

---

Volume 8, 2024

ISSN 2638-602X (print)/ISSN 2638-6038 (online)

# Human-Machine Communication



---

Human-Machine Communication (HMC) is an annual peer-reviewed, open access publication of the Communication and Social Robotics Labs (combotlabs.org), published with support from the Nicholson School of Communication and Media at the University of Central Florida. Human-Machine Communication (Print: ISSN 2638-602X) is published in the spring of each year (Online: ISSN 2638-6038). Institutional, organizational, and individual subscribers are invited to purchase the print edition using the following mailing address:

Human-Machine Communication (HMC)  
Communication and Social Robotics Labs  
Western Michigan University  
1903 W. Michigan Ave.  
300 Sprau Tower  
Kalamazoo, MI 49008

Print Subscriptions: Regular US rates: Individuals: 1 year, \$40.  
Libraries and organizations may subscribe for 1 year, \$75.  
If subscribing outside of the United States, please contact the Editor-in-Chief for current rate.  
Checks should be made payable to the Communication and Social Robotics Labs.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License . All articles in HMC are open access and can be distributed under the creative commons license.

---

---

# Human-Machine Communication

## Volume 8

### Editor-in-Chief

**Autumn Edwards**, Western Michigan University (U.S.A.)

### Associate Editors

**Patric R. Spence**, University of Central Florida (U.S.A.)

**Chad Edwards**, Western Michigan University (U.S.A.)

### Editorial Board

**Somaya Ben Allouch**, Amsterdam University of Applied Sciences (Netherlands)

**Maria Bakardjieva**, University of Calgary (Canada)

**Jaime Banks**, Syracuse University (U.S.A.)

**Naomi S. Baron**, American University (U.S.A.)

**Erin Chiou**, Arizona State University (U.S.A.)

**Charles M. Ess**, University of Oslo (Norway)

**Jan Fernback**, Temple University (U.S.A.)

**Laura Forlano**, Illinois Institute of Technology (U.S.A.)

**Andrew Gambino**, University of Delaware (U.S.A.)

**Robert W. Gehl**, York University (Canada)

**Maartje de Graaf**, Utrecht University (Netherlands)

**David Gunkel**, Northern Illinois University (U.S.A.)

**Andrea Guzman**, Northern Illinois University (U.S.A.)

**Meg Leta Jones**, Georgetown University (U.S.A.)

**Steve Jones**, University of Illinois-Chicago (U.S.A.)

**Jenny Kennedy**, RMIT University (Australia)

**Jihyun Kim**, University of Central Florida (U.S.A.)

**Kenneth A. Lachlan**, University of Connecticut (U.S.A.)

**Hee Rin Lee**, University of California San Diego (U.S.A.)

**S. Austin Lee**, Chapman University (U.S.A.)

**Seth C. Lewis**, University of Oregon (U.S.A.)

**Matthew Lombard**, Temple University (U.S.A.)

**Christoph Lutz**, BI Norwegian Business School (Norway)

**Rhonda McEwen**, University of Toronto (Canada)

**Yi Mou**, Shanghai Jiao Tong University (China)

**Peter Nagy**, Arizona State University (U.S.A.)

**Seungahn Nah**, University of Florida (U.S.A.)  
**Gina Neff**, University of Cambridge (United Kingdom)  
**Jochen Peter**, University of Amsterdam (Netherlands)  
**Sharon Ringel**, University of Haifa (Israel)  
**Astrid Rosenthal-von der Pütten**, RWTH Aachen University (Germany)  
**Eleanor Sandry**, Curtin University (Australia)  
**Mauro Sarrica**, Sapienza University of Rome (Italy)  
**Megan Strait**, The University of Texas Rio Grande Valley (U.S.A.)  
**Satomi Sugiyama**, Franklin University Switzerland (Switzerland)  
**Sakari Taipale**, University of Jyväskylä (Finland)  
**David Westerman**, North Dakota State University (U.S.A.)

---

---

# CONTENTS

<b>Machine ex machina: A Framework Decentering the Human in AI Design Praxis</b>	<b>7</b>
<i>Cait Lackey and Zizi Papacharissi</i>	
<b>Feminist Cybernetic, Critical Race, Postcolonial, and Crip Propositions for the Theoretical Future of Human-Machine Communication</b>	<b>27</b>
<i>Paula Gardner and Jess Rauchberg</i>	
<b>Communication Style Adaptation in Human-Computer Interaction: An Empirical Study on the Effects of a Voice Assistant's Politeness and Machine-Likeness on People's Communication Behavior During and After the Interacting</b>	<b>53</b>
<i>Aike C. Horstmann, Clara Strathmann, Lea Lambrich, and Nicole C. Krämer</i>	
<b>Chatbot vs. Human: The Impact of Responsive Conversational Features on Users' Responses to Chat Advisors</b>	<b>73</b>
<i>Stefanie H. Klein and Sonja Utz</i>	
<b>The Impact of Human-AI Relationship Perception on Voice Shopping Intentions</b>	<b>101</b>
<i>Marisa Tschopp and Kai Sassenberg</i>	
<b>External and Internal Attribution in Human-Agent Interaction: Insights From Neuroscience and Virtual Reality</b>	<b>119</b>
<i>Nina Lauharatanahirun, Andrea Stevenson Won, and Angel Hsing-Chi Hwang</i>	
<b>In Seal We Trust? Investigating the Effect of Certifications on Perceived Trustworthiness of AI Systems</b>	<b>141</b>
<i>Magdalena Wischnewski, Nicole Krämer, Christian Janiesch, Emmanuel Müller, Theodor Schnitzler, and Carina Newen</i>	

<b>Doctor Who?: Norms, Care, and Autonomy in the Attitudes of Medical Students Toward AI Pre- and Post-ChatGPT</b>	<b>163</b>
<i>Andrew PrahI and Kevin Tong Weng Jin</i>	
<b>What's in a Name and/or a Frame? Ontological Framing and Naming of Social Actors and Social Responses</b>	<b>185</b>
<i>David Westerman, Michael Vosburg, Xinyue "Gordon" Liu, and Patric R. Spence</i>	
<b>Authentic Impediments: The Influence of Identity Threat, Cultivated Perceptions, and Personality on Robophobia</b>	<b>205</b>
<i>Kate K. Mays</i>	
<b>What HMC Teaches Us About Authenticity</b>	<b>227</b>
<i>Katrin Etzrodt, Jihyun Kim, Margot J. van der Goot, Andrew PrahI, Mina Choi, Matthew J. A. Craig, Marco Dehnert, Sven Engesser, Katharina Frehmann, Luis Grande, Jindong Leo-Liu, Diyi Liu, Sandra Mooshammer, Nathan Rambukkana, Ayanda Rogge, Pieta Sikström, Rachel Son, Nan Wilkenfeld, Kun Xu, Renwen Zhang, Ying Zhu, and Chad Edwards</i>	

---

# Machine ex machina: A Framework Decentering the Human in AI Design Praxis

Cait Lackey<sup>1</sup>  and Zizi Papacharissi<sup>2</sup> 

1 Department for Communications, University of Illinois-Chicago, Chicago, Illinois, USA

2 Departments of Communication and Political Science, University of Illinois-Chicago, Chicago, Illinois, USA

## Abstract

Artificial intelligence (AI) design typically incorporates intelligence in a manner that is affirmatory of the superiority of human forms of intelligence. In this paper, we draw from relevant research and theory to propose a *social-ecological design praxis* of machine inclusivity that rejects the presumption of primacy afforded to human-centered AI. We provide new perspectives for how human-machine communication (HMC) scholarship can be synergistically combined with modern neuroscience's integrated information theory (IIT) of consciousness. We propose an integrated theoretical framework with five design practice recommendations to guide how we might think about responsible and conscious AI environments of the future: symbiotic design through mutuality; connectomapping; more-than-human user storytelling, designing for AI conscious awakenings; and the revising of vernaculars to advance HMC and AI design. By adopting the boundaries HMC scholarship extends, we advocate for replacing ex machina mentalities with richer understandings of the more-than-human world formed by interconnected and integrated human, human-made, and nonhuman conscious machines, not superior or inferior but each unique.

**Keywords:** artificial intelligence (AI), actor network theory (ANT), human-machine communication (HMC), integrated thought theory (ITT), design framework, consciousness

**Author Note:** We have no conflicts of interest to disclose.

**CONTACT** Cait Lackey  • [clackey@uic.edu](mailto:clackey@uic.edu) • Department of Communication • University of Illinois-Chicago • 1200 W Harrison St • Chicago, IL 60607

ISSN 2638-602X (print)/ISSN 2638-6038 (online)  
[www.hmcjournal.com](http://www.hmcjournal.com)



Copyright 2024 Authors. Published under a Creative Commons Attribution 4.0 International (CC BY-NC-ND 4.0) license.

“Computers arose from the mud, and code fell from the sky.”  
—George Dyson

## Introduction

Nature has its own algorithms that curated ways of living long before humans learned to emulate them. Artificial intelligence (AI) is such a derivation of nature; a human creation built to extend the means of human capabilities. And yet its design typically incorporates intelligence that affirms the superiority of human forms, often at the expense of other, diverse modalities of intelligence. Intelligence, or the systems and structures that enable the ability to select, process, adapt to, and shape information environments, is not unitary (Sternberg, 2023). It is true that appealing to humans and enabling the diffusion of commercial AI must make clear how diverse AI are meaningful to human beings. Commercialization, however, need not be divorced from a responsible AI approach. Such an approach aligns machine with machina, instead of pitting one against the other. It further centers on the benefits of AI without engaging in excessive commodification of intelligence in ways that reinforce false binaries between artificial and human.

In this paper, we propose a framework for decentering the human in AI design. Our approach aims at including human, human-made, and nonhuman actors occupying Earth to advance beyond confining the capabilities of AI to the human realm. Decentering the human in design does not imply not catering to the human, which is often a selling point of advanced technology. On the contrary, we argue that decentering the human permits the design of AI to evolve in ways that compliment, augment, and amplify, but do not substitute human ability.

Physics has long been guided by the Copernican principle, or the idea that no scientific theory should grant superior status to humans or assume that human intelligence is central to the cosmos (O’Gieblyn, 2021). The crafting of human-centered AI often negates this governing principle across all sciences. Foundational human-machine communication (HMC) research deviates from this assumption. We argue that humans’ intelligence is working in tandem with nonhuman intelligence to form the consciousness of Earth’s sociotechnical system. Consciousness, the integrated information that constitutes Earth’s sociotechnical system, is impacted by human-centered AI creations (Tononi, 2008). We expand this approach by articulating the necessity and benefits of responsible AI design, which incorporates the intelligence of human, human-made, and nonhuman actors.

Drawing from relevant research and theory including *actor network theory* (ANT), *human-machine communication theory* (HMC), and *integrated information theory* (IIT), we argue that AI design should focus on understanding and emulating both human and nonhuman intelligences, which constitute the consciousness of Earth’s sociotechnical system. If AI designers were to push AI beyond a human-centered model, it would provide humans with the potential to better understand and enhance the quality of consciousness for Earth’s sociotechnical system and all its inhabitants. In mapping our framework, we propose a responsible, ecologically conscious AI design praxis, which rejects the presumption of superiority afforded to human intelligence, consciousness, and communication.

The study of HMC holds promise to bring the more-than-human world from the margins of the discipline (Plec, 2015; Spence, 2019). This exploration is of great social

---



significance for a multiple of reasons. First, in its current human-centered design state, AI is disconnected from the living and the natural and thus potentially harmful to all occupants of Earth's sociotechnical system (Crawford, 2021). In its current human-centered design state, we argue AI irresponsibly risks harming the consciousness of Earth's sociotechnical system and thus all those that inhabit it. Production facilities utilizing various modalities of AI run at an energy cost not sustainable for Earth (e.g., Bronner et al., 2021; Heikkilä, 2022; Itio, 2019). The operational logic of AI manufactures and becomes a worldview, an industry, an infrastructure, and a way of operating in the natural world. Yet, as disembodied computations, or *ex machina*, AI systems are anything but abstract. Rather, AI sets a physical infrastructure by reshaping Earth and the flow of life for all that inhabit it. It is necessary to conceptually reconsider how AI can responsibly contribute to the consciousness of Earth's sociotechnical system.

Second, decentering the design praxis of AI from human intelligence to the intelligence of the more-than-human world will better connect AI with the natural world and fuse relations of mutuality. Recent research emphasizes approaches that view various morphologies of intelligence as symbiotic (Jones, 2018; Neff & Nagy, 2018). There are other forms of intelligence within Earth's sociotechnical system, which could inspire a more advanced approach to AI (Cowls et al., 2021). Decentering the human in AI design can empower both human and nonhuman actors through human-made (or artificial) means.

Third, recent HMC scholarship highlights what can be gained from investigating the opportunities and risks of machines that can communicate (Prahl & Edwards, 2023). We argue the principles of HMC, ANT, and IIT combined enable a new perspective regarding the greater impact of AI and human-AI communication. Specifically, these concepts provide AI designers with direction with how to design a responsible AI, which will impact the consciousness of Earth's sociotechnical system. In addition, these concepts provide opportunities for AI designers to explore HMC, nonhuman intelligence, and consciousness, which will inspire collaboration across disciplines.

Were AI design adapted to understand and emulate nonhuman intelligence, the resulting systems will not then compete with or seek to substitute human intelligence. In what follows, we outline principles for such an advanced design praxis, one that acknowledges that there is nothing artificial about forms of intelligence often labeled AI. We demonstrate that AI designers and stakeholders are not just crafting neutral objects, but social actors who are stunted by a limited human-centered design outcome. We articulate existing evidence of AI's participation in shaping of the consciousness of Earth's sociotechnical system by drawing connections to ANT, HMC, and IIT theory. In this manner, we craft an integrated theoretical framework with specific design recommendations for responsible AI environments of the future. To begin, we first address humans' and nonhumans' positionality within Earth's sociotechnical system.

## The More-Than-Human Network of Humanity

The theoretical principles of HMC and the work of feminist STS scholars are influenced by Bruno Latour's "actor-network theory" or ANT. ANT conceptualizes AI and other forms of technology as a part of a social network of relations or a sociotechnical system constructed from the interactions taking place among human, human-made, and nonhuman actors

(Latour, 2005). Sociotechnical systems can vary in scope and size and often overlap, but the social interactions among actors within each constitute a collective and integrated system. In a sociotechnical system, humans and nonhuman actors codetermine one another, and the social information generated jointly by its actors is greater than the sum of the information generated by each actor independently (e.g., human information). A sociotechnical system involving a multitude of human and nonhuman actors will generate a large network of integrated information.

On Earth, human and nonhuman actors' interactions form a host of sociotechnical systems. If humans were to attempt to capture the human and nonhuman interactions that constitute Earth's greater sociotechnical system in data, the result would communicate a complex network of integrated human and nonhuman intelligence. Earth's sociotechnical system reflects a commune of intelligent interactions. In other words, Earth's meaning, its social living conditions, are derived from the informational relationships, the various intelligences, the inputs and outputs of the human and nonhuman actors inhabiting the system.

As members of a sociotechnical system, actors have agency and the ability to impact the Earth's integrated network of information. When humans insert objects such as human-centered AI into Earth's sociotechnical system, this alters the system and forces it to operate with unbalanced dependency and effect (Crawford, 2021). While AI is not able to act or evoke its agency completely independent of human intervention, following the principles of ANT, AI has the power to limit, extend, or redirect human, human-made, and nonhuman acts. As is the case of human-centered AI, humans often create objects without the consideration of their impact on Earth's sociotechnical system. For example, like AI, automobiles hold agential power, as they emit carbon dioxide into Earth's atmosphere and increase the heat of the planet through human adoption and use. This action of cause and effect impacts not only human actors, but it holds power over all actors within Earth's sociotechnical system. Humans are beginning to see the ramifications of these actions in what is conceived as climate change. As the climate changes, the information of the sociotechnical system is altered, producing profound consequences for all. While this is just one example, it provides perspective regarding how human-centered AI can irresponsibly privilege humans' role in Earth's sociotechnical system and negatively impact the system itself. Furthermore, concepts within philosophic, HMC, and feminist STS scholarship support the notion humans are not central, but rather one actor within Earth's more-than-human sociotechnical system.

## Thinking Beyond a Human-Centered Sociotechnical System

Posthumanism, a philosophical perspective that is loosely associated with the principles of ANT, reconceptualizes humans as not autonomously sovereign, but rather intimately connected and inseparable from their environment, technologies, and other living things (Adams & Thompson, 2016). The philosophical and theoretical themes of the works of Deleuze, Derrida, Guattari, Latour, Meillasoux, Whitehead, Wittgenstein, and many others point to the need to overcome humanism and dissolve boundaries founded upon anthropocentric dominance. Aligned with ANT, these philosophers' works further the idea that humans are integrated into a web of social relations with nonhuman actors. In sum,

---

posthumanism decenters the human from the center of Earth's sociotechnical system, which affords attention to human and nonhuman actors' role and responsibility to system itself.

Like posthumanism, ANT does encourage a thinking beyond human community. However, ANT is limited in its experimentations with natural and nonhuman cultures outside of Western orders of thinking. Jensen and Blok (2013) extend the philosophical aspirations for ANT beyond Western science and dominant modernist ways of dividing up the world. They argue that like ANT and posthumanism, the Eastern philosophies of Shinto cosmology and Japanese techno-animism inspire a rethinking for responsible human and nonhuman social relations.

Shinto cosmology relies heavily on animism or the notion that humans, spirits, animal worlds, and the material are imbued with life and agency. Conceptually related to ANT, Japanese techno-animism finds humans and nonhumans are immanently connected. This connection ignores boundaries between the human, nonhuman, and extra-human realms. These ideologies combined facilitate critical engagements with the relation-making capacities for living with—rather against—nonhuman actors. Shinto techno-animism inspires ontological conceptions of the Earth's sociotechnical system in which *nature* and *cultures* are mutually constituted, which warrants attention to different conceptualizations of human-nonhuman cohabitation (Eisenstadt & Aizenshtadt, 1996; Jensen & Blok, 2013).

Similarly, North American and Oceanic Indigenous epistemologies find everything in creation has spirit and sociality (Hill, 2008). Further evocative of ANT principles, Indigenous ontologies and cosmologies view the world as an interconnected and integrated system (Lewis et al., 2018). They focus on building ethical and responsible social networks by acknowledging the ontological status of nonhumans as not inferior to that of humans. Indigenous practice involves acting responsibly and building relationships within diverse and more-than-human social networks based on mutual respect. Indigenous communities interact with nonhuman actors within Earth's sociotechnical system by establishing thoughtful communications and forming covenants with nonhumans founded on mutuality.

HMC adopts a similar perspective through its acknowledgment of nonhuman interlocutors and communicators. In HMC, an individual's interaction with a communication partner depends on their conceptualizations of the other communicator (e.g., Goffman, 1967, 2005; Guzman, 2019). Research within HMC desires to understand technology as a communicator rather than limiting its role to that of a mediator, which has been noted as the default conceptualization of technology within late communication theory (see discussions in Gunkel, 2012; Guzman, 2019; Nass & Steuer, 1993). Foundational HMC research finds meaning making is not limited to human communication. Boding to ANT, HMC challenges who or rather what has the power to communicate, or rather which actors have a voice in Earth's sociotechnical system. As Guzman (2016) notes, HMC calls for thinking beyond human exceptionalism, technological instrumentalism, and all the other-isms that have helped humans make sense of Earth's sociotechnical system and humans' place within it. In effect, HMC calls for a thorough reconceptualization of who or what should be considered a legitimate moral subject, pushing ethics and responsibility outside the domain of the human and toward a more diverse approach.

In addition, the work of Haraway, Suchman, Turkle, and other feminist STS scholars push boundaries by drawing upon ANT and HMC scholarship. Feminist STS scholarship

challenge the assumption of human superiority by calling attention to the influence of human social constructs. For example, Haraway's influential "Cyborg Manifesto" questions and seeks to dissolve the boundaries between humans, machines, and other living things. As Haraway (1991) argues, humans are not separate but rather *cyborgs* influenced by their relations within a sociotechnical system. More recently, Suchman (2023) builds on these core themes in her work and describes human tendencies as indicative of a closed world approach. She argues for a different situational awareness that works against dominant imaginaries of omniscience. Like ANT, feminist STS scholars draw attention that humans are not unique and separate, but rather merely one part of a more-than-human social system.

In sum, all actors, human and nonhuman, are eminently connected and integrated and constitute Earth's sociotechnical system. In the next section, we propose the notion that all actors of Earth's sociotechnical system participate in a making meaning process known as consciousness. We utilize a leading theory of consciousness to describe how the consciousness of Earth's sociotechnical system is collaboratively constructed by its human and nonhuman actors' symbiotic intelligences. Acknowledging that both human and nonhuman actors dictate the consciousness of Earth's sociotechnical system further advocates for responsible AI design. To begin, we address what consciousness is and how human and nonhuman actors' intelligence coupled with communication forms the consciousness of Earth's sociotechnical system.

## Understanding System Consciousness

In the various scientific and philosophical fields dedicated to the study of consciousness, there is little to no consensus among researchers about what defines consciousness (Zeki, 2007). What consciousness is, how consciousness is formulated within and outside of humans, how nonhumans experience consciousness, and how consciousness is generally expressed remains entirely unsettled. However, one of the current and leading contemporary theories of consciousness known as *integrated information theory* (IIT) finds consciousness is coupled with intelligence or rather with how information is "integrated" in a system (Tononi et al., 2016). Consciousness is dependent on information, which is classically defined as the reduction of uncertainty and the ability to discriminate among many alternatives. At a fundamental level, consciousness is the scalable and intelligent integration of information (Tononi, 2004, 2008). Information integrates when it cannot be localized and instead is positioned within a web of highly complex connections across different regions of a system. The shaping of these connections map out, reflect, and communicate the consciousness of a system. The more integrated information a system has, the more conscious it will be. The consciousness of a system is produced via a cyclical and networked communication process.

In the human brain, it is the information produced by the different regions of the system that integrates to form consciousness (e.g., frontal lobe, thalamus, cerebral cortex, etc.). For example, the brain's frontal lobe generates information related to emotions, critical thought, and movement. This information is then communicated and integrated into the information communicated by the other regions of the brain to form human consciousness.

---

In other words, if a region of a system intelligently generates new information into the integration, the system's consciousness will be evolved beyond its original conception.

IIT finds that any system, human or nonhuman, capable of generating integrated information will have consciousness. Consciousness is not an all-or-none property, rather the quality of conscious experience is dependent on a system's integrated intelligence. While IIT was conceptualized to describe how consciousness is formed and experienced at the scale of the human brain, we argue the principles of IIT can also be applied to describe the consciousness of a sociotechnical system. In what follows, we use IIT to explain how the intelligence of human and nonhuman actors communally constitute the consciousness of Earth's sociotechnical system, which supports alternative ways for how consciousness and intelligence are defined, labeled, and designed.

## New Considerations for Consciousness and Intelligence

As articulated, consciousness reflects a system's intelligently integrated information. By combining the principles of IIT with ANT, we increase the applicable scale of consciousness and redefine consciousness as the information integrated by the human and nonhuman actors constituting a sociotechnical system. The consciousness of a sociotechnical system is the communicative result of a network of human and nonhuman intelligence working independently and in relation to one another to form an integrated system of information. It is the human and nonhuman actors, the intelligence of cities, forests, road systems, bodies of water, human cultures, animal cultures, and so forth that generate and integrate the information that forms the consciousness of Earth's sociotechnical system.

