 (0.10)*	-0.10 (0.09)	-0.14 (0.09)	-0.09 (0.08)	-0.27 (0.11)*	-0.20 (0.11)	-0.19 (0.11)
Previous Experience	0.24 (0.13)	0.24 (0.12)	0.14 (0.08)	0.05 (0.07)	0.20 (0.11)	0.05 (0.09)	0.09 (0.10)	0.04 (0.08)	0.13 (0.12)	0.02 (0.10)	0.02 (0.10)	-0.02 (0.10)	0.13 (0.13)	0.09 (0.12)	0.09 (0.12)
Privacy Concerns	-0.02 (0.08)	0.00 (0.08)	-0.14 (0.05)*	-0.13 (0.04)**	0.05 (0.07)	0.06 (0.06)	0.15 (0.06)*	0.10 (0.06)	-0.17 (0.08)*	-0.06 (0.07)	-0.20 (0.07)**	-0.11 (0.06)	-0.20 (0.09)*	-0.15 (0.08)	-0.12 (0.08)
Extroversion	0.14 (0.08)	0.13 (0.08)	0.03 (0.06)	-0.03 (0.05)	0.18 (0.07)*	0.09 (0.06)	0.16 (0.06)*	0.10 (0.06)	0.02 (0.08)	0.00 (0.07)	-0.08 (0.07)	-0.06 (0.06)	-0.05 (0.09)	-0.05 (0.08)	-0.07 (0.08)
Independent Variables															
CD		0.53 (0.27)*													0.74 (0.26)**
AP				0.38 (0.04)**		0.58 (0.05)**		0.47 (0.06)**							
CT							0.70 (0.08)**	0.28 (0.09)**		0.81 (0.09)**		0.55 (0.09)**			
AT											0.58 (0.06)**	0.36 (0.07)**			
TIC														0.33 (0.07)**	0.35 (0.07)**
R^2	0.10	0.12	0.10	0.40	0.10	0.49	0.34	0.52	0.08	0.37	0.34	0.44	0.08	0.17	0.20
Adjusted R^2	0.08	0.09	0.07	0.38	0.08	0.47	0.32	0.50	0.05	0.34	0.32	0.42	0.05	0.14	16.8
ΔR^2		0.02*		0.30**		0.39**	0.24**	0.42**		0.29**	0.26**	0.36**		0.09**	0.12**

Notes:

a. Key: CD: CA Self-Disclosure, AP: Anthropomorphism, CT: Cognition-Based Trustworthiness, AT: Affect-Based Trustworthiness, TIC: Trust in CA, UD: User's Self-Disclosure

b. N=208

*c. * p<0.05; ** p<0.01; one-tailed tests were used for directional hypotheses and two-tailed tests for the rest of the relationships.*

d. Unstandardized regression coefficients are shown.

e. Numbers in parentheses are the standard errors

f. All values for ΔR^2 were calculated compared to the models with only control variables.

Chapter 5: Conclusion

According to the industry reports, the presence of AI agents in our daily lives will only increase in the foreseeable future (Columbus 2018). AI agents are used in a variety of forms such as personal assistants (e.g., Amazon's Alexa, Apple Siri, Google Assistant, Microsoft Cortana), business assistants (e.g., Alexa for Business), smart doctors (e.g., Ada the AI doctor), and personal companions (e.g., Replika: My AI Friend). But users cannot control every small detail in the behavior of AI agents. So as long as people depend on AI agents to fulfill their tasks, the concepts of trust and distrust remain relevant in human-AI interaction.

In this dissertation, we drew on theoretical approaches in psychology, neuroscience, communication, artificial intelligence, and information systems to better understand why people trust and distrust AI agents. We not only studied people's trusting and distrusting beliefs and intentions, but also examined their actual behavior in real interactions with such agents.

To do so, we developed an AI agent named Amanda. We used our agent to conduct the discussed experiments in chapters 2 and 4. We designed Amanda as a platform for conducting a range of studies on human-AI interaction. This platform includes a server-side dashboard (Figure 1) in which we designed each of the experiments, a JavaScript client with which we conducted our web-based experiments, and a voice-enabled Android app with which we conducted our mobile experiments. We open-sourced all the code developed for this dissertation (more than 20,000 original lines of code) on GitHub (https://github.com/saffarizadeh/) to encourage more research on conversational agents.