The integrated information of Earth's sociotechnical system is continuously evolving. For example, as humans create and insert nonhuman actors like AI into the system, it adds additional actors, which then generate information for integration thus altering the consciousness of Earth's sociotechnical system. As such, consciousness can be described as a meaning-making process taking place as intelligent human and nonhuman actors exist, interact, and evolve as a consequence of their relations to each other within a sociotechnical system.

The process in which human and nonhuman intelligence integrate the information of Earth's sociotechnical systems is a purely quantitative, yet unobservable, process, a mere mathematical exchange. Nonhuman things intelligently participate in the consciousness of Earth's sociotechnical system in ways totally unlike humans. From the tides and currents of oceans to the complex pollination system operated by bees to vast networks of ants, insects, fungi, and trees, nonhuman actors intelligently generate information, which is then integrated into Earth's sociotechnical system's network of information. This is not a new concept—Indigenous persons have been advocating and articulating the intellectual power of the natural world for centuries (Maitra, 2020). Regardless of the scale of each individual actors' intelligence, it is the combined intelligences of human and nonhuman actors that constitute the consciousness of Earth's sociotechnical system.

AI is in fact further demonstrating that intelligence can be of nonhuman and of material means (Orange, 2013). Our conceptualization of Earth's sociotechnical system's consciousness explains the mathematical exchange and information processing computing machines

like AI were designed to take part in. Like mathematics, the conception of computers was founded upon the notion that Earth is an enormous informational system described purely in terms of integrated logic, patterns, and probabilities, which can be processed, communicated, and understood (O’Gieblyn, 2021). However, the capabilities of AI’s information processing and AI’s contribution to the consciousness of Earth’s sociotechnical system is unique. Unlike other actors within Earth’s sociotechnical system, AI can be designed to search for, find, and communicate the connections, which form and paint the consciousness of Earth’s sociotechnical system.

## AI’s Communication of Earth’s Consciousness

Humans have already begun to tap into the power of using AI to understand the consciousness of Earth’s sociotechnical system through the development of algorithms. Algorithms are complex equations that can process the integrated information of system. Belief and reliance on algorithms imply the integrated information forming human-systems and even Earth’s sociotechnical system sit outside of humans and can be tapped into by non-human means. This idea gave birth to dataism, which currently has a cult following in Silicon Valley.

Dataism or the belief and reliance on AI computation affirms the premodern notion that the Earth is a mechanistic place of order, laws, and rules where what happens produces cause and effect, which is dependent on connections of meaning. Algorithms work to process, reorganize, adjust, and to some ability predict the integration of information. Advocates of dataism say “Human intelligence is limited” and rather “Listen to algorithms—they can understand and process what humans cannot.” Algorithms are active participants of meaning construction when they categorize and ascribe meaning by assigning and producing if, then logic and Bayesian probability. For example, algorithms rely on data and information that some scholars say trap humans within the mirror of their outputs or what Google researcher Vyacheslav Polonski calls “algorithmic determinism” (O’Gieblyn, 2021). In other words, algorithms’ mapping of integrated information constructs meaning by drawing parameters around what is and what is not. When algorithms communicate information to humans, it then impacts the information humans use to process, operate, and exist within Earth’s sociotechnical system. Algorithmic determinism is one example of how nonhuman intelligence coupled with human-nonhuman communication can intervene and impact the consciousness of Earth’s sociotechnical system.

## AI Actors Impacting Consciousness With HMC

AI acts as an active symbiotic meaning-maker that can alter the consciousness of Earth’s sociotechnical system. Specifically, HMC affords perspective and provides explanations for the meaning-making process, the informational exchange, the alteration of consciousness that can take place between two actors within Earth’s sociotechnical system. We argue HMC acts as an intervention where humans’ communication with AI shapes and shifts the consciousness of Earth’s sociotechnical system. This concept positions HMC as not an anomaly of communication, but instead provides enriched context for the discipline of HMC and its greater contribution for advancing understandings of communication and consciousness.

---



The principles of IIT provide new insights and challenges for the field of HMC. First, the principles of IIT and HMC combined highlight the necessity for AI designers to work with HMC scholars. If AI designers better understand the impact of HMC, design can evolve to focus on how AI can responsibly participate in the consciousness of Earth's sociotechnical system. Specifically, HMC explains how human intelligence and AI can communally impact the consciousness of Earth's sociotechnical system. Following the meaning-making power of HMC, it is easier for AI designers to conceptualize the importance of AI's ability to understand the intelligence of other nonhuman actors and communicate its findings with humans. Next, we argue it is necessary to HMC scholarship to explore the human-inflicted limitations of AI intelligence and communication. In its current design state, human-centered AI learns from human intelligence and communication and focuses only on the algorithms that exist to communicate human-based system. Following IIT, AI currently operated with little consideration and concern for how HMC impacts the consciousness of Earth's sociotechnical system. If AI were designed to follow the principles of IIT and HMC, it would provide avenues for humans and AI to reach new communicative potentials and responsibly engage with and alter the consciousness of Earth's sociotechnical system.

## Moving Toward Ecological-Conscious Machines

We draw inspiration from these arguments to make the case for moving beyond simplistic renderings of AI as automated intelligence. This distinction can help advance morphologies of AI beyond mimesis of human qualities, described richly in Turkle's (2021) analysis of pretend empathy. By blurring the boundaries between human and nonhuman, these philosophies and frameworks work to undo dominant assumptions surrounding human-superiority. This does involve processes of unlearning and reimagining, so as to create responsible and trustworthy AI models (Hine et al., 2023). Were AI designers to conceptualize human and nonhuman actors as interconnected and integrated, they could advance more quickly toward a responsible, inclusive, and symbiotically driven AI tropes of being.

Feminist STS advocates for AI designers to confront and address imbalances of power in the relations between AI and the natural world (Wagman & Parks, 2021). By removing constraints pre-determining what communication is and who or rather what is considered an interlocutor, HMC has also paved the way for us to challenge how things are or should be. HMC challenges humans to reconsider how they want to interact with Earth's sociotechnical system. As such, building from Wagman & Parks's (2021) "social machine model" we call for a design of a social, responsible, and inclusively considerate AI or what we term *social-ecological machine actors*.

By opting for a less predetermined orientation that is considerate and conceptually inclusive of the intelligence of all actors, humans will allow space for AI to adopt a responsible role within Earth's sociotechnical system. As social-ecological machine actors, AI will work to understand the intelligent, informational, and communicative contributions of nonhuman actors before inserting their agency on Earth's sociotechnical system. This will create a system that is more inclusive, mutual, and equitable for all involved in the consciousness of Earth's sociotechnical system. In other words, as social-ecological machines, AI decenters the human, thus creating reciprocity. In what follows, we provide a radical

approach to responsible AI design through recommendations that demand the agency, intelligences, and meaning-making power of all actors be considered.

## 1. Symbiotic Design: AI and Mutuality

To create a more responsible AI, designers should rigorously reflect upon and engage with the relations of mutuality in their work. The guiding principle of mutuality is symbiosis. Mutuality directs designers away from design outcomes seeking to substitute. It further abandons any effort to reproduce hierarchies of intelligence. Mutuality aims to create social-ecological machines that can responsibly contribute to Earth's consciousness in symbiotic ways.

To create social-ecological machines, designers can implement actionable design interventions. To do so, it is necessary for designers to interrogate every step of the AI design process. Data collection, data labeling, data training, model design, and decisions on how to responsibly integrate an AI into Earth's sociotechnical system will require the implementation of an investigatory framework. Every step of the framework should question and analyze the AI design pipeline. At each step, designers must audit their processes and ask: Is every design decision embracing the diverse modalities of human and nonhuman intelligence and communication constituting to the consciousness of Earth's sociotechnical system? Is AI utilizing HMC in ways that are considerate of human-AI communication's impact on the consciousness of Earth's sociotechnical system? This proposed critical design process will require reflection and attention at every angle of making, designing, and iterating.

An example of this practice can be found in how designers are beginning to apply the principle of kinship to thinking about practices of reciprocal learning (Lewis et al., 2018). Many disciplines consider kinship or "mutuality of being" to be a cultural and social construction. Kinship bonds form interpersonally through "intersubjective belonging" as kin are "intrinsic to one another's existence" (Sahlins, 2011, p. 2). Following ITT, kinship networks establish the integrated information of Earth's sociotechnical system. Like consciousness, in kinship networks, what one does or suffers also happens to others. This intersubjective belonging has warranted Lewis et al. (2018) to advocate for the acceptance of AI as kin and for the inclusion of Indigenous practice into design. AI design praxis could benefit from Indigenous practice, which embraces human and nonhuman kinship and acting responsibly within diverse and more-than-human networks founded on mutuality.

However, to best implement an investigatory framework guided by a lens of mutuality, designers will need to establish an investigatory community to responsibly determine the mutual needs within Earth's sociotechnical system. Not one person or single entity should be responsible for meaning-making in a community fostered on mutuality. HMC scholars, nonhuman experts, and interdisciplinary scholars are needed to aid AI designers as they interweave nuanced understandings of HMC and various forms of nonhuman intelligence into the design of a social-ecological machine. Through communal design that operates to acknowledge the needs of a system of diverse actors, AI can more responsibly alter the consciousness of Earth's sociotechnical system. To best determine the breadth of representation needed for a social-ecological machine's communal design community, it is first necessary to map out the kinship networks constituting the consciousness of Earth's sociotechnical system.

---



## 2. Connectomapping as Connective AI

AI designers can gain a better understanding of Earth's sociotechnical system if they were to engage with the task of connecting and mapping out how human and nonhuman intelligence are connected and integrated on Earth. This process is referred to by Orange (2013) as "connectomapping," where designers map out "connectomes" or the intelligent connection points between human and nonhuman actors. Connectomapping communicates a global map of connections, a network, which can help designers decipher the ubiquitous intelligent entanglements, intentions, actions, and communications forming the consciousness of Earth's sociotechnical system. Connectomapping will reveal what forms of intelligence constitute Earth's consciousness, what gaps social-ecological machines and HMC can fill, and how designers might responsibly govern human influence and intention in AI design.

Connectomapping reveals a cyborg of interrelations, which constitute the consciousness of Earth's sociotechnical system. As such, connectomapping can provide inspiration for AI design beyond a human-centric lens. This principle resurfaces in the work of MIT roboticist and AI developer Rodney Brooks, whose work lends support to our framework. Brooks (1991) argues that to best facilitate artificial intelligence, it is necessary to move past the notion that human intelligence is superior and all-knowing. Brooks advocates for and produces AI design that utilizes nonhuman actors' intelligence, including plant and insect intelligence. In addition, Íñiguez (2017) a robot developer for the U.S. government, has moved past the limitations of using a human brain as a model for achieving artificial intelligence. Íñiguez instead prioritizes the value of octopi's distributed approach to problem-solving for AI design. Similarly, the collective intelligence of forests is inspiring AI designers to imagine new potentials for neural networks and AI (Wang et al., 2018). These examples of AI moving beyond human intelligence highlight what can be gained if AI designers utilize connectomapping as design inspiration. Following IIT, if AI can better understand and utilize the intelligence of nonhuman actors, AI can better understand the consciousness of Earth's sociotechnical system.

Specifically, connectomapping enhances feminist STS agendas which promote multi-species flourishing (Haraway, 2016) and more responsible, inclusive, and respectful human and nonhuman relations. All actors within Earth's sociotechnical system are embedded in material conditions and power structures, or what Haraway (2016) refers to as the *informatics of domination*. Connectomapping will reveal the human and nonhuman actors that AI's current human-centric design most affects. By peeling back and looking at the layers of AI's potential influence, designers can identify the enormous ramifications of a human-centric AI design and its impact on the consciousness of Earth's sociotechnical system.

Connectomapping has tremendous implications for AI design, and the scope of such a project will take a significant amount of effort, skill, insight, collaboration, and creativity. Shifting the AI design perspective from human-centered to rather an integrated web of human and nonhuman intelligences affords designers the ability to construct not only a tool or device, but rather a responsible networking relationship. This is a huge undertaking. Connectomapping requires shifting the priorities intended for AI design to instead possibilities of greater mutuality, inclusion, and diversity.

### 3. More-Than-Human AI Storytelling

It is necessary for AI designers to consider their own identity, perspective, values, intelligence, and positionality as well as those of the social-ecological machines they seek to design. When designing a responsible AI, it is important not to fall into the habit of designing AI in a single image given it will find place in a complex and integrated web of human and nonhuman relations. As is a common practice in design, the designers of a social-ecological machine will need to develop actor or *user* stories to redirect AI designers' human-centered focus. User stories will help guide designers as they conceptualize an AI that will responsibly impact the consciousness of Earth's sociotechnical system.

Like connectomapping, user stories are collaborative design tools (Cohn, 2004). User stories are short, specific, and goal oriented. User stories help AI designers focus on producing concrete and tangible outcomes for a diversity of users. By shifting design focus from the *human* to the *more-than-human*, a diverse set of user stories create a guiding project mental model. When developing AI user stories, it's important for designers to consider AI as its own actor that intelligently contributes to the consciousness of Earth's sociotechnical system. The user story format forces AI designers to think about nonhuman actors and keep nonhumans' contributions to consciousness in focus. Designers must consider all of what could go wrong with a social-ecological machine. What harm could come to the consciousness of Earth's sociotechnical system if nonhuman intelligences are not considered in the AI design process? What harm could come to Earth's sociotechnical system if nonhumans' intelligences and contributions to consciousness are considered second to humans? As such, the development of more-than-human user stories requires AI designers to engage in dialogue with their creation at all stages of their design process. By adopting HMC theory and methodologies, designers can engage in meaning-making discourse with their AI creations and assess if their design outcomes can responsibly contribute to the consciousness of Earth's sociotechnical system from a position of mutuality.

Again, this is a huge creative undertaking. What will AI conceptualized beyond human-centered design think like, sound like, or look like, and what kind of presence will it evoke? The development of a communal design community, connectomaps, and user stories provide some of the necessary support and creativity needed to produce such an outcome. However, to create a more responsible AI for the Earth's sociotechnical system, it is also necessary for designers to explore, experiment, and expand the space of AI potentiality.

### 4. AI Consciousness Awakenings and Art

To evolve AI, some scholars call for and recommend designers consider a new category of classification for AI (De Graaf, 2016; Edwards, 2018; Kahn Jr et al., 2011). A new category of classification for AI could free AI from the limited scope of AI's current human-centric lens and some human power dynamics at its inception (Wagman & Parks, 2021). As we have described, consciousness is beyond human, but humans' limitations require effort to accept and engage with the nonhumans' contributions to conscious experience. In addition, the principles of IIT imply it is possible to construct highly conscious artifacts. A new category of classification could highlight and account for how nonhuman intelligence can constitute consciousness in nonhuman systems like AI (e.g., ChatGPT). If designers approach AI

---

---

following the principles of IIT, it would provide humans with a tool to explore how it might be possible to create and adapt conscious systems. Furthermore, enriched understandings of consciousness would encourage discussions of the responsibility of AI and how to best hold AI accountable for their impact on Earth's sociotechnical system.

The intentional development of a new category of classification for AI will require a creative methodology, extensive research, and design practices aimed at creating conscious systems. Such projects will require AI designers to collaborate with both HMC and IIT researchers. Research-driven art provides the collaborative space for intent-driven research and critical exploration to take place. Developed by design researcher and digital anthropologist Caroline Sindors (2018), research-driven art starts with an intent like creating a new category of classification for AI, and then uses art as a tool to enable the research and exploration around an idea. A research-driven art outcome explores and uncovers hidden possibilities and truths. Through research-driven art AI designers can explore the potential of HMC and IIT synergistically combined. In sum, research-driven art provides a lens to focus on how the current human-centric limitations for communication, consciousness, and the categorization of AI impact AI design outcomes and humans' understanding of conscious systems.

To launch a research-driven art project, AI designers need to intentionally question what the design of a social-ecological machine will require. Specifically, research-driven art is accomplished in three stages. First, designers set their intention and research their idea. In this stage, AI designers set the intention to research the possibilities for the intelligence and consciousness of AI beings. Designers must research what consciousness is and how consciousness is impacted by communication and intelligence. Designers must ask what it would be like to lack consciousness, and how consciousness is experienced by AI and other nonhuman actors within Earth's sociotechnical system. The middle stage of the research-driven art methodology focuses on shaping and crafting an idea. In this stage, designers must follow where their research and exploration lead. Here designers will explore the potentials and the boundaries for a new category of classification for AI. In the final stage, designers communicate the body of knowledge they've accrued. The design outcome or art produced is shaped primarily by the research accrued and the question designers are ultimately seeking to answer: How might a new category of classification for AI create opportunities to create conscious and responsible AI systems? A research-driven artwork can be a workshop, a presentation, a class, and/or a conference that manifests research and knowledge. The goal of research-driven art is to create a dialogue exploring possibilities, and the result is a breadth of new potentials.

In its current state, it is easy to say a conscious AI is an impossible reality. Rather, we hope to encourage designers to birth new design possibilities and practices by challenging why a conscious AI is impossible. However, perhaps it is possible AI designers will be unable to conceive a conscious system, a social-ecological machine, separate from anthropomorphic elements given designers' embeddedness in human language and cultural meanings. Perhaps a new category of classification for AI will require new vernaculars to better embrace and regulate a responsible AI. To better connect with the more-than-human world as the disciple of HMC seeks, we advocate that AI designers and HMC scholars alike would benefit from the creation and adoption of more-than-human linguistic terminology.

---

## 5. Trans-Post-Human Epistemic Vernaculars

It is important to remember AI designers cannot take on the task of creating something new by using the same kind of thinking or terminology of the past. It is difficult to recognize, articulate, and measure AI's contribution to the consciousness of Earth's sociotechnical system with vernaculars that favor human intelligence, communication, and conscious experience. As Albert Einstein (1946) noted when he introduced a new conception for how psychists approach the structuring of the universe—we need new terms in order to embrace new ways of thinking. Coding languages, terminologies, and classifications produce and limit ways of knowing and being in the world. As Geoffrey Bowker and Susan Leigh Star (2000) find classifications and labeling embed working political infrastructures in a manner that is relatively invisible but warrants powerful consequences. If humans wish to embrace that Earth's sociotechnical system is conscious, that humans and AI impact consciousness via HMC, and that AI can be designed to responsibly impact consciousness, it is necessary to adopt new vernaculars to express new ways of thinking.

All terminology contains a worldview, and our current AI vernacular impacts the potential of AI design (Crawford, 2021). As such, designers cannot seek to create a social-ecological machine founded on principles of mutuality, one that is free from the purview of human classification and labeling without creating new vernaculars to better embrace the role intelligences, AI, and HMC play in constituting the consciousness of Earth's sociotechnical system. These new vernaculars can be inspired by the epistemologies of cultures that already respect the more-than-human world. For example, a core belief of many Indigenous epistemologies is that man is not the center of creation. Indigenous communities worldwide utilize languages, protocols, and ways of knowing to engage in dialogue with nonhumans. These intelligible discourses acknowledge Earth as a conscious sociotechnical system, which is mutually inclusive of human and nonhuman actors. Vernaculars developed via Indigenous cultural frameworks would drastically shift the social and communicative potentiality of AI and HMC.

In addition, posthuman vernaculars place humans intimately inseparable from the complex web of intelligently integrated information, which constitutes the consciousness of Earth's sociotechnical system. For example, Braidotti & Hlavajova's (2018) *Posthuman Glossary* works to “de-segregate the different and highly specialized spheres of knowledge production” by drawing connections to different generations of scholarship and users of human and nonhuman technologies (p. 5). Like the investigatory design community we advocate for in our first design recommendation, the *Posthuman Glossary* brings together thinkers, experts, and practitioners who might not otherwise conceptualize connections with each other. As a result, the *Posthuman Glossary* can help establish new terminology for both AI designers and HMC scholars as they attempt to approach the task of creating and communicating with a social-ecological machine.

## Conclusion

In this paper, we argue AI designers must recognize and correct a flawed logic presuming the superiority of humans' role within Earth's sociotechnical system. In so doing, we combine ANT, philosophy, HMC, and STS research traditions with the work of IIT scholars

---

to construct the foundation for this argument. We further draw from past and ongoing research to present examples of AI design that advances the notion of AI's ability to impact the consciousness of Earth's sociotechnical system. These examples help build understanding that designing AI as substitutes for human functions or intelligence is a practice that underestimates the relevance of nonhuman intelligence and communication.

We propose five design practices that must guide how humans think about the future of responsible AI: symbiotic design through mutuality; connectomapping as connective AI; more-than-human storytelling; designing for AI conscious awakening; and revising our design vernacular to advance language that opens new possibilities and helps address human-centered limitations. The core principles underlying these practices recognize that no actor is superior, and that Earth's sociotechnical system is comprised of intelligences that are multimodal but integrated. *Mutuality* thus invites constant and consistent exercises in reciprocity. These gradually pave the way to design practices that are ecologically responsible and not just human, but also humane. We do not expect change to be instant, but rather build for gradual and durable change to occur about stable and just foundations. *Connectomapping* permeates design and architecture mentalities as it is. We do not seek to make a new point, but rather to center and normalize a practice that is often an afterthought. By rendering connectomapping the foundational step in a design approach, we build a reflexive yet sturdy foundation. Focusing on the *stories of nonhuman being* impacted by AI creates new demands for responsible design to compliment human and commercial needs. Such a foundation can support mutuality and guide toward more-than-human and responsibility-driven approaches. Here, the measuring test for *consciousness* eschews the human to progress toward more inclusive definitions of what is conscious and what is not. Moreover, consciousness and intelligence are understood as nonbinary concepts. Therefore, humans do not construct bi-modal tests that measure the absence of presence of either, but rather the modality, the texture, the tonality, the physicality, and in general, the form that consciousness takes on (and by consequence, the form intelligence embalms itself in). Finally, advancing and possibly creating new vernaculars (or languages) that can be shared between human, human-made, and nonhuman agents presents an egalitarian approach to communication that further decenters the human. Code could be presumed to be one such example of language if it advances to incorporate the form and manner of other communication mechanisms encountered in nature. Here, we propose both a vernacular for design that de-emphasizes human prevalence and the subsequent cultivation of new languages that permit communication that advances orality to include imagery, tactility, and a broader spectrum of mechanisms for listening and speaking with the world surrounding us.

In closing, we challenge the validity of claims to artificiality and intelligence. In speaking with engineers when we collaborate, we often hear a justified complaint that AI is not intelligent enough yet. Perhaps it is not intelligent enough, but if that is the case, then neither are humans, for humans are the ones who designed it. We have made the point in this paper, and elsewhere (Papacharissi, 2015), that there is not much artificial about artificial intelligence. Crawford (2021) further proclaims that AI is neither artificial nor intelligent. Perhaps people are the ones with artificial, human-made blinders on, ones that prevent humans from evolving out of creating things in human-likeness. Yet it is by designing for the other that humans will be eventually able to come out with self-destructive and discriminatory logics that term certain things intelligent, certain artificial, and some neither.

Designing for the other, in the broadest sense of that big word replacing out *ex machina* mentalities with richer understandings of a world populated by all, the human and nonhuman, as sentient machines, or better yet, living, complex, and interconnected organisms, not superior or inferior but each unique.

## Author Biographies

**Cait Lackey** (MA, Purdue University Northwest) is a doctoral student within the Department of Communication at the University of Illinois Chicago (UIC). As an interdisciplinary scholar with a background in cognition, communication, and psychology, Cait's research focuses on the social dynamics of artificial intelligence, human-machine communication, and human-A.I. relationships.

 <https://orcid.org/0000-0002-6362-692X>

**Zizi Papacharissi** (PhD, University of Texas at Austin) is Distinguished Professor of Communication and Political Science at the University of Illinois-Chicago and Department Head of Communication. She is also University Scholar and affiliate faculty with the Discovery Partners Institute at the University of Illinois System. She has published 10 books, over 80 journal articles and book chapters, and serves on the editorial board of 15 journals. Zizi is the founding and current Editor of the open access journal *Social Media & Society*, and has collaborated with Apple, Facebook/Meta, Microsoft, Tencent, and Oculus.

 <https://orcid.org/0000-0001-7301-4620>

## References

- Adams, C., & Thompson, T. L. (2016). *Researching a posthuman world: Interviews with digital objects*. Springer. <https://doi.org/10.1057/978-1-137-57162-5>
- Bowker, G. C., & Star, S. L. (2000). *Sorting things out: Classification and its consequences*. MIT Press. <https://doi.org/10.7551/mitpress/6352.001.0001>
- Braidotti, R., & Hlavajova, M. (Eds.). (2018). *Posthuman glossary*. Bloomsbury Publishing.
- Bronner, W., Gebauer, H., Lamprecht, C., & Wortmann, F. (2021). Sustainable AIoT: How artificial intelligence and the internet of things affect profit, people, and planet. *Connected Business: Create Value in a Networked Economy*, 137–154. [https://doi.org/10.1007/978-3-030-76897-3\\_8](https://doi.org/10.1007/978-3-030-76897-3_8)
- Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, 47(1–3), 139–159. [https://doi.org/10.1016/0004-3702\(91\)90053-M](https://doi.org/10.1016/0004-3702(91)90053-M)
- Cohn, M. (2004). *User stories applied: For agile software development*. Addison-Wesley Professional. <https://dl.acm.org/doi/abs/10.5555/984017>
- Cowls, J., Tsamados, A., Taddeo, M., & Floridi, L. (2021). The AI gambit: Leveraging artificial intelligence to combat climate change—Opportunities, challenges, and recommendations. *AI & Society*, 1–25. <https://doi.org/10.2139/ssrn.3804983>
- Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press. <https://doi.org/10.2307/j.ctv1ghv45t>



- De Graaf, M. (2016). An ethical evaluation of human–robot relationships. *International Journal of Social Robotics*, 8(4), 589–598. <https://doi.org/10.1007/s12369-016-0368-5>
- Dyson, G. (2012). *Turing's cathedral: The origins of the digital universe*. Pantheon.
- Edwards, A. (2018). Animals, humans, and machines: Interactive implications of ontological classification. In A. Guzman (Ed.), *Human-machine communication: Rethinking communication, technology, and ourselves* (pp. 29–50). Peter Lang.
- Einstein, A. (1946, June 23). The real problem is in the hearts of men. *New York Times Magazine*. <https://web.archive.org/web/20180922112353/https://www.nytimes.com/1946/06/23/archives/the-real-problem-is-in-the-hearts-of-men-professor-einstein-says-a.html>
- Eisenstadt, S. N., & Aizenshtadt, S. N. (1996). *Japanese civilization: A comparative view*. University of Chicago Press.
- Goffman, E. (1967, 2005). *Interaction ritual: Essays in face to face behavior*. Routledge. <https://doi.org/10.4324/9780203788387>
- Gunkel, D. J. (2012). Communication and artificial intelligence: Opportunities and challenges for the 21st century. *communication+ 1*, 1(1), 1–25. <http://doi.org/10.7275/R5QJ7F7R>
- Guzman, A. L. (2016). Making AI safe for humans: A conversation with Siri. In R. W. Gehl & M. Bakardjieva (Eds.), *Socialbots and their friends* (pp. 85–101). Routledge. <https://doi.org/10.4324/9781315637228-11>
- Guzman, A. L. (2019). Voices in and of the machine: Source orientation toward mobile virtual assistants. *Computers in Human Behavior*, 90, 343–350. <https://doi.org/10.1016/j.chb.2018.08.009>
- Haraway, D. J. (1991). *Simians, cyborgs, and women: The reinvention of nature*. Routledge. <https://doi.org/10.4324/9780203873106>
- Haraway, D. J. (2016). *Staying with the trouble: Making kin in the Chthulucene*. Duke University Press. <https://doi.org/10.2307/j.ctv11cw25q>
- Heikkilä, M. (2022). We're getting a better idea of AI's true carbon footprint. *MIT Technology Review*. <https://web.archive.org/web/20221114182611/https://www.technologyreview.com/2022/11/14/1063192/were-getting-a-better-idea-of-ais-true-carbon-footprint/>
- Hill, D. (2008, September 1). Listening to stones. *Alberta Views*, 40–45. <https://web.archive.org/web/20170522233829/https://albertaviews.ca/listening-to-stones/>
- Hine, E., Novelli, C., Taddeo, M., & Floridi, L. (2023, November 24). Supporting trustworthy AI through machine unlearning. SSRN. <http://dx.doi.org/10.2139/ssrn.4643518>
- Íñiguez, A. (2017). The octopus as a model for artificial intelligence: A multi-agent robotic case study. *Proceedings of the 9th International Conference on Agents and Artificial Intelligence*, 2, 439–444. <https://doi.org/10.5220/0006125404390444>
- Itoi, N. G. (2019, September 19). AI and sustainability: Will AI help or perpetuate the climate crisis? *Stanford University Human-Centered Artificial Intelligence*. <https://web.archive.org/web/20220919212452/https://hai.stanford.edu/news/ai-and-sustainability-will-ai-help-or-perpetuate-climate-crisis>
- Jensen, C. B., & Blok, A. (2013). Techno-animism in Japan: Shinto cosmograms, actor-network theory, and the enabling powers of nonhuman agencies. *Theory, Culture & Society*, 30(2), 84–115. <http://doi.org/10.1177/0263276412456564>

- Jones, S. (2018). Untitled, no. 1 (Human Augmentics). In Z. Papacharissi (Ed.), *A networked self and human augmentics, AI and sentience*. Routledge. <https://doi.org/10.4324/9781315202082-14>
- Kahn Jr, P. H., Reichert, A. L., Gary, H. E., Kanda, T., Ishiguro, H., Shen, S., Ruckert, J. H., & Gill, B. (2011, March). The new ontological category hypothesis in human-robot interaction. In *Proceedings of the 6th International Conference on Human-Robot Interaction*, 159–160. <https://doi.org/10.1145/1957656.1957710>
- Latour, B. (2005). *Reassembling the social: An introduction to actor-network-theory*. Oxford University Press.
- Lewis, J. E., Arista, N., Pechawis, A., & Kite, S. (2018). Making kin with the machines. *Journal of Design and Science*, 3(5). <http://doi.org/10.21428/bfafd97b>
- Maitra, S. (2020, February). Artificial intelligence and Indigenous perspectives: Protecting and empowering intelligent human beings. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 320–326. <https://doi.org/10.1145/3375627.3375845>
- Nass, C., & Steuer, J. (1993). Voices, boxes, and sources of messages: Computers and social actors. *Human Communication Research*, 19(4), 504–527. <https://doi.org/10.1111/j.1468-2958.1993.tb00311.x>
- Neff, G., & Nagy, P. (2018). Agency in the digital age: Using symbiotic agency to explain human-technology interaction. In Z. Papacharissi (Ed.), *A networked self and human augmentics, AI and sentience*. Routledge. <https://doi.org/10.4324/9781315202082-8>
- O’Gieblyn, M. (2021). *God, human, animal, machine: Technology, metaphor, and the search for meaning*. Knopf Doubleday Publishing Group.
- Orange, E. (2013). Understanding the human-machine interface in a time of change. In R. Luppini (Ed.), *Handbook of research on technoself: Identity in a technological society* (pp. 703–719). IGI Global. <https://doi.org/10.4018/978-1-4666-2211-1.ch036>
- Papacharissi, Z. (2015). *Affective publics: Sentiment, technology, and politics*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199999736.001.0001>
- Plec, E. (2015). *Perspectives on human-animal communication*. Routledge. <https://doi.org/10.4324/9780203082935>
- Prahl, A., & Edwards, A. (2023). Defining dialogues: Tracing the evolution of human-machine communication. *Human-Machine Communication*, 6, 7–16. <https://doi.org/10.30658/hmc.6.1>
- Sahlins, M. (2011). What kinship is (part one). *Journal of the Royal Anthropological Institute*, 17(1), 2–19. <https://doi.org/10.1111/j.1467-9655.2010.01666.x>
- Sinders, C. (2018, August 16). *How to make research-driven art*. The Creative Independent. <https://web.archive.org/web/20210925155839/https://thecreativeindependent.com/essays/how-to-make-research-driven-art/>
- Spence, P. R. (2019). Searching for questions, original thoughts, or advancing theory: Human-machine communication. *Computers in Human Behavior*, 90, 285–287. <https://doi.org/10.1016/j.chb.2018.09.014>
- Sternberg, R. J. (2023). Intelligence. In V. P. Glăveanu & S. Agnoli (Eds.), *The Palgrave Encyclopedia of the Possible* (pp. 793–800). Palgrave MacMillan. [https://doi.org/10.1007/978-3-030-90913-0\\_187](https://doi.org/10.1007/978-3-030-90913-0_187)
-



- Suchman, L. (2023). Imaginaries of omniscience: Automating intelligence in the US Department of Defense. *Social Studies of Science*, 53(5), 761–786. <https://doi.org/10.1177/03063127221104938>
- Tononi, G. (2004). An information integration theory of consciousness. *BMC neuroscience*, 5, 1–22. <https://doi.org/10.1186/1471-2202-5-42>
- Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. *The Biological Bulletin*, 215(3), 216–242. <https://doi.org/10.2307/25470707>
- Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17(7), 450–461. <https://doi.org/10.1038/nrn.2016.44>
- Turkle, S. (2021). *The empathy diaries*. Penguin.
- Wagman, K. B., & Parks, L. (2021). Beyond the command: Feminist STS research and critical issues for the design of social machines. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1–20. <https://doi.org/10.1145/3449175>
- Wang, S., Aggarwal, C., & Liu, H. (2018). Random-forest-inspired neural networks. *ACM Transactions on Intelligent Systems and Technology*, 9(6), 1–25. <https://doi.org/10.1145/3232230>
- Zeki, S. (2007). The disunity of consciousness. *Progress in Brain Research*, 168, 11–18, 267–268. [https://doi.org/10.1016/S0079-6123\(07\)68002-9](https://doi.org/10.1016/S0079-6123(07)68002-9)
-



# Feminist Cybernetic, Critical Race, Postcolonial, and Crip Propositions for the Theoretical Future of Human-Machine Communication

Paula Gardner<sup>1</sup>  and Jess Rauchberg<sup>2</sup> 

1 Department of Communication Studies and Media Arts, McMaster University, Hamilton, Ontario, Canada

2 Department of Communication, Media, and the Arts, Seton Hall University, South Orange, New Jersey USA

## Abstract

The authors review theoretical trends in HMC research, as well as recent critical interventions in the *HMC* journal that usefully reshape and expand our research terrain. Conventional research such as positivist and quantified approaches are identified as restraining research questions and delimiting understandings of concepts including subjects, agency, and interactivity. Feminist cybernetic, critical race, postcolonial, and crip theoretical approaches are offered, examining how they fill research gaps in HMC, expanding content areas explored, and addressing diverse intersectional pressures, situated, and time/space dynamics that impact human-machine interaction. The authors suggest these shifts are essential to expanding HMC research to address diverse populations, regional realities around the globe, and to engage in vibrant scholarly debates occurring outside HMC. They contend these shifts will outfit HMC to weigh in on important issues of justice, equity, and access that arise with emerging technologies, climate change, and globalization dynamics.

**Keywords:** crip, critical digital race studies, feminist cybernetics, human-machine communication, postcolonial feminism

**Author Note:** This research was supported by funding from the Asper Foundation, the Social Science and Humanities Research Council, and Microsoft Research.

**CONTACT** Paula Gardner  • [gardnerp@mcmaster.ca](mailto:gardnerp@mcmaster.ca) • Department of Communication Studies and Media Arts • McMaster University • Togo Salmon Hall 331 • 1280 Main Street • West Hamilton • ON L8S 4L8, Canada.

ISSN 2638-602X (print)/ISSN 2638-6038 (online)  
[www.hmcjournal.com](http://www.hmcjournal.com)



Copyright 2024 Authors. Published under a Creative Commons Attribution 4.0 International (CC BY-NC-ND 4.0) license.

## Introduction

Human-machine communication scholars have long been developing our research in diverse journals of communication and human-machine interaction, focusing on the relationship between technologies and communication. Since 2017, scholars have sought to demark human-machine communication (HMC) as a coherent subfield, notably with interventions in the early volumes of *Human-Machine Communication* and the *SAGE Handbook*; these represent scholarly traditions as well as shifts reflecting contemporary critical turns. As the HMC subfield expands, we propose action items to expand and innovate the theoretical trajectory of HMC scholarship.<sup>1</sup> Specifically, we propose engaging feminist, including cybernetic, critical race approaches, postcolonial, and crip approaches, particularly the work of esteemed communication colleagues (conspicuously absent from HMC research), that can enrich and extend burgeoning HMC research. In engaging greater criticality, including ontological, phenomenological, and constructivist approaches, HMC research can attend more carefully to the contexts that condition machines and humans, and refine analyses regarding how human-machine interactions make possible diverse forms of subjectivity, interaction, power, identity, agency, and communication. Our intervention reflects Iliadis (2023), who contends that potent scholarship that can arise in HMC when combining “humanistic and qualitative tools, theories, methods, and frameworks” with the “long rich histories” of critical and cultural approaches in communication studies (Iliadis, 2023, pp. 117–118). Referencing canonical literature from critical and cultural studies and contemporary HMC scholars, Iliadis (2023) defines criticality as rejecting objective or universalizing views of science and technology, and relativizing subject positions and political orientations, emphasizing culture, relativity, subjectivity, standpoints, and situated interactions (p. 199).

We concur with this definition of critical research, and point readers toward lesser referenced critical approaches, particularly intersectional scholarship. Critical intersectional approaches, particularly feminist, cybernetic, critical race, postcolonial and crip conceptual frameworks, can attune research to the complexities that enable or restrain human-machine communication. Feminist cybernetic scholarship presents opportunities for HMC scholars to critically probe the relationship between gendered and racialized flows of labor and technology. Moreover, crip approaches demonstrate how the generative insights on building, hacking, and creating that emerge in disability cultures offer machines new ways of reading, organizing, and interacting with human-created data that shape unique relationships beyond ableism (see Brilmyer & Lee, 2023; Rauchberg, 2022). Moreover, recognizing the intersectional (dual micro and macro frameworks) digital, globalized, networked dynamics in which our technologies and practices operate will nuance research findings. This will enable HMC scholarship to take a more prominent role in these important scholarly discussions in and beyond communication studies. Our intervention seeks to ensure that the emerging landscape of HMC is outfitted to engage with emerging human-machine issues and realities in a shifting terrain where digital technologies and practices fluctuate regionally and globally in response to crises including industrial change, climate change, political conflicts, and globalization dynamics.

---

## ***The Theory Behind Conventional HMC Methods***

HMC scholars maintain that while communication brings a distinct, valuable approach to the study of human-machine discourse and interaction, HMC has employed a restrained collection of methods and theoretical approaches. As Wilson (2017) argues, HMC researchers choose methods based upon several considerations: “opportunities and access, resource constraints, disciplinary traditions, and ethics as well as the types of data desired, plans for data analysis, and broader assumptions about the research process” (p. 1020). Methods, of course, are binded to theoretical assumptions and dispositions. Predominant methods in HMC have included content analysis, experimental, or ethnographic methods. Critical approaches such as discourse, visual, material, ideological, aesthetic, and cultural analysis, as well as co-design and research creation, have been perceptibly peripheral in HMC research. Such absences close off experimentation and the potential of HMC to demonstrate its full potential. Engaging critical and, crucially, intersectional approaches can produce findings impactful in both communication and neighboring academic spaces (e.g., sociology, anthropology, digital humanities, science and technology studies [STS]) and expand HMC’s strength in offering policy recommendations valuable to government, industry, and cultural organizations. Scholars in these spaces have had to contend with similar reckonings. STS, for example, has successfully pressed scholars to bring greater “social thickness and complexity” to the study of technological systems (Jasanoff, 2015, p. 2).

## **The Case for Future Robust and Expansive Intersectional HMC Research**

A robust HMC field is one where scholars engage in reflexivity and trouble our theoretical assumptions; engage with contemporary theory to consider human-machine dynamics more generously; interrogate our digital and networked conditions across diverse global regions, contexts, and practices; and address questions focused on the political, justice, and climate impacts of human-machine interaction. This includes addressing the various ways in which humans and machines interact in ways that create barriers to or foster equity, diversity, access, inclusion, ethics, justice, and sustainability. Taking up these approaches should be key action items for HMC scholars—theoretical hurdles we must jump through if HMC is to ably contend with issues vital to our research terrain, engage with issues currently addressed rigorously in other areas of communication, and to confidently weigh in on key challenges facing our planet in the twenty-first century and propose steps that chart ways forward.

This essay proceeds as follows: First, we review trends in historic HMC research, followed by recent calls for innovation in HMC from various researchers, including the editors of *Human-Machine Communication (HMC)*, recognizing the journal as a primary site of emerging HMC research. We then highlight examples of HMC scholarship that demonstrates innovation, and, in turn, that which illustrates ongoing limitations, particularly in regard to theoretical breadth and intersectionality.<sup>2</sup> As well, we reflect on current efforts to (re)frame the field, focusing on the *HMC* journal, noting the editors’ calls for new types of primary research questions, content foci, and analytic lenses. In our review, we focus on noted absences that can be filled by feminist (cybernetic and critical race), postcolonial and crip approaches—potent approaches that offer innovative theoretical inquiry and

paradigmatic<sup>3</sup> orientations, in conversation with critical shifts outside of HMC. Notably, this article is a theoretical extension of our *SAGE Handbook of Human-Machine Communication* chapter (Gardner & Rauchberg, 2023). In this essay, we offer theory (arising from critical methods) that enables inventive analysis and argumentation in HMC.

## Limitations of Conventional HMC Research

Scholars have tracked the theoretical commitments of conventional HMC research as focusing on interpersonal interaction (Fortunati & Edwards, 2020) and (post-)positivist research, often employing interpersonal theory, survey-based instruments, and quantitative measures (Spinda, 2017). These theoretical foci normalize methods that capture cues and patterns of (assumedly monolithic) human subjects as they interact with computers, which glosses details of identity and cultural contexts. Such approaches limit the ways in which user/subjects are contextualized and assume that human-human communication interactions guide human-machine communications, neglecting the opportunity to complicate machine interlocutors. Stahl and Edwards (2017) review HMC research as relying on positivist and post-positivist theories, and quantitative research methodologies, such as experimental and survey research, and only minimally engaging with critical or qualitative research methods (e.g., humanistic or critical methods). In emphasizing interpersonal theories, crucial distinctions between humans and computers are blurred; the focus is often on evaluating social scripts utilized by computers upon human users, restraining consideration of the diverse types of human-computer social interaction. These limited approaches often constrain research on mobile technologies and artificial intelligence (AI) to addressing relationships, speech acts, nonverbal cues, and/or measuring the gratification computers might offer to humans. A traditional user-centric focus is overly narrow, often problematically embracing technological determinism or utopianism.

Makady and Liu's (2022) quantitative study reviewed 444 peer-reviewed empirical studies published between 2010 and 2021 across journals with the highest impact factors in the Social Sciences Citation Index (SSCI). The study tracked terms employed in articles noting their coherence and prevalence, aiming to note trends in HMC scholarship that included subject matter, and theoretical and methodological approaches. However, the study did not investigate intersectional approaches, nor track the terms we used in our review of early HMC issues. The authors instead tracked the use of the term "power"—singularly (rather than in relation to other terms) and as content rather than analytic lens—finding it was employed to inquire into the "critical role of AI in Journalism," which recurred in the *Journal of Broadcasting and Electronic Media (JBEM)*, and *Journalism Studies (JS)*. Despite such methodological limitations, the findings show HMC research in these journals addressed a limited set of (emerging) devices and gave only marginal attention to others such as wearables. The study also echoes assertions that HMC research on emerging technology research needs to work to further develop theoretical HMC-focused frameworks.

In much HMC research, scholars objectify machines and homogenize users, failing to note how biases (regarding disability, gender, race, ethnicity, and other signifiers) are embedded in both machines and in social structures, and work to condition, and impact experiences (Gardner & Kember, 2021). While some HMC scholars recognize that technology itself has become a communicator (Guzman, 2018), many still overlook important

---

feminist cybernetic scholarship recognizing practices by which humans and technology inter-inform (Haraway, 1987) or entangle (Barad, 2007) to complicate communication dynamics, which is further discussed below.

The circumvention of feminist and other critical research trends also appears in HMC scholarship in areas of interface design and ubiquitous computing research, which overwhelmingly rely on convenience sampling and experimental research design. Here research often takes a human-computer interaction (HCI) approach, seeking to improve usability via storing, retrieving, and manipulating information from interfaces in seamless manners (Stahl & Edwards, 2017). There is ample opportunity in HMC to engage a critical communication framework that complicates the notion of the universal user, foregrounding how different histories, experiences, and expectations of subject's condition and impact human-machine dynamics. In interface design and ubiquitous computing, such an approach disavows the concept of a homogenous user, addressing micro and macro contexts of use to complicate and situate research. In social computing studies, which tests how computers and interfaces facilitate interactions, this approach would contextualize the "social" in time and space.

HMC research that engages with information-processing theory (how information is processed by humans, driven predominantly by psychology) and agent goal theories (what motivates users toward a goal or activity) also tends to omit attention to user difference. In turn, these missing cultural and identity signifiers could profitably complexify analysis. While such critical lenses remain infrequent in HMC research, some approaches *do* problematize understandings of the social and the human. For example, social interaction theory in HMC attends to: "culture, situation, time, organization, physical setting, and others that are all socially embedded within each individual" (Stahl & Edwards, 2017, pp. 3–5). As well, Computers as Social Actors (CASA) framings probe human communication to understand how and why humans might respond to computers as social actors (see Nass et al., 1994).

That is to say, lessons in innovation are readily available from within the HMC community. Some HMC scholars engage critical and interpretivist paradigmatic approaches that problematize the reductive framing of human subjects, offering theoretical models that complicate notions of interaction and communication. For example, referencing machine-actor dynamics, which garners much attention in HMC, Dehnert & Leach (2021) found that humans interpreted video game scripts via ableist lenses, reading machines, for example, as sub- or superhuman which worked to manifest a sense of control or anxiety. The authors make a plea to HMC researchers to address the social biases (e.g., heteronormativity, whiteness, ableism, etc.) that condition how humans communicate with machines.