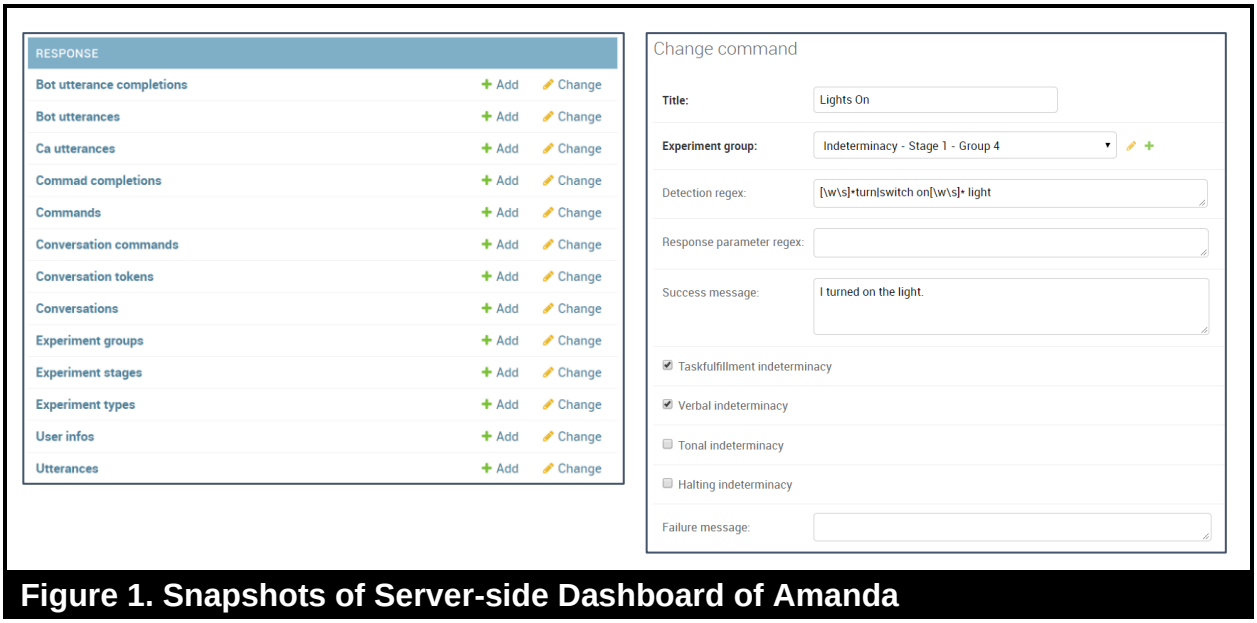


Figure 1. Snapshots of Server-side Dashboard of Amanda

We employed randomized experiments as our identification strategy in all essays. Randomized experiments, as the gold standard of internal validity, are especially useful for establishing cause-effect relationships. We used a range of analysis methods, including structural equation modeling, regression, and ANOVA, to evaluate the results of our experiments.

Without reiterating each essay's theoretical and practical contributions, which were discussed in detail in their corresponding chapters, here, we discuss how the cumulative findings provide a clearer picture of human-AI interaction. Figure 2 provides a summary of our findings in the form of a conceptual framework.

A compilation of the findings in the three essays shows that: (1) users anthropomorphize AI agents in order to make sense of AI-specific *behaviors*, (2) users' perceptions of AI-specific *characteristics* are shaped based on both the AI and its creator, (3) perceived AI characteristics, AI behaviors, and anthropomorphism provide cognitive and affective evidence that drives trust and distrust, and (4) trust and distrust in AI drive users' risk-taking behaviors such as disclosure and delegation to AI.