Upon reflection, while some exceptional HMC scholarship engages critical frameworks of analysis, much HMC research to date has often neglected to rigorously incorporate historical, social, regional, or cultural contexts relevant to the human-machine experience, or to critically reflect upon how theoretical framings are employed. In the next section, we review calls to action to address such absences in the early journal issues of *Human-Machine Communication*, where editors Fortunati and Edwards rigorously solicit new material aiming to reinvigorate HMC scholarship. We then provide a review of key innovative scholarship published consecutively in *HMC* Volumes I–V. We note this research as important advancements, particularly in ontological and constructivist research, innovations



in CASA, Actor-Network Theory (ANT), and in engaging interdisciplinary approaches. Finally, the essay reviews remaining gaps or weaknesses that, we propose, can be filled by feminist, crip, critical race, and postcolonial approaches.

## **Calls for HMC Innovation: Early Volumes of *Human-Machine Communication***

We reviewed the first five volumes of *Human-Machine Communication* to capture research that engaged theoretical or methodological approaches from feminist, critical race, post-colonial, critical disability, or crip approaches to HMC subject matter. We scanned these articles manually, searching for keywords including critical, feminist, cyber, colonial, post-colonial, queer, disability, crip, race, and power. We also reviewed the articles' bibliographies and citations seeking authorial references from feminist, critical race, postcolonial, anti-colonial, critical disability, and crip studies scholarship. When such evidence was found, we conducted a critical/cultural close reading of the article to assess how and in what ways the arguments and findings espoused key principles, aims, and objectives common in these approaches. In the following discussion we reference findings of evidence as well as significant deficits of these approaches in HMC.

The *HMC* editors recognize absences of critical research in the field and have stridently solicited research to the journal that engages in complex critical, contextualized, and interdisciplinary scholarship. Their calls invite "big" research questions that offer complexity beyond mere engagement with interdisciplinary methods, and provide alternative ways to analyze complex interactions (Fortunati & Edwards, 2022, p. 11) and to shift attention more rigorously toward the analysis of emerging technologies. Their appeals have advanced with each issue; Volume I (2020) and II (2021) called for critical and innovative research.<sup>4</sup> Volume IV proposed new psychosocial and cultural frameworks able to tarry with key ideas such as hybridity, otherness, relations of work, labor, and gender, which have given rise to important shifts in the social sciences. Finally, Volume V (2022b) invited nuanced research on gender in HMC with attention to historical and political dynamics that shapes it, a clear recognition that we must update HMC research to reflect advancements in gender-machine research elsewhere in communication, sociology, cultural anthropology, digital humanities, STS, and beyond.

### **Key Critical Interventions and Gaps in Early Volumes of HMC**

*HMC* Volumes I–V include the editors' introductions with inspiring arguments for theoretical advancements in HMC. Our review below is generally organized by Volume number, summarizing the editors' priorities for future HMC research, and highlighting selected innovative interventions that correct HMC's pervasive focus on human-human communication, and engage in more critical, historical, ontological, and constructivist research approaches. These essays additionally include notes regarding ongoing classic HMC approaches that can benefit by incorporating critical frameworks and contexts.

Articles in *HMC* Volume I propose a broad redefinition of HMC scholarship. For instance, HMC can address communication theories and practices with and about digital interlocutors, including the context of machine spaces, human-machine configurations,

---



and how humans and machines are constructed through discourses and interactions. The editors call for more ontological inquiries to innovate HMC, noting as example research on interactor and inter-agent communication, reflecting humans' emotional investments in relations with digital interlocutors, which productively troubles classic interpersonal theories in HMC (Fortunati & Edwards, 2020, p. 9). *HMC* Volume I also includes articles that engage classic sender/receiver models that problematically assume disembodied signaling, and communication science approaches in dialogue with (often automated) computers and robots, social robots, and conversation versus dissemination. At the same time key articles in Volume I make great strides, reflecting the editors' ambitions for the field. Banks and De Graaf (2020), for example, propose replacing the outdated transmission model of communication with an agent-agnostic transmission model that recognizes blurred ontological differences between humans and machines. They contend that scholars should focus on how machines themselves communicate, to address the "missing mass [of] . . . emerging, unintuitive, and surprising ways that humans and machines make meaning together" (Banks & De Graaf, p. 20).

In Volume II (2020), "Moving Ahead with Communication," the editors praise the interdisciplinarity approaches of articles in the issue, with notable pieces that pressure paradigmatic HMC boundaries and theoretical habits. Recognizing the central position in HMC occupied by mediated communication, the media equation, and Computers as Social Actors (CASA) (Nass et al., 1994), the editors challenge scholars to develop CASA and Media as Social Actors (MASA) approaches with historical, sociological, semiotic, and hermeneutic approaches (Fortunati & Edwards, 2021, p. 9). We concur, noting that, while HMC is indeed rich in CASA, MASA, and Actor Network Theory (ANT) approaches, much research in this terrain fails to contextualize the social, political, or embodied state of "actors" in networks. Moreover, it does not differentiate between "humans" in the human-machine dyad, and social actors in human-machine networks. A strong contribution is offered in this volume by Gibbs and colleagues' (2021) analysis of structuration theory, which addresses both micro- and macro-communication processes in the negotiation of control between human and machine agents, qualifying human experience with attention to institutional, social, cultural, and personal contexts. Such approaches, they note, shift attention from technology as object to technology as agent, allowing analysis of the roles played by agency and control to better understand HMC in organizational processes (Gibbs et al., 2021, p. 161).

Other important contributions in these *HMC* issues trouble interaction research that focuses on outcomes and glosses over deep understandings of interactivity or how human communication complicates HCI approaches, machines as social actors, and media agents (see Banks & De Graaf, 2020; Fortunati & Edwards, 2020; Guzman & Lewis, 2020; Lombard & Xu, 2021). For example, Gunkel's (2022) subsequent *HMC* Volume IV intervention, in response to Banks et al.'s (2021), demonstrates the usefulness of ontological approaches to consider ethical questions (and how we ask subjects about them) in HMC, rather than relying on applied approaches. The piece interrogates the diverse mental models and social representations people use to create perceptions, opinions, and attitudes in human-machine interactions. Such scholarly exchanges offer productive debate that is essential to keeping HMC research accountable and relevant. Volume IV (2022), engaging in psycho-social and cultural approaches to HMC, offers scholarship engaging narrativity, content analysis, and philosophical and empirical approaches. The editors praise the contributions as proactively

addressing emerging issues, and wading into fresh territory—articles, for example, that explore machines as potential moral subjects or sites of otherness and hybridity (Gunkel, 2022). To illustrate potentials for theoretical inventiveness, the editors propose that scholars might resurrect James’s (1991) pragmatic social theory of meliorism. The concept probes our human future—not via an inflexible binary of optimism/pessimism that asks what is—but rather via an “in-between” position that asks what-if (Gunkel, 2022, p. 11). The call for such innovative shifts in HMC is repeated in the volume with Richards and colleagues’ (2022), whose review of journal articles about HMC decries outdated research approaches, worrying the current research trajectory (namely laboratory cross-sectional experiments) “will lead to naivete in our understanding of HMC” (Richards et al., 2022, p. 56). As a solution, the authors call for interdisciplinary research that engages in intersectionality, to address “marginalized individuals and communities (e.g., ethnicity, class, gender identity, sexuality, sexual orientation, physical disability), critical/cultural (e.g., prejudice, discrimination), relational and group development” (Richards et al., p. 56).

Successively, in Volume V (2023), *Gender and Human-Machine Communication*, the editors’ introductory essay presents diverse theoretical approaches to gender from philosophy, women’s studies, and communication, to introduce gender as a constructed phenomenon. They review research, largely empirical, showing that power, embedded in social, industry (particularly ICTs), language, and other structures and systems, enforces and normalizes particular gendered practices. As examples, the editors cite analyses of gender perceptions (e.g., in human-robot interactions) and representation (how technologies assume a normative male subject in design).

While questions of gender representation and perception are important areas of communication research, important feminist intersectional and cybernetic approaches are not well reflected in this or previous *HMC* issues. An intersectional approach, for example, could add weight to Liu’s (2021) feminist mixed-methods study in Volume II, of advertisements marketing a holographic bride substitute in Japan. Blending visual semiotic analysis and an ANT framework, the study finds that ontological assumptions—the passive, subordinated female subject/wife—are attached to the machinic bride, glorifying the ideal. The guarded summary contends that humanized objects reflect social practices of objectification. While the editors reinforce the importance of the finding—that machines are reflective and productive of human gender relations (Fortunati & Edwards, 2021, p. 19)—we propose that a feminist intersectional approach that explores how gender power articulates to age, regional customs, and family values (in Japan) could offer a thicker reading regarding the cultural conditioning and communication with impact of female-identified machines.

An outlier in *HMC* Volume V, authored by Jarvis and Quinlan (2022), productively employs a feminist intersectional lens (addressing gender, race, class, and sexuality) to effectively analyze how Instagram shadow banning (a belief referring to a platform company’s opaque algorithmic suppression of user-generated content) impacts infertility hashtags. The authors found that hashtag patterns prioritized by Instagram worked to construct in-vitro fertilization (IVF) experiences as most accessible to White women and administered in wealthy medical spaces, thus reinforcing stratified access to IVF. This unique article integrates emerging research on shadow banning and racialized algorithm studies to create an important research question and effective critical analytic lens—an exemplary lesson for *HMC* scholars invested in contemporary critical gender analysis.

---

In *HMC* Volumes I–V, the editors succeed in laying out changes necessary in HMC, some of which are underway, to refresh our scholarship particularly calling for ontological, constructivist, feminist, and intersectional frameworks of analysis. Scholars in these volumes offer important interventions that refresh classic models, including engaging ontological approaches to understand subjects and ethics, and offering heightened constructivist approaches. Along with the editors, many authors in these volumes plead for theoretical invention and experimentation, and greater dialogue with emerging trends in communication and ancillary fields. In the following, we eagerly embrace these recommendations, including the *HMC* editors' interest in a "what if" future (Fortunati & Edwards, 2022) that can be possible with theoretical shifts, particularly injecting key critical conceptual frameworks from communication scholars that are often overlooked in HMC research.

## **Toward a "What If" Future of Human-Machine Communication: Propositions for Theoretical Shifts in the Field**

Here we introduce key challenges posed to HMC by feminist cybernetic, critical race, crip, and postcolonial approaches, and offer distinct propositions for invigorating research in HMC. Each section addresses a specific example of *HMC* research representing an innovation or a gap, and then offers propositions for integrating feminist, including critical race frameworks, postcolonial and crip approaches. The propositional sections discuss how and why these are essential interventions for this subfield, the types of new research questions that invite and how they complicate analyses to open up HMC terrain to new ideas and possibilities. We offer this intervention in the spirit of the editors' ambitions for the field, proposing that engagement with such approaches can kindle scholarship that expands the breadth of HMC subject matter, approaches, and build theory, and speculate new future questions to be asked in HMC.

### **What Can Critical Feminist, Race, Postcolonial, and Crip Lenses Bring to HMC?**

First, we summarize the key commitments that feminism, including critical race approaches, crip and postcolonial research offer that can update and innovate HMC. Our concerns are that much HMC research in the field, and a significant portion published in the *HMC* journal (despite interventions by the editors), continues to reflect conventional approaches that often engage with reductive, narrow approaches. In sum, these often: support epistemological approaches that reify essentialisms and binaries; assume technological and information systems are neutral or objective; and reify technological determinism, technological utopianism, or techno-futurism (assuming technology produces advanced humans). Such choices flatten power and ontological differentials that distinguish humans and machines, failing to complexify agency, subjectivity, and affect by neglecting to explore critical and time/space dimensions. In this way, such approaches do not consider local and intersectional contexts that impact communication. Crucially, while some HMC research nods to intersectionality, we call for more contextual and situational intersectional approaches. Our concept of intersectionality recognizes the varying dimensions by which identity signifiers

attached to subjects, including race, class, gender, disability, and colonization, interconnect to create compounding systems of discrimination. We identify these gaps and propose new theoretical frameworks in HMC research to disrupt the routinized replication of habituated theories that restrain research questions, analyses, and findings. In this way, we understand our intervention to push back against the disregarding of important innovations in scholarship outside of HMC. Below we offer our assessments of key residual theoretical gaps and analytic weaknesses in HMC scholarship, and examples that demonstrate how the addition of critical feminist, race, postcolonial and crip approaches can enable evocative research that updates HMC scholarship. We do so by placing our recommendations in conversation with prominent diverse scholarly communities to make it more relevant and impactful.

### **Proposition 1: Engage Feminist Critical Digital Race and Postcolonial Studies in HMC**

HMC spaces sparingly engage with feminist critical digital race and postcolonial theoretical frameworks, despite that much of this research comes from within communication and fields that directly feed HMC, including STS, Internet Studies, and digital humanities. These crucial approaches support analysis of how racial, gender, class, and colonial values embedded in social structures and cultural practices are inscribed in technologies, and become replicated or transformed in human-machine interactions. Intersectional scholarship offers essential critical race-informed approaches that unpack how layered forms of bias attached to identity signifiers (race, class, gender, colonialism, ableism) infuse technologies and social systems. Such understandings complicate HMC theories that assume systems and machine and human actors are innocent or homogenous and unpack how social bias impacts how humans and machinic systems interact.

In groundbreaking work, for example, feminist scholars have revealed the internet as a space where social racism moved to online (Nakamura, 2002), manifesting cybertype (racial stereotypes) structures that became part of the online experiences. Like gender, race itself is understood as a technology (Benjamin, 2019; Coleman, 2009) that amplifies racial hierarchies, replicating social divisions. At the same time, race can also be a resistive position; Bailey and Trudy (2018) coined the term *misogynoir* to illustrate how Black women's agency is systemically mocked, erased, and plagiarized in interactions with machines and platforms; Bailey and Trudy subsequently documented Black women's online responses to disrupt racial stereotypes and confer agency to human actors.

Nakamura and Chow-White's (2012) anthology offers scholars diverse methods to illustrate how race works as code, image, and interaction in non-innocent digital networks (articulated to race and other biases) to distribute privilege. In information studies, Noble (2018) has demonstrated that search engine algorithms imbue generalized racism on the internet to guide searches that reinforce racism, while Buolamwini and Gebru (2018) have shown that facial databases that feed common recognition tools are White-dominant, reflecting history technologies that have worked to surveil Blackness (Browne, 2015), particularly. TallBear (2013) offers a close reading of DNA lab science, showing that material (blood) and semiotic (race or tribe) data are conflated via "markers" that segregate Indigenous peoples in distinct genetic categories, with tragic consequences for land claims and sovereignty. The approach shows how science and social systems mutually inform to denaturalize race

---

and ethnicity, in this case, indigeneity. Machinic designs, contends Benjamin (2019), act as a “New Jim Code” that encodes inequity in machinic interactions. Similarly, Coleman (2021) writes that artificial intelligence (AI) possesses a pathological insistence on racial categories that automate the sorting of race, place, and objects (Coleman, 2021, p. 6). At that same time, race is also identified as a tool that can stratify and sanctify or support liberation and social injustice.

## The Value of Critical Digital Race Scholarship for HMC

As feminist scholars have noted, no social (including machine or platform) space is free of gender, race, and other operations of power and we must beware of assuming in our research that White or male actors are deraced or degendered. That is, intersectional critical digital and platform research frameworks apply expansively to HMC research. They can assist researchers to engage in what STS scholar Suchman (2006), among others, refer to as situated research—that which reflects on the micro and macro practices of power that inform human-machine communication in distinct spaces and times. Similarly, feminist and critical digital race scholarship shows the value of addressing historic social practices of intersectional bias to reveal often invisible, colluding White and masculinist forms of power that necessarily imbue technological tools, structures, and practices. Feminist scholars also offer metatheoretical directives—frameworks to transform colonial practices within the academy. Tallbear’s (2013) “promiscuous” standpoint approach, or objectivity in action for example, invites scholars across disciplines to work collaboratively to co-constitute research claims and outcomes; such an approach supports scholars to check biases embedded in lenses and method, and ensure ethical values reflect diverse dispositions. Sandoval (2000) redeploys Haraway’s (1987) idea of oppositional consciousness in a method constructed to aid scholars to transform theory into social action, to confront academic colonialism. Critical race, ethnicity, and indigenous approaches correct biased ontological and epistemological approaches, including those within HMC, which have historically neglected and undertheorized the intersectional dynamics of power attendant to gender, race, ethnicity, colonialism, and more.

Here we offer an example of how Gardner, co-author of this paper, engages a critical intersectional approach in her current study, which probes how and why young women (aged 18–20) navigate cyberviolence on social media platforms in regional communities in Canada and South Africa. An uncritical HMC approach might focus on how platforms such as Instagram are programmed with terms to capture cyberviolence, but fail to explore the local terms (language, emojis, etc.) recognized in youth subcultures as gender-based biases or slurs. Conversely, a critical HMC approach would address how users understand the machine’s communication nature, which in turn impacts their communication acts (Edwards, 2018) in cyberviolence scenarios. Our study, for example, probes how young women’s engagement in chat groups might be impacted by their expectations that platform algorithms might censor or delete violent gender-based cyberviolence. An uncritical study might collect data to quantify percentages of (undifferentiated) young women who use likes or shares in acts that seem to amplify cyberviolence. In contrast, our study queries how, in such cases, subjects may be navigating their identity, reputation and agency and gender



power (via culture, religion, community standards) alongside expectations regarding how platform algorithms function. Alper (2017) cautions HMC scholars to avoid assuming that technologies generally empower any (universal) subject; similarly, we can not assume that perpetrators use universal practices to harass and disempower. At the same time, we query subjects' use of technologies that appear resistive, but may instead indicate other aims. Local communities' interactions with digital technologies may be guided by their expectations of these tools, combined with distinct definitions of gender-based cyberviolence. Our study thus probes how personal belief systems (informed by local family, religious, or cultural values) may inform how young women calculate the power that social media tools render in their local social groups, and how those understandings may impact when and how they respond to cyberviolence on social media platforms. For example, young women may choose to engage confrontationally or passively with cyber perpetrators in order to avoid appearing weak, which might increase their vulnerability, or they may agree to share a sexualized photo to win community approval or enhance social status. This case study illustrates how considering a subject's assumptions about machines, regional understandings of gender power, and cultural epistemologies of gender violence produces richer understandings of how and why actor-subjects engage in communications mediated by machines.

An excellent example of such intersectional research in *HMC*, noted earlier, is Jarvis and Quinlan's (2022) study, which carefully interrogates the ways whiteness shapes gender, class, and sexuality within reproductive health messaging on Instagram. While others such as Dehnert and Leach (2021), in addition to the *HMC* editors, have called for more critical studies, we challenge HMC scholars to engage with and cite the scholarship of feminist and critical digital race scholars whose work is prominent in communication and neighboring fields, to engage in thicker analyses that more carefully link histories (past and present) of bias and prejudice to the technologies and practices we analyze.

### **Postcolonial Feminist Contributions to Human-Machine Communication**

Nearly absent in HMC scholarship are studies using feminist postcolonial media and technology approaches that articulate feminist interests to transnational, colonial, and nationalist relations, with focused attention on regional histories. Exceptional postcolonial feminist communication scholars offer blended micro and macro frameworks able to recognize the colonial values embedded in technology and networks, and actor practices, with attention to how technological flows to and within the global South impact access, uptake, and interaction. These approaches correct research that essentializes subaltern subjects (Kumar & Parameswaran, 2018) and denaturalizes North/Western research that universalizes the concept of networks, to expand understandings of how technologies and subjects arise relationally and in transnational dynamics (Shome & Hegde, 2002).

Shome (2016) seeks to expand conversations across media and postcolonial studies to unsettle the prominence of Eurocentric biases within media studies, particularly the universalization of White, Northern subjects and a history failing to recognize the complexities of colonialism. Shome's analysis shows how colonized peoples, in this case referencing India, have historically preferred different value systems (e.g., religious over secular) than colonizers in the design and uptake of media and other technologies. The article criticizes Northern scholarship that assumes technological development and use follows a coherent,

---

linear path over time and space, for example, failing to recognize the ways in which colonized peoples, often covertly, engage values in media/technology in histories that are circuitous and messy. For Shome (2016), convergence is an example of a poorly theorized Northern idea that is insensible in India, particularly among the majority with no technology access. She writes: “. . . convergence . . . obscures issues of (and is often built upon) divergences and disconnections of peoples situated in, or excluded by, contemporary capitalist mediated relations that are imbricated in geopolitics and postcoloniality” (Shome, 2016, p. 250). Shome’s appeal is akin to the one we are proposing here—that postcolonial approaches can help HMC scholars to regionalize studies of human practices with technologies, with attention to how diverse social and cultural values condition them and to understand development histories that are distinct from the North. Such analyses will be more fine-tuned and accurate and contribute to theoretically sophisticated understandings of the geopolitical dimensions in which technologies operate and flow.

Many fine examples of feminist postcolonial research in communication studies serve as excellent models for HMC. Employing online ethnography in Second Life research, Gajjala (2010) has shown that digital diasporic cultures condition subjects to manifest “authentic” cultural positions to enable their success in emergent transnational economies (p. 523). Hegde (2011) offers a groundbreaking collection of feminist transnational media and network studies addressing how globalization dynamics impact networked labor, media consumption and regulation, and identity practices (e.g., sexuality and gender). Parameswaran’s (2011) ethnographic study shows that cosmetic whitening creams are technologies that both offer Indian women cultural currency—white skin that reflects Eurocentric standards of beauty, while also reifying racial and caste biases in India. These intersectional studies produce rich, often contradictory, findings that productively complicate analysis.

These foundational intersectional, transnational studies in communication are rarely evoked or employed in HMC research. The aforementioned feminist critical race and postcolonial research scholarship has obvious relevance to HMC in exploring relations between media technologies, networks, and issues of human (including audience) consumption, and representation. However, this research also productively pressures HMC scholars to expand our conceptions of gender and race to what feminists, in the Foucauldian (post-structuralist) sense, term technologies—tools and practices. This conceptualization supports the analysis of how times/spaces and other conditions produce and reproduce gender and race in ways that might support or deny access, agency, and so forth. While some HMC scholars recognize technologies as practices, we encourage that application to race and gender, bodies and subjectivity, via intersectional frameworks, to expand attention to how micro and macro power dynamics surround and often produce human-machine relations and communication.

With great appreciation for the *HMC* editors and their broad solicitation attempts, we find little evidence of postcolonial, let alone intersectional feminist postcolonial approaches in the journal to date. The editors, in the introduction to Volume II (2021), note the importance of recognizing colonialism in theoretical work that evaluates the nature of the human being (p. 16); as well, Jarvis and Quinlan (2022) note colonization as an identity signifier that denaturalizes human subjects, and Denhart (2022) crafts human-machine sexualities, as “communicative sexuotechnical-assemblages” noting the historic exclusion of “others” from sexual science as something that has compounded colonization (Denhart, 2022, p. 131).

These brief references aside, postcolonial frameworks have not, to date, been deeply employed in HMC. In correcting this absence, HMC can move its subject matter and approaches toward greater attention to diverse global actors and agents and unique human-machine dynamics, while remaining astute and responsive to emerging—and constantly shifting—technological, sociocultural, political, and environmental global dynamics.