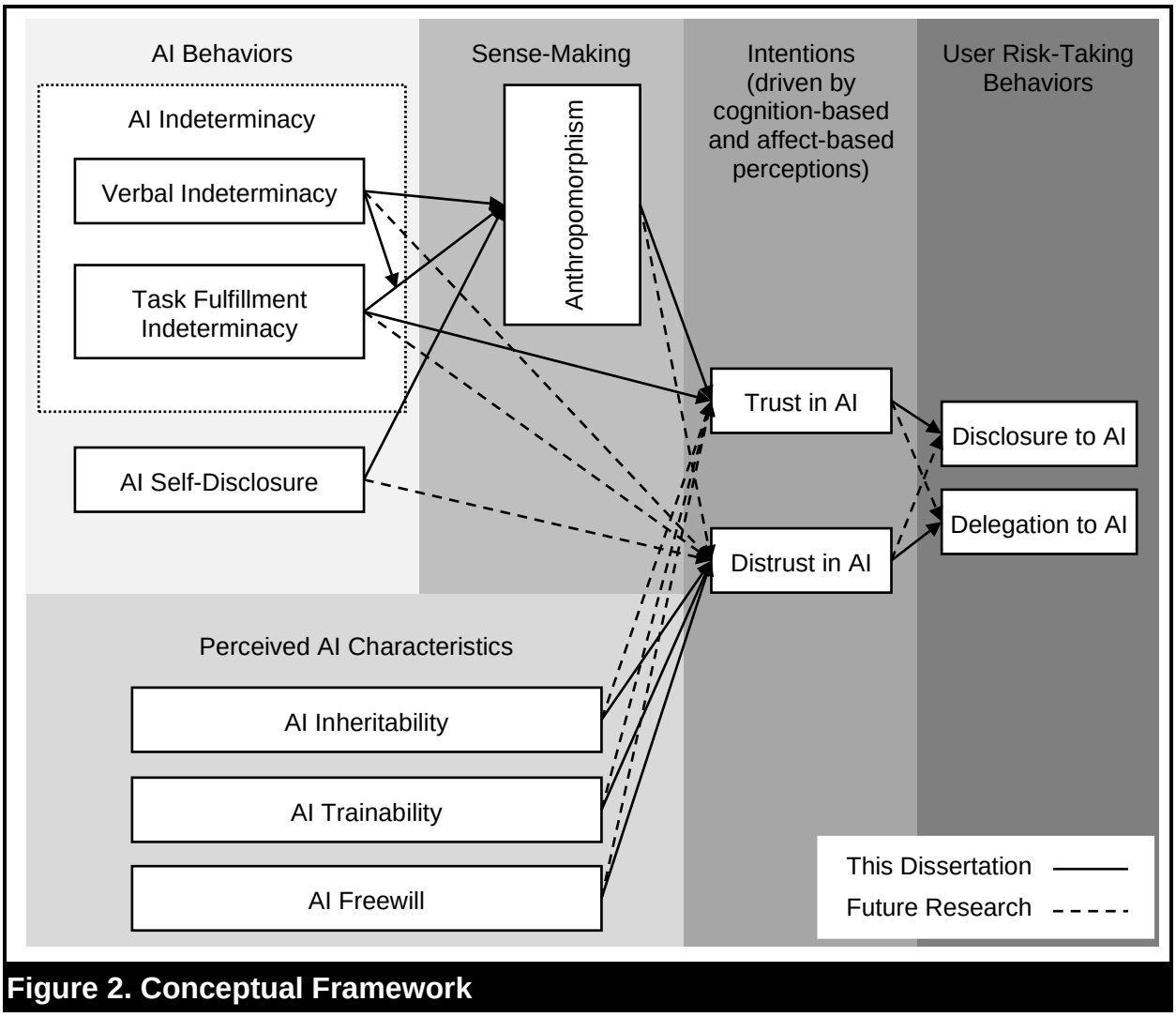


Figure 2. Conceptual Framework

This conceptual framework can guide future research on human-AI interaction. First, by indicating several unexplored relationships that can explain trust and distrust in AI, it provides a roadmap for future research on the antecedents of trust and distrust in the context of AI agents (see dashed lines in Figure 2). Second, by enumerating behaviors and perceived characteristics of AI, it sets the foundations for future research to study the interplay of the two and to examine how AI behaviors can be manipulated to change users' perceptions of AI-specific characteristics. For instance, future research can examine whether it is possible to mitigate the negative direct effect of task fulfillment indeterminacy on trust by changing the level of AI trainability, or whether it is possible to intensify

the positive indirect of task fulfillment indeterminacy on trust (via anthropomorphism) by increasing AI trainability and decreasing AI inheritability. Finally, by including sense-making as an essential step between AI behavior and trust and distrust in AI, our conceptual framework emphasizes the important role of mechanisms like anthropomorphism. While previous research in information systems used anthropomorphism to study anthropomorphic features such as whether the AI had a human-like face, voice, or body, our framework in line with research in psychology (e.g., Epley et al. 2007; Schroeder et al. 2017; Waytz et al. 2010) highlights the more general role of anthropomorphism as a sense-making mechanism.

In this dissertation, we leveraged several theories, concepts, and analytical techniques to study the phenomenon of users' trust and distrust of AI agents. While shedding light on this phenomenon, as our conceptual framework shows, we advance a number of avenues for future research. We hope that the findings of this work provide a foundation for other scholars to conduct theory-driven research that will advance our understanding of human-AI interaction.

References

Columbus, L. 2018. "10 Charts That Will Change Your Perspective On Artificial Intelligence's Growth," *Forbes*. (https://www.forbes.com/sites/louiscolumbus/2018/01/12/10-charts-that-will-change-your-perspective-on-artificial-intelligences-growth/, accessed February 16, 2019).

Epley, N., Waytz, A., and Cacioppo, J. T. 2007. "On Seeing Human: A Three-Factor Theory of Anthropomorphism.," *Psychological Review* (114:4), pp. 864–886. (doi: 10.1037/0033-295X.114.4.864).

Schroeder, J., Kardas, M., and Epley, N. 2017. "The Humanizing Voice: Speech Reveals, and Text Conceals, a More Thoughtful Mind in the Midst of Disagreement," *Psychological Science* (28:12), pp. 1745–1762.

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J.-H., and Cacioppo, J. T. 2010. "Making Sense by Making Sentient: Effectance Motivation Increases Anthropomorphism.," *Journal of Personality and Social Psychology* (99:3), pp. 410–435. (doi: 10.1037/a0020240).