## **Proposition 2: Critical Disability and Crip Challenges to Human-Machine Communication**

Akin to connections in feminist and postcolonial studies, HMC is uniquely positioned to engage with innovative crip and disability justice approaches. Derived from the interstices of critical disability studies, feminist analysis, and queer theory, crip theory rejects curative and deficit conceptualizations of disability (Kafer, 2013). Instead, it presents disability as a whole, political-cultural identity always in flux and contextualized by economic, political, and cultural ideologies (McRuer, 2006). Following Fortunati and Edwards's call (2021) for work that disrupts disabled/nondisabled binaries (p. 20), we note crip, critical disability, and disability justice approaches as essential points of extension to feminist and postcolonial studies of human-machine interaction.

To date, HMC has only sparingly engaged in disability and crip research. Such practices create oversights for the ways digital technologies, such as internet-hosted platforms, are hubs for disability cultures—particularly disability justice making and organizing (Sins Invalid, 2019, p. 25). Often referred to as the “second wave” of disability rights, disability justice is a practice led and guided by the expertise of Black, Brown, Indigenous, queer, and trans disabled people across North America in the early 2000s (Sins Invalid, 2019). Committed to intersectionality (Crenshaw, 1990), disability justice and crip approaches to computing foreground the importance of understanding how disability status is negotiated by its interactions with race, gender, sexuality, class, nationality, and other political categories of identity in digital or computer-mediated spaces. The digital space is crucial for disability justice activism, art practice, archiving, and other human-machine engagements. Disability justice perspectives articulate the need to address access as a frictive, always incomplete goal that users, machines, and other interlocutors must collectively strive for to create many possibilities for human-machine engagement (Hamraie & Fritsch, 2019, p. 4). Crip approaches also interrogate the relationship between imperialism, disability, and technology (Coráñez Bolton, 2023; Jerreat-Poole, 2022). Influenced by feminist, critical race, and postcolonial analyses of technology, crip approaches to HMC equally articulate boundary-pushing research of understanding the role of cultural contexts in platforms, systems, and human-machine interactions through various methodological orientations and approaches. Some of these projects offer challenges to ableist ideas about human-computer relationalities through crip and neuroqueer technoscience (Banner, 2019; Hamraie & Fritsch, 2019; Rauchberg, 2022; Sterne, 2019), collective access-making (Gotkin, 2019; Hamraie & Fritsch, 2019; Jackson et al., 2022), crip HCI and information studies (Brilmyer & Lee, 2023; Shew, 2020; Sum et al., 2022; Williams et al., 2021); and participatory digital arts-based approaches (Britton & Paehr, 2021; Lazard, 2018; Sick in Quarters, 2020).

While existing HMC work lacks in quantity, early work in the *HMC* journal on disability offers critical beginnings to design and usability through analyses of human-machine

---



relations as they are represented in new media texts. For instance, Dehnert and Leach (2021) call for more critical approaches in their critical constructivist case study, probing how gamers' scripts reveal ableist views of the normal body and ableist stigmas. The pair call upon researchers to challenge our methodological habits, questioning for example, how human interaction scripts might embed harmful principles and instigate harmful relations with machines. Davis and Stanovsek's (2021) discussion of disabled users on the virtual reality platform Second Life address the use of avatars as digital embodied identity, and the concurrent benefits and limitations disabled platform users face. For instance, though the platform provides benefits for disabled people to connect and build community (particularly in a pandemic), some forms of virtual communication, such as typed gestures, are inaccessible to blind/low vision users and those accessing the platform with screen readers (p. 131). Though they do not use the term collective access (see Hamraie & Fritsch, 2019), Davis and Stanovsek's (2021) digital ethnography provides crucial insight on the frictive nature of accessibility, challenging the mainstream assumption that accessibility is universally experienced by all disabled people everywhere. Additionally, Denhert's (2022) new materialist study of sex robots through an HMC lens rallies researchers in the subfield to consider crip and critical disability analyses of human-machine relationalities.

While this existing HMC work addresses the violent encoding of ableism in human-machine relations, previous writing does not identify how crip computational and design practices can mutually inform human-machine relations in complex, expansive ways. We call for an HMC approach that imagines disability as a theoretical and methodological intervention for broadening and deepening our understanding of human-machine relations. For instance, both Fritsch and Hamraie's (2019) articulation of crip technoscience and Rauchberg's (2022) extended provocation of neuroqueer technoscience offer exciting possibilities for HMC researchers. Notably, co-author of this paper, Rauchberg (2022)'s, invocation of neuroqueer technoscience offers salient nodes for empirical researchers to study disability and self-expression in human-machine relations. Her provocations call for integrating disabled expertise and leadership in the development of human-machine relationships (p. 383). Such methodological offerings can support HMC scholarship to think beyond siloed user-machine divides, and begin to think through the nuanced, complex relationships emerging from computer and human engagements.

Prioritizing human communication and social interactionist approaches, we propose that previous work in the field can also be nuanced with critical feminist situated (Haraway, 1987; Suchman, 2006) and crip approaches to technology and user-experience. Williams et al.'s (2023) introduction of counterventions draws from feminist standpoint theory and crip HCI (Williams et al., 2021) to develop practices for addressing ableism in intervention-based computing systems. The authors identify five steps for engaging in feminist and crip counterventions to substantiate more ethical human-computer engagements: reflexively engaging with stakeholders; critically examining the disconnects between a researcher's intervention and a user's access needs; interrogating the intervention's ideological orientations; developing an intervention that engages in self-critique; and privilege stakeholder experience and leadership in the design and intervention process (Williams et al., 2023, p. 7). Williams et al.'s (2023) discussion of counterventions demonstrates how our propositions for feminist and crip approaches to the study of HMC are mutually constitutive—used together, these

critical theoretical framings introduce exciting possibilities for HMC research to consider questions of power and justice.

Finally, the invocation of crip time transcends past nondisabled notions of time, embodiment, and technology, offering theoretical and paradigmatic contributions to HMC scholarship. Crip time (Kafer, 2013) departs from able-bodied and neurotypical conceptualizations of time: bending the clock to meet people where they are (p. 26). Instead, crip time works alongside technology to provide interdependence for disabled users. Crip time disrupts technoableist (Shew, 2020) uses of assistive tech as a curative measure. Doing so reorients them toward an interdependent flow of relationality between machine and disabled users. Crip HCI considers interdependent transformative alternatives for assistive tech, establishing important nodes for HMC. For example, as a way to challenge assimilative practices in machine learning in “ABLE,” a participatory gaming project for older adults with dementia, Gardner et al. (2021) propose training their prototype’s inertial measurement unit (IMU) sensors to understand multiple types of movements instead of forcing users to assimilate toward a “normative” style. Moreover, crip time as a theoretical framing offers creative, critical methodologies for interrogating the relationship between ableism, colonialism, and human-machine relationalities through digital storytelling (Dion-Fletcher, 2019), video performance (Lazard, 2018), and autoethnography (Forlano, 2017; Rauchberg, 2022). This practice departs from postpositivist and quantitative work, offering multi-perspectival, critical, and context-specific possibilities for the future of HMC.

### **Proposition 3: Feminist Posthuman Approaches Addressing Gender, Embodiment, and Interaction in Human-Machine Studies**

Critical cyberfeminism<sup>4</sup> is a rich area of scholarship within and beyond the field of communication that probes the relationship between feminism and cyberspace, the internet, and digital technologies, beginning with new media but advancing to consider platforms, networks, and systems. It is rarely addressed in HMC, excepting occasional references to Haraway’s (1987) famous concept of the cyborg where its usages tend to dismiss the term’s grounding in critical feminist race approaches. While cyberfeminism may be considered a densely theoretical framework, we work here to expose key considerations that will make the frameworks approachable.

Where some forms of cyberfeminism address the internet as a space that liberates subjects from social constructs (gender, race, disability), and levels access, critical cybernetic feminism exposes these ideas as mythology. Haraway (1987) establishes the cyborg, referencing Third World feminists’ strategic work at the margins, that trounce patriarchal power operating through technologies. The cyborg human-machine hybrid rejects humanist binaries that falsely polarize humans and machines, and positions women (and others) as lacking, deficient, natural, weak, and irrational, and machines as unlively and inert. For Haraway, the networked worlds of computers, infected by origin stories (e.g., Christianity and patriarchy) and the informatics of domination (structural and theoretical forces devoted to binaries) offer potentials for potent human-machine fusions, and transgressive freedoms.

---

## From Feminist STS to Patterns of Intra-Action

Cybernetic feminism shares with Feminist STS approaches informed by situated and robust sociocultural analyses that complicate understandings of interactivity and debunk techno-determinist assumptions. In her landmark book, *Situated Actions*, Suchman (2007) shows that users rely on human conversational norms, rather than machinic instructional logic, to understand how to interact with machines (p. 283). This revelation, only sometimes referenced in HMC, should inform how researchers set up studies of humans reading and responding to machinic scripts.

Many cybernetic feminists, particularly Hayles (1999), Barad (2007), and Braidotti (2013) have expanded upon Haraway's cyborg. Their scholarship offers epistemological challenges to how networks are imagined, referencing the distributed system model as one where subjects and actors mutually or intra-inform, in ongoing dynamics that tend to reproduce embedded social and structural bias. These approaches, further discussed below, offer metaphysical challenges to how scholars imagine networks, actors, and interaction and troubles HMC research that assumes systems and networks communications are static, universal, or exist within singular spheres of power. Specifically, Hayles contends that machine and human cognition inter-form networks in a process of distributed cognition (or deep attention). Haraway (2006) disrupts the idea of mutually informed intelligibility in network studies, offering an alternative where humans and machines inter-inform to create meaning and knowledges over time. Barad (2007) counters with the provocative concept of intra-action, derived from quantum physics, contending that humans, machines (and all stuff) co-evolve in disparate, unpredictable ways that reflect the layers of (emerging) context that inform all (animate and inanimate) actors and objects. There is great relevance here to HMC: Barad's (2007) "ethico-onto-epistemological" approach complicates ANT by interrogating the apparatus within material and social realities that evolve in shifting relations. The potency of the concept of intra-action is illustrated in Gardner and Jenkins (2015), who used it to understand how participants read data visualized by consumer biometric devices; they discovered that participants engaged in complex intra-actions with the machinic representations, including converting them into narratives inspired by their embodied experience, and the *virtual* pasts of their own lives.

As well, Braidotti (2013) and Barad (2007) disrupt assumptions that communication (or interaction) dynamics occur in stable time/space realities. Instead, they show that geopolitical relations impact all human-machine interactions. Challenging our understanding of matter as inert, Bennett (2010) complicates it as vibrant, engaging an ecological sensibility, and expanding Latour's (2007) ANT approach with Deleuze and Guattari's (1980) assemblage theory.<sup>6</sup> Her feminist, situated, embodied approach augments ANT theory, enabling analysis of how machines and technologies impact intelligibility, agency, interaction, and innovation. These conceptual frameworks disavow coherent networks and any universal, objective, or innocent subject (commonly assumed in HMC). These approaches can be used to explore, practically, how the layered dynamics of power and/or privilege can impact human-machine interactions and communications, subjectivity, to produce (or otherwise inform) embodiment, agency, or automation, or in metaphysical studies speculating how subjects come into being or becoming. These interventions challenge well-used approaches, such as ANT, and offer innovative frameworks that complicate how we address context

(e.g., adding geopolitical and other time/space dimensions), and finally, inject greater attention to how embodiment impacts agency and interaction, opening HMC into these vibrant theoretical conversations within and beyond communication studies.

## Evidence and Potentials of Feminist Posthumanism in HMC

HMC has not rigorously engaged feminist cybernetic theory and continues to engage with critical feminist approaches only sparsely. Still, we are encouraged by the editors' call to theorize beyond "binary" gender and discourse models, to probe discourses of power and privilege, and engage feminist and disability frameworks, which will bring more critical analysis of the normative body to HMC. As well, the post humanist challenge to the antiquated human-machine dyad is well represented in some ANT studies in HMC and researchers have pressured traditional ontological and epistemological assumptions in ANT. Banks and De Graaf's (2020) study of robots, for example, probes the ontological nature of nonhuman actors' understanding of linguistic capability. Guzman (2020) presses ontological questions regarding how social representations of machines impact human experiences with machine's potential communication abilities. Additionally, Sandry's (2015) challenge to ontological habits of ascribing human to human communication patterns to robots engages Hayles (1999), recognizing the messy reality of human communication as both distinct from and entwined with robot communication. The authors reevaluate the human-robot boundary as permeable (Fortunati & Edwards, 2021, p. 15, quot. Sandry, 2015), provoking Hayle's (1999) interest in understanding humans and computers as dynamic partnerships.

We propose more such challenging feminist ANT approaches in HMC, which complicate essentialist and binary gender assumptions and asymmetric framings of gender to technology (Lagesen, 2012). They work to destabilize key analytic concepts in HMC (life, object, agent) and address how material (e.g., biological, physiological) and social relations intra-inform, to trouble how we understand subjectivity, perception, and cognition in human-machine interactions and spaces. Usefully, a feminist post humanist approach can also posit flaws in post-anthropological assumptions. An example is Braidotti's response (2013) to Verbeek's (2008) popularly cited theory of nonhuman agency, whereby technologies actively contribute to how humans conceptualize power and address ethical questions in human-machine relations. Braidotti challenges that Verbeek problematically applies human ethics to technology, shifting moral intentionality from an autonomous transcendental consciousness to technological artifacts, suggesting this devalues complex (and diverse) human positions. This type of intervention exemplifies the potentials for feminist cybernetics to challenge theory habitually referenced and reified in HMC, again providing useful pressure that tests, deepens, and expands the terrain of HMC research.

## Conclusion

As human-machine communication (HMC) scholarship seeks to expand its theoretical and paradigmatic approaches, there is an unprecedented opportunity to learn from and engage with feminist, critical race, postcolonial, and crip frameworks, arising from within and beyond communication studies. We propose that HMC researchers should expand the repertoire of both theory and paradigm to complicate normative conceptualizations of

---

actors, interactivity, interaction, agency, to challenge habituated HMC theory, and engage micro and macro contexts to trace the messy operations of power vis a vis various forces, and diverse temporal and spatial planes. The feminist, critical race, postcolonial, and crip scholarship we have offered assists scholars to locate and trouble conventional ontological and epistemological assumptions; we recommend these approaches to update references to conventional HMC cannon and to oft-cited Western critical and postmodern theories in HMC research.

While decidedly underutilized in HMC, feminist, critical race, postcolonial, and crip approaches offer strategies to interrogate material artifacts, data, technologies, practices, and framings that can innovate research designs, methods, and insert new ethical considerations. This research would expand HMC terrain to include greater and richer considerations of gender, race, disability, and postcolonial manifestations of human-machine dynamics. These dynamic interventions enable scholars to address the material, ontological, and epistemological realities and contexts shaping regional and global human-machine dynamics, thus encouraging HMC research to be more global, situated, nuanced, and relevant. In moving more intentionally into the experiential and situational world of diverse global actors and dynamics, HMC shifts our work into the space of emerging human and communication practices. HMC scholarship reflecting this breadth and depth would outfit scholars with ongoing agility, and to have greater relevance and impact within communication and allied fields, including HCI, digital humanities, STS, and beyond.

## Notes

1. We follow Lindlof and Taylor's (2017) definition of theory as "... any systematically developed account of communication that seeks to explain what it is and how it works" (p. 50).
2. Our use of the word intersectional recognizes both Crenshaw's (1990) coining of the term and formative scholarship by Third World feminists (Anzaldúa, 1987; Combahee River Collective, 1977; Lorde, 1984) describing how layered social identity factors generate exponential practices and systems of bias.
3. We present paradigm as "fundamental . . . frames of reference that we use to justify our choices in designing and conducting communication research" (Lindlof & Taylor, 2017, p. 6).
4. Notably guest editors of *HMC* Volume III (2021) sought research emerging from the COVID-19 pandemic; the issue took a more practical approach, asking scholars to produce *holistic discourse* analyzing how partnerships with humans make possible, recognize, or shape communicable machines. Because the *HMC* editors did not inject a call for innovation into Volume III, we do not address its content in this article.
5. Cyberfeminism was a term invented by Sadie Plant, as explained by Bassett (1997) to denote a post-human insurrection, where an emergent system of women and computers revolts against patriarchy as a worldview and material reality that seeks to subdue them.
6. Bennett (2010) seeks to understand how all things are connected, complicating traditional notions of relationality via a feminist material analysis of embodiment

(desire, sensations). Her positive ontology approach probes the vibrancy of matter, challenges life/matter boundaries, and understands the political contributions of nonhuman matter, as stretching “received concepts of agency, action, and freedom” (p. viii).

## Author Biographies

**Paula Gardner** (PhD) is Professor and Asper Chair in Communication in the Department of Communication Studies and Media Arts at McMaster University, and directs the [Pulse Lab](#), creating art/technology with diverse publics for social change. Gardner is a media and technology scholar-practitioner, engaging feminist, postcolonial, and critical disability frames to address digital literacy, access and inclusivity; technological bias; gender-based oppression; and foster ethical collaboration. Her current co-design project is ABLE Village, an interactive arts/game platform for diverse older adults to enhance discovery, wellness, and kinship. Gardner’s work is published in *Communication*, *Digital Humanities*, *Feminist, STS*, and *HMC/I spaces*. <https://paulagardner.ca> and <https://pulselab.humanities.mcmaster.ca>.

 <https://orcid.org/0000-0002-2190-8021>

**Jess Rauchberg** (PhD, McMaster) is an Assistant Professor of Communication Technologies at Seton Hall University. Her scholarship investigates the relationship among disability, platform ideologies, and cultural production in the creative economy. Rauchberg’s work appears in *New Media & Society*, *Feminist Media Studies*, and *First Monday*, in addition to other journals and edited collections. <https://www.jessrauchberg.com>.

 <https://orcid.org/0000-0003-2513-5107>

## References

- Alper, M. (2017). *Giving voice: Mobile communication, disability and inequality*. MIT Press.
- Anzaldúa, G. (1987). *Borderlands/La Frontera: The New Mestiza*. Aunt Lute Books.
- Bailey, M., & Trudy. (2018). On misogynoir: citation, erasure, plagiarism. *Feminist Media Studies*, 18(4), 762–768. <https://doi.org/10.1080/14680777.2018.1447395>
- Banks, J., & De Graaf, M. M. A. (2020). Toward an agent-agnostic transmission model: Systematizing anthropocentric and technocentric paradigms in *Communication*. *Human-Machine Communication*, 1, 19–36. <https://doi.org/10.30658/hmc.1.2>
- Banks, J., Koban, K., & Chauveau, P. de V. (2021). Forms and frames: Mind, morality, and trust in robots across prototypical interactions. *Human-Machine Communication*, 2, 81–103. <https://doi.org/10.30658/hmc.2.4>
- Banner, O. (2019). Technopsyence and Afro-Surrealism’s criptistemologies. *Catalyst: Feminism, Theory, Technoscience*, 5(1), 1–29. <https://doi.org/10.298968/cftt.v5i1.29612>
- Barad, K. (2007). *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning*. Duke University Press.



- Bassett, C. (1997). With a little help from (our) new friends? *Mute*, 1(8). <https://web.archive.org/web/20121113103214/https://www.metamute.org/editorial/articles/cyberfeminism-spl-little-help-our-new-friends>
- Benjamin, R. (2019). *Race after technology: Abolitionist tools for the New Jim Code*. Polity.
- Bennett, J. (2010). *Vibrant matter: A political ecology of things*. Duke University Press.
- Braidotti, R. (2013). *The posthuman*. Polity.
- Brilmyer, G., & Lee, C. (2023). Terms of use: Crip legibility in information systems. *First Monday*, 28(1–2). <https://doi.org/10.5210/fm.v28i1.12935>
- Britton, L., & Paehr, I. (2021). Con(fuse)ing and re(fuse)ing barriers. *APRJA: A Peer-Reviewed Journal about Research Refusal*, 1(1), 1–14. <https://doi.org/10.7146/aprja.v10i1.128188>
- Browne, S. (2015). *Dark matters: On the surveillance of Blackness*. Duke University Press.
- Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, 81, 77–91. [http://proceedings.mlr.press/v81/buolamwini18a.html?mod=article\\_inline](http://proceedings.mlr.press/v81/buolamwini18a.html?mod=article_inline)
- Coleman, B. (2009). Race as technology. *Camera Obscura*, 24(1), 177–207. <https://doi.org/10.1215/02705346-2008-018>
- Coleman, B. (2021). Technology of the surround. *Catalyst: Feminism, Theory, Technoscience*, 7(2), 1–21. <https://doi.org/10.29868/cftt.v7i2.35973>
- Combahee River Collective. (1977). The Combahee River collective statement. <https://web.archive.org/web/20201109143613/https://www.blackpast.org/african-american-history/combahee-river-collective-statement-1977/>
- Corañez Bolton, S. (2023). *Crip colony: Mestizaje, US imperialism, and the queer politics of disability in the Philippines*. Duke University Press.
- Crenshaw, K. (1990). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, 43, 1241–1293.
- Davis, D. Z., & Stanovsek, S. (2021). The machine as an extension of the body: When identity, immersion and interactive design serve as both resource and limitation for the disabled. *Human-Machine Communication*, 2, 121–135. <https://doi.org/10.30658/hmc.2.6>
- Dehnert, M., & Leach, R. B. (2021). Becoming human? Ableism and control in Detroit: Become human and the implications for human-machine communication. *Human-Machine Communication*, 2, 137–152. <https://doi.org/10.30658/hmc.2.7>
- Deleuze, G., & Guattari, F. (1980). *A thousand plateaus*. University of Minnesota Press.
- Denhart, M. (2022). Sex with robots and human-machine sexualities: Encounters between human-machine communication and sexuality studies. *Human-Machine Communication*, 4, 131–151. <https://doi.org/10.30658/hmc.4.7>
- Dion-Fletcher, V. (2019). Own your cervix. *Canadian Journal of Disability Studies*, 8(1), 160–163. <https://doi.org/10.15353/cjds.v8i1.475>
- Edwards, A. P. (2018). Animals, humans, and machines: Interactive implications of ontological classification. In A. L. Guzman (Ed.), *Human-machine communication: Rethinking communication, technology, and ourselves* (pp. 29–59). Peter Lang. <https://doi.org/10.3726/b14399>
- Forlano, L. (2017). Data rituals in intimate infrastructures: Crip time and the disabled cyborg body as an epistemic site of science. *Catalyst: Feminism, Theory, Technoscience*, 3(2), 1–28. <https://doi.org/10.28968/cftt.v3i2.28843.17>



- Fortunati, L., & Edwards, A. (2020). Opening space for theoretical, methodological, and empirical issues in human-machine communication. *Human-Machine Communication, 1*, 7–18. <https://doi.org/10.30658/hmc.1.1>
- Fortunati, L., & Edwards, A. (2021). Moving ahead with human-machine communication. *Human-Machine Communication, 2*, 7–28. <https://doi.org/10.30658/hmc.2.1>
- Fortunati, L., & Edwards, A. (2022). Framing the psycho-social and cultural aspects of human-machine communication. *Human-Machine Communication, 4*, 7–26. <https://doi.org/10.30658/hmc.4.1>
- Gajjala, R. (2010). Placing South Asian digital diasporas in second life. In T. K. Nakayama & R. T. Halualani (Eds.), *The handbook of critical intercultural communication* (pp. 517–533). Wiley & Sons.
- Gardner, P., & Jenkins, B. (2015). Bodily intra-actions with biometric devices. *Body & Society, 22*(1), 1–28. <https://doi.org/10.1177/1357034X15604030>
- Gardner, P., & Kember, S. (2021). Introduction: Probing the system: Feminist complications of automated technologies, flows, and practices of everyday life. *Catalyst: Feminism, Theory, Technoscience, 7*(2), 1–15. <https://doi.org/10.28968/cftt.v7i2.36962>
- Gardner, P., & Rauchberg, J. (2023). Feminist, postcolonial, and crip approaches to human-machine communication methodology. In A. Guzman, R. McEwen, & S. Jones (Eds.), *The SAGE handbook of human-machine communication* (pp. 252–260). SAGE.
- Gardner, P., Surlin, S., Akinyemi, A., Rauchberg, J., Zheng, R., McArthur, C., Papaioannu, A., & Hao, Y. (2021). Designing a dementia-informed, accessible, co-located gaming platform for diverse older adults with dementia, family, and carers. In Q. Gao & J. Zhou (Eds.), *Human aspects of IT for the aged population: Supporting everyday life activities* (pp. 58–77). Springer, Cham. [https://doi.org/10.1007/978-3-030-78111-8\\_4](https://doi.org/10.1007/978-3-030-78111-8_4)
- Gibbs, J. L., Kirkwood, G. L., Fang, C., & Wilkenfield, J. N. (2021). Negotiating agency and control: Theorizing human-machine communication from a structurational perspective. *Human-Machine Communication, 2*(1), 153–171. <https://doi.org/10.30658/hmc.2.8>
- Gotkin, K. (2019). Crip club vibes: Technologies for new nightlife. *Catalyst: Feminism, Theory, Technoscience, 5*(1), 1–7. <https://doi.org/10.28968/cftt.v5i1.30477>
- Gunkel, D. J. (2022). The symptom of ethics: Rethinking ethics in the face of the machine. *Human-Machine Communication, 4*, 67–83. <https://doi.org/10.30658/hmc.4.4>
- Guzman, A. L. (2018). What is human-machine communication, anyway? In A. L. Guzman (Ed.), *Human-machine communication: Rethinking communication, technology, and ourselves* (pp. 1–29). Peter Lang.
- Guzman, A. L. (2020). Ontological boundaries between humans and computers and the implications for human-machine communication. *Human-Machine Communication, 1*, 37–54. <https://doi.org/10.30658/hmc.1.3>
- Guzman, A. L., & Lewis, S. C. (2020). Artificial intelligence and communication: A human-machine communication research agenda. *New Media & Society, 22*(1), 70–86. <https://doi.org/10.1177/1461444819858691>
- Hamraie, A., & Fritsch, K. (2019). Crip technoscience manifesto. *Catalyst: Feminism, Theory, Technoscience, 5*(1), 1–31. <https://doi.org/10.28968/cftt.v5i1.29607>
- Haraway, D. J. (1987). A manifesto for cyborgs: Science, technology, and socialist feminism in the 1980s. *Australian Cultural Studies, 2*(4), 1–42. <https://doi.org/10.1080/08164649.1987.9961538>
-

- Haraway, D. J. (2006). *When species meet*. University of Minnesota Press.
- Hayles, N. K. (1999). *How we became posthuman: Virtual bodies in cybernetics, literature, and informatics*. University of Chicago Press.
- Hegde, R. S. (2011). *Circuits of visibility: Gender and transnational media cultures*. NYU Press.
- Iliadis, A. (2023). Critical and cultural approaches to human-machine communication. In A. Guzman, R. McEwan, & S. Jones (Eds.), *The SAGE handbook of human-machine communication*. SAGE.
- Jackson, L., Haagaard, A., & Williams, R. M. (2022, April 19). Disability dongle. *Platypus: The CASTAC blog*. <https://blog.castac.org/2022/04/disability-dongle/>
- James, W. (1991). *Pragmatism*. Prometheus Books.
- Jarvis, C. M., & Quinlan, M. M. (2022). IVF so White, so medical: Digital normativity and algorithm bias in infertility on Instagram. *Human-Machine Communication*, 5, 133–149. <https://doi.org/10.30658/hmc.5.6>
- Jasanoff, S. (2015). One. Future imperfect: Science, technology, and the imaginations of modernity. In S. Jasanoff & S. Kim (Ed.), *Dreamscapes of modernity: Sociotechnical imaginaries and the fabrication of power* (pp. 1–33). University of Chicago Press. <https://doi.org/10.7208/9780226276663-001>
- Jerreat-Poole, A. (2022). Virtual reality, disability, and futurity: Crippling technologies in Half Life: Alyx. *Journal of Literary and Cultural Disability Studies*, 16(1), 59–75. <https://muse.jhu.edu/article/847103/summary>
- Kafer, A. (2013). *Feminist, queer, crip*. Indiana University Press.
- Kumar, S., & Parameswaran, R. (2018). Charting an itinerary for postcolonial communication and media studies. *Journal of Communication*, 68(2), 347–358. <https://doi.org/0.1093/joc/jqx025>
- Lagesen, V. A. (2012). Reassembling gender: Actor-network theory (ANT) and the making of the technology in gender. *Social Studies of Science*, 42(3), 442–448.
- Latour, B. (2007). *Reassembling the social: An introduction to actor-network-theory*. Oxford University Press.
- Lazard, C. (2018). Crip time [Video file]. <https://vimeo.com/clazard>
- Lindlof, T. R., & Taylor, B. C. (2017). *Qualitative communication research methods* (4th ed.). SAGE.
- Liu, J. (2021). Social robots as the bride? Understanding construction of gender in a Japanese social robot product. *Human-Machine Communication*, 2, 105–120. <https://doi.org/10.30658/hmc.2.5>
- Lombard, M., & Xu, K. (2021). Social responses to media technologies in the 21st century: The media are social actors paradigm. *Human-Machine Communication*, 2, 29–55. <https://doi.org/10.30658/hmc.2.2>
- Lorde, A. (1984). *Sister outsider: Essays and speeches*. Crossing Press.
- Makady, H., & Liu, F. (2022). The status of human-machine communication research: A decade of publication trends across top-ranking journals. In M. Kurosu (Ed.), *Human-computer interaction: Theoretical approaches and design methods*. HCII 2022. Lecture notes in computer science (pp. 83–103). Springer. <https://doi.org/10.1007/978-3-031-05311>

- McRuer, R. (2006). *Crip theory: Cultural signs of queerness and disability*. NYU Press.
- Nakamura, L. (2002). *Cybertypes: Race, ethnicity, and identity on the internet*. Routledge.
- Nakamura, L., & Chow-White, P. (2012). *Race after the internet*. Routledge.
- Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *CHI '94: The 1994 ACM Conference on Human Factors in Computing Systems*, 72–78. <https://doi.org/10.1145/191666.191703>
- Noble, S. U. (2018). *Algorithms of oppression: How algorithms reinforce racism*. NYU Press.
- Parameswaran, R. (2011). E-Race-ing color: Gender and transnational visual beauty economies in India. In R. S. Hegde (Ed.), *Circuits of visibility: Gender and transnational media cultures* (pp. 68–88). NYU Press.
- Rauchberg, J. S. (2022). Imagining a neuroqueer technoscience. *Studies in Social Justice*, 16(2), 370–388. <https://doi.org/10.262522/ssj.v16i2.3415>
- Richards, R. J., Spence, P. R., & Edwards, C. C. (2022). Human-machine communication scholarship trends: An examination of research from 2011 to 2021 in communication journals. *Human-Machine Communication*, 4, 45–65. <https://doi.org/10.30658/hmc.4.3>
- Sandoval, C. (2000). *Methodology of the oppressed*. University of Minnesota Press.
- Sandry, E. (2015). *Robots and communication*. Palgrave-MacMillan.
- Shew, A. (2020). Ableism, technoableism, and future AI. *IEEE Technology and Society Magazine*, 31(2), 40–85. <https://doi.org/10.1109/MTS.2020.2967492>
- Shome, R. (2016). When postcolonial studies meets media studies. *Critical Studies in Media Communication*, 33(3), 245–263. <https://doi.org/10.1080/15295036.2016.1183801>
- Shome, R., & Hegde, R. S. (2002). Postcolonial approaches to communication: Charting the terrain, engaging the intersections. *Communication theory*, 12(3), 249–270. <https://doi.org/10.1111/j.1468-2885.2002.tb00269.x.21>
- Sick In Quarters. (2020, December 31). SiQ for 8Ball Community TV [Video file]. <https://www.youtube.com/watch?v=3nRjDyXmK2c>
- Sins Invalid. (2019). *Skin, tooth, and bone: The basis of movement is our people—A disability justice primer* (2nd ed.). Sins Invalid.
- Spinda, J. W. (2017). Communication and technology. In M. Allen (Ed.), *The SAGE encyclopedia of communication research methods* (pp. 174–177). SAGE.
- Stahl, B., & Edwards, C. (2017). Human-computer interaction. In M. Allen (Ed.), *The SAGE encyclopedia of communication research methods* (pp. 671–674). SAGE.
- Sterne, J. (2019). Ballad of the dork-o-phone: Towards a crip vocal technoscience. *International Journal of Interdisciplinary Voice Studies*, 4(2), 179–189. [https://doi.org/10.1386/ijvs\\_00004\\_1](https://doi.org/10.1386/ijvs_00004_1)
- Suchman, L. (2006). *Human-machine reconfigurations: Plans and situated actions*. Cambridge University Press.
- Sum, C. M., Alharbi, R., Spektor, F., Bennett, C. L., Harrington, C. N., Spiel, K., & Williams, R. M. (2022). Dreaming disability justice in HCI. *CHI EA '22: Extended Abstracts of the 2022 Conference on Human Factors in Computing Systems*, 1–5. <https://doi.org/10.1145/3491101.3503731>
- TallBear, K. (2013). *Native American DNA: Tribal belonging and the false promise of genetic science*. University of Minnesota Press.
-

- Verbeek, P. (2008). Cyborg intentionality: Rethinking the phenomenology of human-technology relations. *Phenomenology and the Cognitive Sciences*, 7, 387–395. <https://doi.org/10.1007/s11097-008-9099-x>
- Williams, R. M., Boyd, L. E., & Gilbert, J. E. (2023). Counterventions: A reparative reflection on interventionist HCI. *CHI '23: Proceedings of the 2023 Conference on Human Factors in Computing Systems*, 1–11. <https://doi.org/10.1145/3544548.3581480>
- Williams, R. M., Ringland, K., Gibson, A., Mandala, M., Maibaum, A., & Guerreiro, T. (2021). Articulations toward a crip HCI. *Interactions*, 28(3), 28–37. <https://doi.org/10.1145/3458453>
- Wilson, S. R. (2017). Selection of methodology. In M. Allen (Ed.), *The SAGE encyclopedia of communication research methods* (pp. 1020–1023). SAGE.
-



# Communication Style Adaptation in Human-Computer Interaction: An Empirical Study on the Effects of a Voice Assistant's Politeness and Machine-Likeness on People's Communication Behavior During and After the Interacting

Aike C. Horstmann<sup>1</sup> , Clara Strathmann<sup>1</sup> , Lea Lambrich<sup>1</sup> , and Nicole C. Krämer<sup>1</sup> 

<sup>1</sup> Department of Social Psychology: Media and Communication, University of Duisburg-Essen, Germany

## Abstract

Humans adapt their communication style when interacting with one another. With interactive technologies such as voice assistants taking over the role of an interaction partner, the question arises whether and to what extent humans also adapt to their communication style. The adaptation could have a grounding function, ensuring efficient communication with the current interaction partner, or be based on priming which could endure and influence subsequent interactions. In a pre-registered experimental lab study, 133 participants interacted with a voice assistant whose communication style varied regarding politeness (polite vs. non-polite) and machine-likeness (machine-like vs. natural). Participants' verbal behavior during and in a subsequent communication situation was analyzed. Politeness as well as machine-likeness adaptation was observed during the interaction but not afterward, supporting the grounding hypothesis. Furthermore, the adaptation process appears to be unconscious as the voice assistant's different communication styles did not affect conscious evaluations.

**Keywords:** voice assistant, communication styles, adaptation, politeness, machine-likeness

**CONTACT** Aike C. Horstmann  • [aike.horstmann@uni-due.de](mailto:aike.horstmann@uni-due.de) • University of Duisburg Essen • Bismarckstraße 120 • 47057 Duisburg, Germany

ISSN 2638-602X (print)/ISSN 2638-6038 (online)  
[www.hmcjournal.com](http://www.hmcjournal.com)



Copyright 2024 Authors. Published under a Creative Commons Attribution 4.0 International (CC BY-NC-ND 4.0) license.

**Acknowledgments:** The study has been funded by the Volkswagen Foundation in the project “IMPACT: The implications of conversing with intelligent machines in everyday life on people’s beliefs about algorithms, their communication behavior and their relationship building”, project number 95 836.

**Authors Note:** We have no known conflict of interest to disclose.

## Introduction

With the progressing prevalence of interactive technologies, various questions regarding their effects on human interaction behaviors arise. Particularly, voice-activated intelligent personal assistants (in the following referred to as voice assistants), which are integrated in smartphones or smart speakers such as Amazon’s Alexa and Apple’s Siri, are well-known and widespread (López et al., 2018). Voice assistants are used for diverse services such as playing music, setting an agenda or to do-lists, retrieving news, weather information, or directions. They are operated via voice commands, which is considered intuitive and related to natural human behavior (López et al., 2018). In human-human interaction, people tend to adapt their verbal and nonverbal behavior to match the person they are interacting with (Burgoon et al., 1995; Giles et al., 1991). Since voice assistants are operated in an interactive way which resembles a human-human interaction, the question arises to what extent these adaptation processes also take place here. As previous research has shown, people tend to react socially to interactive technologies, for instance by applying politeness or responding to flattering (Nass & Moon, 2000; Reeves & Nass, 1996). In line with that, there is first evidence that people adapt their communication behavior when talking to machines (L. Bell et al., 2003; Branigan & Pearson, 2006; Branigan et al., 2003; Branigan et al., 2010; Oviatt et al., 1998; Suzuki & Katagiri, 2007). This could be due to a grounding function, where mutual understanding is established via adaptation to ensure efficient communication (Clark & Brennan, 1991), or be based on priming, where a certain communication style may activate contextual interaction scripts leading to an adaptation of that communication style (Hoey, 2007). While a mere grounding function would not influence subsequent interactions, a priming effect could entail that voice assistant users carry over negative communication patterns (e.g., a more non-polite or machine-like communication style) into human-human conversations. This would have crucial implications for the dialogue design of these devices. To shed further light on the question of whether and to what extent the communication style of a voice assistant has the potential to affect the communication style of the human interacting with it, we investigate potential adaptation processes *during* and *after* the interaction. The voice assistant’s perceived competence and sociability are considered as influencing factors to receive further evidence why people adapt their communication behavior to machines (Branigan & Pearson, 2006; Riordan et al., 2014).

At the beginning of this paper, we review related work on alignment processes in human-human as well as human-machine interaction, concluding with our hypotheses and research questions. Next, we describe the methods of our experimental lab study and the results we obtained from our analyses. We conclude by discussing the findings, elaborating their importance for the field, and giving an outlook for future research.

---



## Related Work

People adapt to each other verbally and nonverbally (e.g., proximity, gaze, smiling, silences, response latency, utterance length) as well as behaviorally (e.g., helping, global intimacy, affect, resources; Burgoon et al., 1993). These adaptation processes were given many names such as accommodation, alignment, convergence, congruence, synchrony, or reciprocity (Giles et al., 1991). According to the Communication Accommodation Theory (CAT), the aim is to “index and achieve solidarity” via “realignments of patterns of code or language selection” (Giles et al., 1991, p. 2). The Interaction Adaptation Theory (IAT) by Burgoon et al. (1995) describes the process as matching or synchronizing the timing of behavior. In the current work, we focus on communication accommodation in terms of the alignment of verbal aspects which we refer to as communication style adaptation.

## Communication Style Adaptation in Humans

The adaptation of communication styles is argued to build the basis for successful social communication situations (Pickering & Garrod, 2006). It is assumed to take place automatically and unconsciously with the goal of establishing a joint semantic concept for the persons involved in the interaction, which reduces the need to exchange explicit information (Garrod & Anderson, 1987). For instance, in a study by Garrod and Anderson (1987), participants were tasked to navigate a labyrinth together. Here, if one speaker described where they were located by saying “third row two along,” the other would typically use a subsequent description such as “second row three along.” In several studies, communication partners were observed to converge in sentence structure and choice of words which facilitates appropriate reference to something or someone without precise knowledge of the partner or their experiences (Bock, 1986; Branigan et al., 2000; Brennan & Clark, 1996).

There are two prominent theories offering explanations for these processes: grounding and priming (Riordan et al., 2014). Grounding stands for the establishment of mutual knowledge and reciprocal understanding to facilitate an efficient conversation (Clark & Brennan, 1991). Consequently, interactions are seen as collaborations between speaker and listener: the listener indicates understanding and the speaker considers the listener’s knowledge, beliefs, and abilities and monitors their understanding (A. Bell, 1984; Riordan et al., 2014). For instance, when speaking to children, people tend to use simpler vocabulary and shorter sentences (Riordan et al., 2014). Other researchers argue that priming may offer a better explanation for the observed alignment processes (for a review, see Ferreira & Bock, 2006). Following this argumentation, verbal and nonverbal alignment result from the interlocutors priming each other (Riordan et al., 2014). A speaker activates, for instance, an expression for a listener who then uses the same or a closely related expression when becoming the speaker. Following priming theory, the usage of certain words, sentence structures, and language style will activate certain contextual interaction scripts (Hoey, 2007) which may remain activated in a subsequent interaction with a different interlocutor. While these adaptation processes are well-investigated in the human-human context, the growing prevalence of social and communicative technologies, such as voice assistants, unveils new types of interaction partners which may also elicit communication style adaptation.

## Communication Style Adaptation in Human-Machine Interaction

As we know from media equation theory (Nass & Moon, 2000; Reeves & Nass, 1996), minimal social cues such as interactivity, natural speech, and the fulfillment of a social role are sufficient to elicit unconscious social reactions toward machines that are typical for human-human interactions. In this vein, research has shown that people also adapt their behavior to nonhuman interaction partners such as computers (cf. Fogg & Nass, 1997), robots (cf. Lorenz et al., 2016; Sandoval et al., 2016), and virtual agents (cf. Krämer et al., 2013). This is for instance examined in the context of reciprocal self-disclosure (von der Pütten et al., 2010), establishment of rapport (feeling of being “in sync”; Huang et al., 2011), mimicry of facial expressions (Krämer et al., 2013), and game and negotiation strategies (Asher et al., 2012; Mell et al., 2018). Besides these behavioral and social adaptations, the convergence of communication behaviors is of particular interest when studying interactions between humans and machines. Previous research has shown that humans adapt to computers in terms of their speech rate (L. Bell et al., 2003), their loudness of speech and response latency (Suzuki & Katagiri, 2007), their syntax (Branigan et al., 2003), as well as their linguistic alternation, articulation, speech segments, pauses, use of final falling contours, and linguistic variability (Oviatt et al., 1998). Lexical alignment was observed when the computer deliberately used different terms than the human interaction partner (Brennan, 1996). Branigan et al. (2010) conclude in their review that communication style adaptation occurs in interactions with machines, often to an even greater extent than in interactions with other humans.

The communication style of a machine mostly differs regarding two aspects: its politeness and machine-likeness. Politeness is a universal social norm and a powerful mechanism that was developed to facilitate efficient interactions between individuals (Ribino, 2023). Most interactive devices employ politeness strategies as they help to facilitate the perception of trustworthiness and reliability as well as the social acceptance of the device (see review by Ribino, 2023). Machine-likeness can relate to several aspects such as appearance, behavior, and communication style, which varies extensively in different types of interactive devices. A machine-like communication style was found to lead to less perceived social presence, competence, and warmth (Dautzenberg et al., 2021; Kim et al., 2021) and further to inhibit or even suppress social responses (Lee, 2010). Therefore, a more natural communication style is often strived for.

Referring back to the two theories that offer explanations for communication style adaptation observed in human-human interaction (grounding and priming theory, Riordan et al., 2014), we aim to investigate which theory may be adequate to explain communication style adaptation that occurs in human-machine interaction. Some researchers argue that the observed adaptation in human-machine interaction could be a case of audience design (Riordan et al., 2014), which goes in line with the grounding theory. According to this assumption, participants adapt by using words and phrases the computer uses to facilitate an efficient conversation (Riordan et al., 2014). In case the adaptation has a grounding function to ensure efficient communication with the current interaction partner (Branigan & Pearson, 2006), it will occur *during* the current communication situation. However, considering the priming theory, it could also be that the machine’s communication style

---

activates contextual interaction scripts (Hoey, 2007). By using a certain communication style, such as politeness or machine-likeness, this style is also triggered for the human interaction partner who then applies it in the following. The activated communication style may remain activated in a subsequent interaction with a different interlocutor. Consequently, the adapted communication style may also be observable *after* interacting with a machine (Ferreira & Bock, 2006; Hoey, 2007; Pickering & Garrod, 2004; Riordan et al., 2014).

In sum, previous research offers substantial evidence for the occurrence of communication style adaptation when interacting with machines such as voice assistants. According to the grounding theory, this will take place *during* the interaction to ensure an efficient communication (Riordan et al., 2014). Considering the priming theory, the machine's communication style activates a certain communication style of the listening human, which then stays activated and can be observed *after* the interaction. Since politeness and machine-likeness are relevant aspects of the communication style of machines, we postulate the following:

**H1:** Individuals adapt to a voice assistant's communication style *during* the interaction: Individuals interacting with a voice assistant that displays (a) a *polite* (vs. non-polite) communication style will use a more polite communication style and (b) a *machine-like* (vs. natural) communication style will use a more machine-like communication style.

**H2:** Individuals adapt to a voice assistant's communication style *after* the interaction: Individuals who interacted with a voice assistant that displays (a) a *polite* (vs. non-polite) communication style will subsequently use a more polite communication style and (b) a *machine-like* (vs. natural) communication style will subsequently use a more machine-like communication style.

The results by Branigan et al. (2010) suggest that in many cases communication style adaptation occurs to a greater extent in human-machine than in human-human interaction. This is explained with people's goal of establishing mutual knowledge and reciprocal understanding to facilitate an efficient conversation (Clark & Brennan, 1991). When the machine-likeness is more salient, this could trigger concerns regarding the extent to which the machine has the knowledge and ability that is needed for an easy-flowing and successful communication. This could likely lead to a stronger effort to adapt the own communication style to that of the machine. Consequently, we hypothesize that individuals more strongly adapt to the voice assistant's communication style, in this case its politeness, the more machine-like it communicates:

**H3:** There is an *interaction effect* of the voice assistant's polite and machine-like communication style: Individuals interacting with a voice assistant displaying a *polite* (vs. non-polite) communication style will use a more polite communication style when the voice assistant displays a *machine-like* (vs. natural) communication style.

## Perception of the Voice Assistant Influencing the Communication Style Adaptation

Another indication that communication style adaptation behavior is caused by grounding might be delivered when investigating the voice assistant's perceived competence and sociability. Adaptation was found to occur to a greater extent when interacting with computers compared to humans, likely with the goal to enhance communicative success as a response to the perceived limited capabilities of computers (Branigan et al., 2010). Supporting this theory, participants were shown to adapt their communication style more strongly to a computer that is evaluated less competent (Pearson et al., 2006). Since interaction behaviors such as machine-likeness and politeness influence the evaluation and liking of artificial entities as interaction partners (Horstmann & Krämer, 2020, 2022), the voice assistant's communication style is assumed to influence how sociable and competent it is perceived. While the perception of low competence appears to be clearly linked to stronger communication style adaptation behavior (Pearson et al., 2006), the effect which the voice assistant's perceived sociability may have is less clear. Against this background, the following hypothesis concerning the voice assistant's perceived competence and the following research question concerning its perceived sociability are formulated:

**H4:** A voice assistant's communication style influences individuals' communication style adaptation during the interaction via the voice assistant's perceived *competence*: (a) a polite vs. non-polite communication style is perceived more competent leading to less adaptation; (b) a machine-like vs. natural communication style is perceived less competent leading to more adaptation.

**RQ1:** Do individuals interacting with a voice assistant displaying a polite vs. non-polite and a machine-like vs. natural communication style show differences in their communication style adaptation depending on how they perceive the voice assistant's *sociability*?

## Method

An experimental lab study with a 2 (machine-like vs. natural communication style) × 2 (polite vs. non-polite communication style) between-subject design was conducted. We preregistered the study at the OSF platform (<https://osf.io/m8rha>) and the local ethics committee approved the study's procedure. Supplementary study material (experimenter instructions, interaction script, questionnaire, codebook) can be found online: <https://osf.io/grqn4/>.

## Experimental Manipulations and Procedure

First, a cover story was presented explaining that the participants were supposed to test the made-up interaction program SAM running on an Amazon Echo Dot smart speaker. After the procedure and alleged purpose were explained and participants gave their written consent, they filled out some pre-questionnaires on a laptop (sociodemographic

---

background, previous experiences with voice assistants). This was followed by an introduction of the interaction program SAM that was allegedly running on the smart speaker. In reality, pre-recorded audio files were played simulating an interaction with SAM. One Echo Dot was placed in the middle of a table on a black box in front of the participants (see Figure 1). A second Echo Dot was placed underneath the box for the audio output (since it was not possible to turn on the well-known blue light of the device and use it as a speaker for audio output simultaneously). Next, the experimenter pretended to start the interaction program SAM and asked the participants to wait a few seconds. The experimenter left the room, allegedly so that the participants would not feel observed during the interaction. From the adjacent room, the experimenter controlled the voice assistant's output by using a webcam that was installed in the lab (see Figure 1, top right corner) to see and hear the participant and let the voice assistant react accordingly (Wizard of Oz design; see Dahlbäck et al., 1993). The webcam was justified by explaining that in case of errors the developers of SAM could track what went wrong.

**FIGURE 1** Experimental Setup With an Amazon Echo Dot Placed on and One Placed Under a Black Box, a Webcam, and the Cooking Requisites Which Are Needed for the First Interaction Task



Photos taken and owned by authors.

In the first part of the interaction, SAM walked the participants through a salad recipe which they followed by using cooking requisites (see Figure 1). In the following interaction part, the voice assistant asked about dietary restrictions and preferences, allergies, intolerances, as well as the preferred food preparation difficulty and time with the alleged aim to recommend suitable recipes in the future. The cooking task and personalization of recipe suggestions was chosen to represent a plausible everyday application scenario for a voice assistant in the private sphere. In both parts, the voice assistant's communication style was manipulated regarding politeness and machine-likeness (see next subchapter). Then, participants were sent back to the laptop where they were asked to imagine a person to whom they are explaining the recipe from part one and to record how they would go through the recipe step by step. The aim was to measure whether the communication style adaptation lasts beyond the interaction situation with the voice assistant. This was followed by questionnaires including participants' evaluation of the voice assistant's competence and sociability and manipulation checks (and personality variables, which were not used for the current analyses). Upon completion, the experimenter returned to the lab, debriefed the participants, and compensated them for their time with money or course credits.

### ***Politeness and Machine-Likeness Manipulation***

In the polite conditions, the voice assistant used *verbal markers* of politeness (e.g., “please” and “thank you”) and *structural elements* of politeness (e.g., requests formulated as interrogatives versus imperatives; mitigating verbs, e.g., “would” and “could” versus forceful verbs, e.g., “must” and “have to”). For the machine-like conditions, the voice assistant used short and functional sentences with a repetitive structure and a limited range of vocabulary (e.g., saying “okay!” after each executed recipe step or “your answers have been processed” after each reply to the recipe recommendation questions), while lengthy and colloquial sentences in a varying structure and a larger vocabulary range were characteristic for the natural conditions (e.g., alternating between expressions like “thanks for your answer” or “I will remember that” when reacting to replies to the recipe recommendation questions; *structural elements*). Furthermore, there were no hints to having own feelings or intentions in the machine-like conditions, which was different for the natural conditions (e.g., for intentions: “. . . will be taken into account” vs. “I will take into account . . . ,” or by saying “I am glad to meet you” vs. “Now we get to know each other”; *verbal markers*). The entire script can be viewed in the online supplementary material (<https://osf.io/8mn6g>).

### **Sample**

The software G\*Power was used to conduct a power analysis (.80 power, medium effect size of  $f^2(V) = 0.0625$ , standard .05 alpha error probability). The results recommend a minimum of 113 respondents. In total, 137 participated in the experimental lab study of which four were excluded (failure of both attention checks, suspicious answering behavior, heavily restricted language skills), analyses were conducted with 133 participants; 85 of those stated to be female, 47 to be male, and one to be diverse. On average, participants were 23.15 years old, ranging from 18 to 35 ( $SD = 3.64$ ) years. Most of the participants reported to be students (93.2%) and to hold a university entrance level (79.7%) or university degree (18.8%). Most of the participants had interacted with a voice assistant before (84.2%), on



average with a medium frequency ( $M = 2.59$ ,  $SD = 1.32$ ; 1 = “very rarely” to 5 = “very often”) and a rather low intensity of use ( $M = 1.97$ ,  $SD = 1.04$ ; 1 = “little intensively” to 5 = “very intensively”).

## Measurements

### Communication Style Adaptation

To analyze whether and to what extent the participants adapt their communication style to that of the voice assistant, we analyzed how often participants used *structural elements* and *verbal markers* of politeness or machine-likeness, respectively, during and after the interaction. This follows the theoretical basis by Bunz and Campbell (2004) that was also used to design the voice assistant’s different communication styles. The analyses were conducted using the coding software MAXQDA 2022 (the final codebook is available in the online supplementary files; <https://osf.io/yf2j7>). *Structural elements* of politeness include *opening and closing acts* (greeting and saying goodbye) and whether requests are formulated as *imperatives* or *interrogatives* (e.g., “Continue with the next step!” vs. “Can you continue with the next step?”; Bunz & Campbell, 2004; De Jong et al., 2008). *Verbal markers* of politeness comprise *thanking acts*, *saying please*, expressions of *appreciation* (e.g., “I would appreciate if you could repeat the last step”), *flattering* (e.g., “You explained that very well”), *redressing hedges* (words or phrases that diminish the face-threatening force of a speech act, e.g., “I just want to ask if we could continue”), and the use of *mitigating verbs* such as *could*, *would* (like to), and *can* instead of *forceful verbs* such as *must*, *have to*, *need to*, and *want to* (e.g., “I want easy recipes” vs. “I would like easy recipes”; Bunz & Campbell, 2004; De Jong et al., 2008). Additionally, we analyzed whether participants indicated to consider their interlocutor as *social entity* with social needs, for instance, by suggesting group membership (De Jong et al., 2008) or attempts to reduce the other’s uncertainty (e.g., replying to “Let me know when you’re done” with “Will do”).

*Structural elements* of machine-likeness include participants’ *word count* (number of words they used; Hoffmann et al., 2020), *direct address* of their dialogue partner (e.g., “You can start”; Hoffmann et al., 2020), and lexical diversity measured via *Type-Token-Ratio*, the ratio of different words (types) to total words (tokens); Templin, 1957). For *verbal markers* of machine-likeness we checked the communication style for *functionality* (short, functional expressions, e.g., “No meat”) in contrast to *verbosity* (long, copious sentences conveying more information than needed, e.g., “I do not really like meat, so I think I would like recipes that are vegetarian”) and for *list structures* (e.g., “One: no meat, two: no mushrooms, three: spicy”; Hoffmann et al., 2020; Horstmann et al., 2018). Furthermore, expressions suggesting *intentionality* (intentions, thoughts, and opinions, e.g., “no meat because it’s bad for the environment”) and *personal preferences* (e.g., “I’d prefer no meat”) were considered (Horstmann et al., 2018).

### Voice Assistant Evaluation and Manipulation Checks

The voice assistant’s perceived *competence* was assessed via adapted items of the Task Attraction subscale (5 items; e.g., “The voice assistant would be a poor problem solver with regard to speech-based interaction”;  $\alpha = 0.68$ ) of the Interpersonal Attraction Scale (IAS;



McCroskey & McCain, 1974; 1 = “strongly disagree” to 5 = “strongly agree”) and a collection of adjectives from Horstmann and Krämer (2022) rated on a five-point semantical differential (10 items; e.g., “incapable–capable”;  $\alpha=0.83$ ). The voice assistant’s perceived *sociability* was measured with the Social Attraction subscale of the IAS (5 items, e.g., “I think the voice assistant could be a friend of mine”;  $\alpha=0.74$ ) and another collection of adjectives from Horstmann and Krämer (2022; 15 items, e.g., “cold–warm”;  $\alpha=0.87$ ).

To check the *success of the manipulations*, participants were asked to rate the voice assistant’s expressions as either 1 = “rather non-polite,” 2 = “completely neutral,” or 3 = “rather polite” and as either 1 = “rather machine-like,” 2 = “completely neutral,” or 3 = “rather natural.” An ANOVA revealed that the voice assistant displaying a machine-like compared to a natural communication style was perceived as more machine-like ( $F(1, 131) = 4.12$ ,  $p = .044$ ,  $\eta_p^2 = 0.03$ ; machine-like:  $M = 1.61$ ,  $SD = 0.74$ ; natural:  $M = 1.38$ ,  $SD = 0.58$ ). The voice assistant displaying a polite compared to a non-polite communications style was not perceived significantly more polite ( $F(1, 131) = 1.79$ ,  $p = .183$ ,  $\eta_p^2 = 0.01$ ; polite:  $M = 2.84$ ,  $SD = 0.44$ ; non-polite:  $M = 2.72$ ,  $SD = 0.55$ ).

## Results

The statistical analyses were conducted with IBM SPSS Statistics 29 including the PROCESS macro v4.3, significance was determined using the standard  $p < .05$  criterium.

### Communication Style Adaptation in Human-Machine Interaction

To investigate **H1** (Individuals adapt to a voice assistant’s communication style *during* the interaction) and **H3** (There is an interaction effect of the voice assistant’s politeness and machine-likeness communication style), we conducted a MANOVA with the voice assistant’s communication styles (*polite vs. non-polite* and *machine-like vs. natural*) as factors and the participant’s communication style (*structural elements* and *verbal markers* of politeness/machine-likeness) *during* the interaction as dependent variable. Using Pillai’s trace, there was a significant main effect of the *politeness* of the voice assistant’s communication style on the *politeness* of the participants’ communication style,  $V = 0.64$ ,  $F(18, 99) = 9.66$ ,  $p < .001$ . Separate univariate ANOVAs on the different outcome variables revealed a significant effect on participant’s usage of *redressing hedges*,  $F(1, 116) = 8.93$ ,  $p = .003$ ,  $\eta_p^2 = 0.07$ , and *thanking acts*,  $F(1, 116) = 111.55$ ,  $p < .001$ ,  $\eta_p^2 = 0.49$ , their consideration of the voice assistant as *social entity*,  $F(1, 116) = 14.17$ ,  $p < .001$ ,  $\eta_p^2 = 0.11$ , and their usage of *opening and closing acts*,  $F(1, 116) = 10.94$ ,  $p = .001$ ,  $\eta_p^2 = 0.09$ . There was no significant effect on the participants’ usage of *mitigating verbs*,  $F(1, 116) = 1.84$ ,  $p = .178$ ,  $\eta_p^2 = 0.02$ , *forceful verbs*,  $F(1, 116) = 3.08$ ,  $p = .082$ ,  $\eta_p^2 = 0.03$ , the word *please*,  $F(1, 116) = 0.66$ ,  $p = .419$ ,  $\eta_p^2 = 0.01$ , *flattering*,  $F(1, 116) = 2.94$ ,  $p = .089$ ,  $\eta_p^2 = 0.03$ , *interrogatives*,  $F(1, 116) = 2.00$ ,  $p = .160$ ,  $\eta_p^2 = 0.02$ , and *imperatives*,  $F(1, 116) = 2.37$ ,  $p = .126$ ,  $\eta_p^2 = 0.02$ . Expressions of *appreciation* remained uncoded and were therefore not considered. For descriptive values, see Table 1. **H1a** is partly supported.

**TABLE 1** Descriptive Values for the Structural Elements and Verbal Markers of Politeness During and After the Interaction With the Voice Assistant

	During the interaction					After the interaction				
	Polite	Non-polite	Natural	Machine-like	Total	Polite	Non-polite	Natural	Machine-like	Total
<b>Structural elements</b>										
Interrogative M (SD)	0.34 (0.92)	0.59 (1.03)	0.57 (1.03)	0.33 (0.91)	0.46 (0.98)	0.02 (0.13)	0.00 (0.00)	0.02 (0.13)	0.00 (0.00)	0.01 (0.09)
Imperative M (SD)	0.92 (1.28)	1.21 (1.52)	0.46 (0.74)	1.72 (1.66)	1.06 (1.40)	0.11 (0.57)	0.10 (0.55)	0.16 (0.68)	0.05 (0.39)	0.11 (0.56)
Opening/ closing acts M (SD)	0.37 (0.58)	0.72 (0.64)	0.49 (0.56)	0.60 (0.70)	0.54 (0.63)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
<b>Verbal markers</b>										
Mitigating verbs M (SD)	1.42 (1.49)	1.10 (1.20)	1.49 (1.31)	1.02 (1.38)	1.27 (1.36)	0.06 (0.30)	0.07 (0.26)	0.08 (0.27)	0.05 (0.29)	0.07 (0.28)
Forceful verbs M (SD)	0.08 (0.33)	0.26 (0.64)	0.29 (0.63)	0.04 (0.27)	0.17 (0.51)	0.70 (1.19)	0.67 (1.37)	0.68 (1.33)	0.69 (1.22)	0.69 (1.27)
Redressing hedges M (SD)	1.24 (1.39)	0.64 (0.91)	1.24 (1.37)	0.63 (0.94)	0.95 (1.22)	0.43 (0.67)	0.28 (0.70)	0.33 (0.60)	0.38 (0.77)	0.36 (0.68)
Appreciation M (SD)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Thanking acts M (SD)	1.42 (0.88)	0.10 (0.36)	0.81 (1.00)	0.75 (0.89)	0.78 (0.95)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Saying please M (SD)	0.26 (0.54)	0.17 (0.53)	0.16 (0.41)	0.28 (0.65)	0.22 (0.54)	0.00 (0.00)	0.02 (0.13)	0.00 (0.00)	0.02 (0.13)	0.01 (0.09)
Flattering M (SD)	0.00 (0.00)	0.07 (0.32)	0.04 (0.18)	0.04 (0.27)	0.03 (0.22)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Consider. as social entity M (SD)	2.61 (2.04)	1.48 (1.35)	2.49 (2.19)	1.60 (1.16)	2.07 (1.83)	0.38 (0.91)	0.41 (1.06)	0.46 (1.08)	0.33 (0.87)	0.40 (0.98)

Using Pillai's trace, there was a significant main effect of the *machine-likeness* of the voice assistant's communication style on the *machine-likeness* of the participants' communication style,  $V = 0.54$ ,  $F(18, 99) = 6.48$ ,  $p < .001$ . Separate univariate ANOVAs on the different outcome variables revealed a significant effect on participant's disclosure of *personal preferences*,  $F(1, 116) = 42.57$ ,  $p < .001$ ,  $\eta_p^2 = 0.27$ , the *functionality* of their communication style,  $F(1, 116) = 19.29$ ,  $p < .001$ ,  $\eta_p^2 = 0.14$ , their *Type-Token-Ratio*,  $F(1, 116) = 8.71$ ,  $p = .004$ ,  $\eta_p^2 = 0.07$ , and their *word count*,  $F(1, 116) = 15.68$ ,  $p < .001$ ,  $\eta_p^2 = 0.12$ . The effect of the voice assistant's machine-likeness was not significant regarding participants' disclosure of *intentionality*,  $F(1, 116) = 2.96$ ,  $p = .088$ ,  $\eta_p^2 = 0.03$ , their usage of *list-style* communication,  $F(1, 116) = 2.39$ ,  $p = .125$ ,  $\eta_p^2 = 0.02$ , their *verbosity*,  $F(1, 116) = 3.19$ ,  $p = .077$ ,

**TABLE 2** Descriptive Values for the Structural Elements and Verbal Markers of Machine-Likeness During and After the Interaction With the Voice Assistant

	During the interaction					After the interaction				
	Polite	Non-polite	Natural	Machine-like	Total	Polite	Non-polite	Natural	Machine-like	Total
<b>Structural elements</b>										
Direct address	1.21	1.47	1.56	1.08	1.33	0.94	1.05	1.17	0.79	0.99
M (SD)	(1.43)	(1.66)	(1.42)	(1.64)	(1.54)	(2.09)	(2.70)	(2.57)	(2.18)	(2.39)
Word count	67.58	61.88	75.89	52.60	64.83	67.48	65.12	67.59	65.00	66.35
M (SD)	(33.86)	(34.81)	(37.28)	(25.89)	(34.29)	(20.09)	(20.88)	(21.60)	(19.16)	(20.42)
Type-Token-Ratio	0.71	0.69	0.68	0.73	0.70	0.65	0.62	0.63	0.64	0.64
M (SD)	(0.08)	(0.08)	(0.07)	(0.08)	(0.08)	(0.09)	(0.08)	(0.08)	(0.09)	(0.09)
<b>Verbal markers</b>										
Pers. preferences	1.87	1.62	2.48	0.95	1.75	0.00	0.00	0.00	0.00	0.00
M (SD)	(1.51)	(1.49)	(1.50)	(1.01)	(1.50)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Intentionality	0.16	0.16	0.22	0.09	0.16	0.03	0.03	0.03	0.03	0.03
M (SD)	(0.41)	(0.45)	(0.52)	(0.29)	(0.43)	(0.18)	(0.18)	(0.18)	(0.18)	(0.18)
List structure	0.52	0.67	0.70	0.47	0.59	0.11	0.09	0.06	0.14	0.10
M (SD)	(0.74)	(0.85)	(0.87)	(0.68)	(0.79)	(0.32)	(0.28)	(0.25)	(0.35)	(0.32)
Verbosity	0.55	0.41	0.63	0.32	0.48	0.10	0.17	0.19	0.07	0.13
M (SD)	(1.07)	(0.92)	(1.26)	(0.54)	(1.00)	(0.30)	(0.38)	(0.40)	(0.26)	(0.34)
Functionality	7.34	8.24	6.13	9.60	7.78	0.33	0.38	0.35	0.36	0.36
M (SD)	(4.85)	(4.49)	(4.33)	(4.40)	(4.68)	(0.48)	(0.49)	(0.48)	(0.49)	(0.48)

$\eta_p^2 = 0.03$ , and their *direct address* of the voice assistant,  $F(1, 116) = 2.55$ ,  $p = .113$ ,  $\eta_p^2 = 0.02$ . For descriptive values, see Table 2. **H1b** is partly supported. Using Pillai's trace, there was no significant interaction effect of the *politeness* and *machine-likeness* of the voice assistant's communication style on the participants' communication style,  $V = 0.13$ ,  $F(18, 99) = 0.80$ ,  $p = .702$ . Therefore, **H3** needs to be rejected.

To explore **H2** (individuals adapt to a voice assistant's communication style *after* the interaction), another MANOVA was conducted with the voice assistant's communication styles as factors and the participant's communication style that was assessed *after* the interaction as dependent variable. Using Pillai's trace, there was no significant main effect of the voice assistant's communication style, neither of *politeness*,  $V = 0.10$ ,  $F(14, 104) = 0.78$ ,  $p = .687$ , nor of *machine-likeness*,  $V = 0.10$ ,  $F(14, 104) = 0.85$ ,  $p = .614$ , on the participants' communication style. Consequently, **H2** needs to be rejected. For descriptive values, see Table 1 (*politeness*) and Table 2 (*machine-likeness*).

Summing up, *during* the interaction with a voice assistant that displays a polite compared to a non-polite communication style, individuals display less *opening and closing acts* (structural elements of *politeness*), more *redressing hedges*, more *thanking acts*, and more *consideration as social entity* (verbal markers of *politeness*). When interacting with a voice assistant that displays a machine-like compared to a natural communication style, individuals disclose fewer *personal preferences* and less *intentionality*, use less *verbosity*, and more *functionality* (verbal markers of *machine-likeness*). They also display a lower *word count*, but

**TABLE 3** Descriptive Values of the Evaluation of the Voice Assistant

	Polite		Non-Polite		Natural		Machine-Like		Total	
	M	SD	M	SD	M	SD	M	SD	M	SD
Task Attraction (IAS)	3.85	0.56	3.76	0.83	3.78	0.70	3.83	0.70	3.81	0.70
Competence	3.45	0.58	3.41	0.64	3.43	0.61	3.43	0.61	3.43	0.61
Social Attraction (IAS)	2.24	0.77	2.29	0.89	2.31	0.86	2.22	0.80	2.27	0.83
Sociability	3.67	0.55	3.56	0.54	3.73	0.58	3.50	0.50	3.62	0.55

a higher *Type-Token-Ratio* (structural elements of *machine-likeness*). There is no interaction effect of the voice assistant's *politeness* and *machine-likeness* on their human interaction partners' communication style *during* the interaction and no effect of the voice assistant's communication style on individuals' communication style *after* the interaction.

### Perception of the Voice Assistant Influencing the Communication Style Adaptation

To investigate **H4** (A voice assistant's communication style influences individuals' communication style adaptation during the interaction via the voice assistant's perceived *competence*) and **RQ1** (Do individuals [ . . . ] show differences in communication style adaptation depending on how they evaluated the voice assistant's *sociability*?), we first conducted a MANOVA to test for the influence of the voice assistant's communication style on its perceived *competence* (task attraction, competence) and *sociability* (social attraction, sociability). Pillai's trace revealed no significant effect of the voice assistant's *machine-likeness*,  $V = 0.07$ ,  $F(4, 126) = 2.32$ ,  $p = .061$ , and no significant effect of its *politeness*,  $V = 0.02$ ,  $F(4, 126) = 0.70$ ,  $p = .594$  (see Table 3 for descriptive values). Since we found no significant effect of voice assistant's communication style on its perceived competence, **H4** needs to be rejected and **RQ1** needs to be negated. Summing up, the results suggest that neither the voice assistant's perceived competence nor sociability are influenced by its communication style.

## Discussion

Against the background of the rising prevalence of voice assistants, the main question of this paper was whether and to what extent individuals adapt their communication style to the communication style of a voice assistant, during and after the interaction with it. From previous research, two theories that are used to explain communication style adaptation processes are considered and further investigated in the current study: grounding and priming theory (Riordan et al., 2014). While grounding would be based on the aim to ensure an efficient communication with the current communication partner and therefore only take place during the interaction (Clark & Brennan, 1991), priming could endure and influence subsequent interactions (Ferreira & Bock, 2006). We therefore conducted a pre-registered lab study to record and analyze the communication style of 133 participants

during and after interacting with a voice assistant that displays a machine-like vs. natural and a polite vs. non-polite communication style.

### Communication Style Adaptation in Human-Machine Interaction

The results show that communication style adaptation takes place largely during the interaction, but not after. During the interaction, participants were observed to use more *redressing hedges* and more *thanking acts* while they also consider the voice assistant more as *social entity* when it displays a polite compared to a non-polite communication style. When interacting with a voice assistant that displays a machine-like compared to a natural communication style, individuals appear to adapt by using fewer *words*, disclosing fewer *personal preferences* (e.g., “I would like warm dishes” or “I do not like fish”) and *intentionality*, using fewer *verbose* and more *functional* expressions (e.g., one-word phrases).

Two findings contradicted what we expected: there were fewer *opening and closing acts* in the politeness compared to the non-politeness conditions and a higher *Type-Token-Ratio* (indicating a higher lexical diversity) in the machine-like compared to the natural conditions. The occurrence of *opening and closing acts* may have been influenced by the script’s design. For instance, in the polite conditions, the voice assistant concluded the recipe interaction with “Enjoy!” and the entire interaction with “Have a pleasant rest of the day!” Here, people might not have replied with goodbye, but rather thank you (coded as thanking act). In the non-polite conditions, it concluded by saying goodbye, which may have triggered saying goodbye in return resulting in more closing acts. An explanation for having a higher *Type-Token-Ratio* in the machine-like conditions could be that people functionally report their preferences (e.g., “Preferences: tomatoes, mushrooms, dislike: onions, garlic”), thus having few repeated words resulting in a higher Type-Token-Ratio, while users in the natural conditions might repeat sentence structures such as “I like tomatoes and mushrooms, and I don’t like onions and garlic” resulting in a lower Type-Token-Ratio.

The remaining results paint a clear picture of people adapting to a voice assistant’s politeness and machine-likeness *during* the interaction. These findings support the grounding theory (Bock, 1986; Branigan et al., 2000; Clark & Brennan, 1991). The reduction in word quantity, increased functionality, and decreased verbosity are in line with Riordan et al.’s (2014) idea of audience design, according to which users adopt expressions to fit the device’s perceived constraints. Evidence for the priming theory as an explanation for communication style adaptation (Ferreira & Bock, 2006; Riordan et al., 2014) could not be found as adaptations processed were only observed during the interaction with the voice assistant and not in a subsequent interaction with an imagined person. Furthermore, the hypothesized interaction effect leading to greater politeness adaptation when speaking to a voice assistant displaying a more machine-like communication style was not found. Thus, the politeness adaptation appears not to depend on the voice assistant’s machine-likeness. A potential explanation could be that politeness is a concept that runs automatically so that users adapt to it independent of the interaction partner’s perceived constraints. In future studies, it would be interesting to investigate whether other communication style aspects are affected by a technological interaction partner’s machine-likeness.

---

## Perception of the Voice Assistant Influencing the Communication Style Adaptation

The different communication styles did not affect how competent and sociable the voice assistant was perceived. An explanation could be that participants' communication style adaptation to that of the voice assistant might take place on an unconscious level and therefore does not influence how it is evaluated consciously. This is partly in line with the manipulation checks which revealed that, when asked directly, people were not fully aware of the voice assistant's polite versus non-polite communication style. Potentially, the consistently friendly tone of the voice assistant in all conditions may have led the participants to evaluate the voice assistant with the non-polite communication style—on a conscious level—as polite as well. Nevertheless, people in the polite conditions adapted to its communication style for instance by using more *redressing hedges* and more *thanking acts* which implies an unconscious communication style adaptation. We therefore argue that people do not deliberately process and evaluate the interaction partner's communication style before adapting to it. In other words, people register and adapt to a communication style automatically and do not make a conscious decision to accommodate. Since a behavioral change was expected but not necessarily a cognitive evaluation beforehand and significant differences in the participants' communication style depending on the voice assistant's communication style were measured, we are confident that the manipulation was successful. Considering that humans mindlessly treat machines socially when presented with human-like cues such as natural language (Nass & Brave, 2005; Nass & Moon, 2000; Reeves & Nass, 1996), our findings are also in line with the Media Equation Theory.

## Limitations and Future Research

There was no baseline condition which surveyed users' communication behavior without manipulating the voice assistants' communication style, which could have helped with putting the findings into context. Regarding the experimental setting, the focus of this study was on a task to be accomplished, a more openly designed social experience could deliver further insights. Furthermore, the communication with the imagined person was short and resulted in only a few codes, which may have restricted the measurement of communication style adaptation after the interaction with the voice assistant. For future studies, an interaction with a real person instead of an imagined one could be more effective. As in many studies, the sample consisted mainly of students and was conducted in a lab setting, therefore the generalizability of the results for other age groups and in the real world is limited. Especially the last aspect calls for future research. Children, for instance, should be looked at in detail, not least because of the prevailing worry that interactive devices could teach children impolite or machine-like communication behavior. As voice assistants neither require nor encourage politeness (Curry & Rieser, 2018) and even tend to misunderstand copious requests (e.g., including phrases such as “could you” or “if you don't mind”), children who still need to learn the rules of social communication could be particularly prone to adapting negatively connotated communication. While we did not find any evidence for politeness having a strong effect on users' communication behavior, it would be valuable to investigate how people behave over a longer period of interacting with



the device. Furthermore, adding a condition with a clearly impolite communication style could lead to different effects, particularly since this behavior is not common and therefore unexpected. Future research should also investigate whether there are circumstances under which people do not adapt or even diverge with their communication style as it has been observed in interactions between humans (e.g., Giles et al., 1991).

## Conclusion

Our aim was to investigate whether individuals adapt their communication style to a voice assistant's communication style in terms of politeness and machine-likeness and whether (if at all) the communication style adaptation only takes place during the interaction with the voice assistant or also in a subsequent interaction with an imagined person. In line with the grounding theory, which suggests that communication style adaptation serves the purpose of establishing and maintaining a successful communication, individuals were observed to adapt to the voice assistant's politeness as well as machine-likeness during the interaction with it but not in subsequent interactions with others. Furthermore, this adaptation process appears to take place unconsciously as the voice assistant's different communication styles did not affect how it was consciously evaluated.

## Author Biographies

**Aike Horstmann** (PhD, University of Duisburg-Essen) studied Applied Cognitive and Media Science at the University of Duisburg-Essen, followed by her doctoral studies in the field of human-robot/virtual agent-interaction with a focus on the effects on humans' perception of artificial entities. She received her PhD from the University of Duisburg-Essen in early 2021 and started as a senior research associate and project coordinator in the commercial sector. Since 2022, she continues to conduct research in the field of human-machine interaction as a postdoc at the department of Social Psychology: Media and Communication, University of Duisburg-Essen.

 <http://orcid.org/0000-0003-4693-1743>

**Clara Strathmann** (MSc, University of Duisburg-Essen) studied Applied Cognitive and Media Science at the University of Duisburg-Essen, where she was involved in research on voice assistants as a student assistant at the department of Social Psychology: Media and Communication. In September 2023, she started her doctoral studies in the field of privacy online with a focus on vulnerable user groups at the University of Duisburg-Essen.

 <http://orcid.org/0009-0003-6641-0168>

**Lea Lambrich** (BSc, University of Duisburg-Essen) studied Applied Cognitive and Media Science at the University of Duisburg-Essen. In 2022 she started working as student assistant at the Social Psychology: Media and Communication department focusing on research involving voice assistants and technology-mediated connectedness.

 <https://orcid.org/0009-0005-1358-437X>

---



**Nicole Krämer** (PhD, University of Cologne) is Full Professor of Social Psychology: Media and Communication at the University of Duisburg-Essen, Germany, and co-director of the Research Center Trustworthy Data Science and Security. She completed her PhD in Psychology at the University of Cologne, Germany, in 2001 and received the *venia legendi* for psychology in 2006. Dr. Krämer's research focuses on social psychological aspects of human-machine-interaction (especially social effects of robots and virtual agents) and computer-mediated-communication (CMC). She heads numerous projects that received third party funding. She served as Editor-in-Chief of the *Journal of Media Psychology* 2015–2017 and currently is Associate Editor of the *Journal of Computer Mediated Communication (JCMC)*.

 <http://orcid.org/0000-0001-7535-870X>

## References

- Asher, D. E., Zaldivar, A., Barton, B., Brewer, A. A., & Krichmar, J. L. (2012). Reciprocity and retaliation in social games with adaptive agents. *IEEE Transactions on Autonomous Mental Development*, 4(3), 226–238. <https://doi.org/10.1109/TAMD.2012.2202658>
- Bell, A. (1984). Language style as audience design. *Language in Society*, 13(2), 145–204. <https://doi.org/10.1017/S004740450001037X>
- Bell, L., Gustafson, J., & Heldner, M. (2003). Prosodic adaptation in human–computer interaction. In M.-J. Solé, D. Recasens, & J. Romero (Eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 2453–2456). Causal Productions.
- Bock, J. (1986). Syntactic persistence in language production. *Cognitive Psychology*, 18(3), 355–387. [https://doi.org/10.1016/0010-0285\(86\)90004-6](https://doi.org/10.1016/0010-0285(86)90004-6)
- Branigan, H. P., & Pearson, J. (2006). Alignment in human-computer interaction. In K. Fischer (Ed.), *Report Series of the Transregional Collaborative Research Center SFB/TR 8. How People Talk to Computers, Robots, and Other Artificial Communication Partners* (pp. 140–156). Universität Bremen.
- Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition*, 75(2), B13–B25. [https://doi.org/10.1016/S0010-0277\(99\)00081-5](https://doi.org/10.1016/S0010-0277(99)00081-5)
- Branigan, H. P., Pickering, M. J., McLean, J. F., & Nass, C. (2003). Syntactic alignment between computers and people: The role of belief about mental states. In R. Altermann & D. Kirsch (Eds.), *Proceedings of the Twenty-fifth Annual Conference of the Cognitive Science Society* (pp. 186–191). Lawrence Erlbaum Associates, Inc.
- Branigan, H. P., Pickering, M. J., Pearson, J., & McLean, J. F. (2010). Linguistic alignment between people and computers. *Journal of Pragmatics*, 42(9), 2355–2368. <https://doi.org/10.1016/j.pragma.2009.12.012>
- Brennan, S. E. (1996). Lexical entrainment in spontaneous dialog. In *Proceedings of the 1996 International Symposium on Spoken Dialogue (ISSD-96)* (pp. 41–44). Acoustical Society of Japan.
- Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6), 1482–1493. <https://doi.org/10.1037/0278-7393.22.6.1482>

- Bunz, U., & Campbell, S. W. (2004). Politeness accommodation in electronic mail. *Communication Research Reports*, 21(1), 11–25. <https://doi.org/10.1080/08824090409359963>
- Burgoon, J. K., Dillman, L., & Stern, L. A. (1993). Adaptation in dyadic interaction: Defining and operationalizing patterns of reciprocity and compensation. *Communication Theory*, 3(4), 295–316. <https://doi.org/10.1111/j.1468-2885.1993.tb00076.x>
- Burgoon, J. K., Stern, L. A., & Dillman, L. (1995). *Interpersonal adaptation: Dyadic interaction patterns*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511720314>
- Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on Socially Shared Cognition: Revised Papers Presented at a Conference* (pp. 127–149). American Psychological Association. <https://doi.org/10.1037/10096-006>
- Curry, A. C., & Rieser, V. (2018). #MeToo Alexa: How conversational systems respond to sexual harassment. In M. Alfano, D. Hovy, M. Mitchell, & M. Strube (Eds.), *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing* (pp. 7–14). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W18-0802>
- Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of Oz studies—Why and how. *Knowledge-Based Systems*, 6(4), 258–266. [https://doi.org/10.1016/0950-7051\(93\)90017-n](https://doi.org/10.1016/0950-7051(93)90017-n)
- Dautzenberg, P. S. C., Vos, G. M. I., Ladwig, S., & Rosenthal-von der Putten, A. M. (2021). Investigation of different communication strategies for a delivery robot: The positive effects of humanlike communication styles. In *Proceedings of the 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)* (pp. 356–361). IEEE. <https://doi.org/10.1109/ro-man50785.2021.9515547>
- De Jong, M., Theune, M., & Hofs, D. (2008). Politeness and alignment in dialogues with a virtual guide. In L. Padgham & D. Parkes (Eds.), *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems* (pp. 207–214). ACM; AAI.
- Ferreira, V. S., & Bock, K. (2006). The functions of structural priming. *Language and Cognitive Processes*, 21(7–8), 1011–1029. <https://doi.org/10.1080/01690960600824609>
- Fogg, B. J., & Nass, C. (1997). How users reciprocate to computers: An experiment that demonstrates behavior change. In A. Edwards & S. Pemberton (Eds.), *CHI '97 Extended Abstracts on Human Factors in Computing Systems Looking to the Future* (pp. 331–332). ACM. <https://doi.org/10.1145/1120212.1120419>
- Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2), 181–218. [https://doi.org/10.1016/0010-0277\(87\)90018-7](https://doi.org/10.1016/0010-0277(87)90018-7)
- Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. In H. Giles, J. Coupland, & N. Coupland (Eds.), *Contexts of Accommodation: Developments in Applied Sociolinguistics* (pp. 1–68). Cambridge University Press.
- Hoey, M. (2007). Lexical priming and literacy creativity. In M. Hoey, M. Mahlberg, M. Stubbs, & W. Teubert (Eds.), *Text, Discourse and Corpora: Theory and Analysis* (pp. 7–29). Continuum International Publishing.
-

- Hoffmann, L., Derksen, M., & Kopp, S. (2020). What a pity, Pepper! How warmth in robots' language impacts reactions to errors during a collaborative task. In T. Belpaeme, J. Young, H. Gunes, & L. Riek (Eds.), *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction—HRI \_20* (pp. 245–247). ACM. <https://doi.org/10.1145/3371382.3378242>
- Horstmann, A. C., Bock, N., Linhuber, E., Szczuka, J. M., Straßmann, C., & Krämer, N. C. (2018). Do a robot's social skills and its objection discourage interactants from switching the robot off? *PloS One*, *13*(7), e0201581. <https://doi.org/10.1371/journal.pone.0201581>
- Horstmann, A. C., & Krämer, N. C. (2020). Expectations vs. actual behavior of a social robot: An experimental investigation of the effects of a social robot's interaction skill level and its expected future role on people's evaluations. *PloS One*, *15*(8), e0238133. <https://doi.org/10.1371/journal.pone.0238133>
- Horstmann, A. C., & Krämer, N. C. (2022). The fundamental attribution error in human-robot interaction: An experimental investigation on attributing responsibility to a social robot for its pre-programmed behavior. *International Journal of Social Robotics*, *14*, 1137–1153. <https://doi.org/10.1007/s12369-021-00856-9>
- Huang, L., Morency, L.-P., & Gratch, J. (2011). Virtual rapport 2.0. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, H. H. Vilhjálmsson, S. Kopp, S. Marsella, & K. R. Thórisson (Eds.), *Lecture Notes in Computer Science. Intelligent Virtual Agents* (Vol. 6895, pp. 68–79). Springer. [https://doi.org/10.1007/978-3-642-23974-8\\_8](https://doi.org/10.1007/978-3-642-23974-8_8)
- Kim, J., Merrill, K., Xu, K., & Sellnow, D. D. (2021). I like my relational machine teacher: An AI instructor's communication styles and social presence in online education. *International Journal of Human-Computer Interaction*, *37*(18), 1760–1770. <https://doi.org/10.1080/10447318.2021.1908671>
- Krämer, N., Kopp, S., Becker-Asano, C., & Sommer, N. (2013). Smile and the world will smile with you—The effects of a virtual agent's smile on users' evaluation and behavior. *International Journal of Human-Computer Studies*, *71*(3), 335–349. <https://doi.org/10.1016/j.ijhcs.2012.09.006>
- 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), 665–672. <https://doi.org/10.1016/j.chb.2010.01.003>
- López, G., Quesada, L., & Guerrero, L. A. (2018). Alexa vs. Siri vs. Cortana vs. Google Assistant: A comparison of speech-based natural user interfaces. In I. L. Nunes (Ed.), *Advances in Intelligent Systems and Computing. Advances in Human Factors and Systems Interaction* (Vol. 592, pp. 241–250). Springer International Publishing. [https://doi.org/10.1007/978-3-319-60366-7\\_23](https://doi.org/10.1007/978-3-319-60366-7_23)
- Lorenz, T., Weiss, A., & Hirche, S. (2016). Synchrony and reciprocity: Key mechanisms for social companion robots in therapy and care. *International Journal of Social Robotics*, *8*(1), 125–143. <https://doi.org/10.1007/s12369-015-0325-8>
- McCroskey, J. C., & McCain, T. A. (1974). The measurement of interpersonal attraction. *Speech Monographs*, *41*(3), 261–266. <https://doi.org/10.1080/03637757409375845>
-

- Mell, J., Lucas, G. M., & Gratch, J. (2018). Welcome to the real world: How agent strategy increases human willingness to deceive. In M. Dastani, G. Sukthankar, E. André, & S. Koenig (Eds.), *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems—AAMAS '18* (pp. 1250–1257). IFAAMAS.
- Nass, C., & Brave, S. (2005). *Wired for speech: How voice activates and advances the human-computer relationship*. The MIT Press. <https://web.archive.org/web/20211027225327/https://aclanthology.org/j06-3009.pdf>
- Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1), 81–103. <https://doi.org/10.1111/0022-4537.00153>
- Oviatt, S., Bernard, J., & Levow, G. A. (1998). Linguistic adaptations during spoken and multimodal error resolution. *Language and Speech*, 41, 419–442. <https://doi.org/10.1177/002383099804100409>
- Pearson, J., Hu, J., Branigan, H. P., Pickering, M. J., & Nass, C. (2006). Adaptive language behavior in HCI. In R. Grinter, T. Rodden, P. Aoki, E. Cutrell, R. Jeffries, & G. Olson (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems—CHI '06* (pp. 1177–1180). ACM. <https://doi.org/10.1145/1124772.1124948>
- Pickering, M. J., & Garrod, S. (2004). The interactive-alignment model: Developments and refinements. *The Behavioral and Brain Sciences*, 27(2), 212–225. <https://doi.org/10.1017/S0140525X04450055>
- Pickering, M. J., & Garrod, S. (2006). Alignment as the basis for successful communication. *Research on Language and Computation*, 4(2–3), 203–228. <https://doi.org/10.1007/s11168-006-9004-0>
- Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.
- Ribino, P. (2023). The role of politeness in human-machine interactions: A systematic literature review and future perspectives. *Artificial Intelligence Review*, 56(S1), 445–482. <https://doi.org/10.1007/s10462-023-10540-1>
- Riordan, M. A., Kreuz, R. J., & Olney, A. M. (2014). Alignment is a function of conversational dynamics. *Journal of Language and Social Psychology*, 33(5), 465–481. <https://doi.org/10.1177/0261927X13512306>
- Sandoval, E. B., Brandstetter, J., Obaid, M., & Bartneck, C. (2016). Reciprocity in human-robot interaction: A quantitative approach through the prisoner's dilemma and the ultimatum game. *International Journal of Social Robotics*, 8(2), 303–317. <https://doi.org/10.1007/s12369-015-0323-x>
- Suzuki, N., & Katagiri, Y. (2007). Prosodic alignment in human-computer interaction. *Connection Science*, 19(2), 131–141. <https://doi.org/10.1080/09540090701369125>
- Templin, M. C. (1957). *Certain language skills in children: Their development and interrelationships*. University of Minnesota Press. <https://www.jstor.org/stable/10.5749/j.ctttv2st.16>
- von der Pütten, A., Krämer, N. C., Gratch, J., & Kang, S.-H. (2010). “It doesn't matter what you are!” Explaining social effects of agents and avatars. *Computers in Human Behavior*, 26(6), 1641–1650. <https://doi.org/10.1016/j.chb.2010.06.012>
-

# Chatbot vs. Human: The Impact of Responsive Conversational Features on Users' Responses to Chat Advisors

Stefanie H. Klein<sup>1</sup>  and Sonja Utz<sup>1,2</sup> 

1 Everyday Media Lab, Leibniz-Institut für Wissensmedien, Tübingen, Germany

2 Department of Psychology, Eberhard Karls Universität Tübingen, Germany

## Abstract

As educational organizations increasingly consider supporting or replacing human chat advisors with chatbots, it is crucial to examine if users perceive a chatbot differently from a human. Chatbots' conversational features may signal responsiveness and thus improve user responses. To explore this, we conducted three online experiments ( $N_{\text{total}} = 1,005$ ) using a study advising setting. We computed pooled data analyses because the individual study results did not provide clear support for our hypotheses. Results indicate that users prefer human agents regarding competence and intention to use but not perceived enjoyment. Responsiveness increased likability, warmth, and satisfaction. Perceptions of the interaction mediated the responsiveness effects. Our findings suggest that educational organizations can support their study advising departments with well-functioning chatbots without eliciting negative user responses.

**Keywords:** agent type, responsiveness, chatbot, user response, human-machine communication

**CONTACT** Stefanie H. Klein  • [s.klein@iwm-tuebingen.de](mailto:s.klein@iwm-tuebingen.de) • Everyday Media Lab • Leibniz-Institut für Wissensmedien • Schleichstrasse 6 • 72076 Tübingen, Germany

ISSN 2638-602X (print)/ISSN 2638-6038 (online)  
[www.hmcjournal.com](http://www.hmcjournal.com)



Copyright 2024 Authors. Published under a Creative Commons Attribution 4.0 International (CC BY-NC-ND 4.0) license.



## Introduction

Chatbots are text-based automated agents interacting with users through natural language (Shawar & Atwell, 2007). They increasingly rely on artificial intelligence (AI-)based technologies, including large language models (LLMs) like GPT-3 (Kasneji et al., 2023). Already a part of many organizations' communication policies, AI-based chatbots are also increasingly used in educational contexts (e.g., for learning support or academic advising; Karrenbauer et al., 2021; Kasneji et al., 2023). Academic advising can be prescriptive, providing information about administrative issues, or developmental, aiming at defining and exploring study or career goals (Gordon, 1994; Mottarella et al., 2004). Developmental advising holds great potential for using AI-based chatbots but has received little attention in previous research (Meyer von Wolff et al., 2020). However, because developmental advising entails more than simply answering information requests, we deem it crucial to systematically investigate whether potential students prefer to be advised by a human or a chatbot (Sundar, 2020). Although developers have made significant technological progress in developing conversational agents, increasing chatbot acceptance among users remains a challenge (Neururer et al., 2018). Responsiveness, in the form of backchanneling cues, is a promising conversational feature that has been positively linked to organizational and relational outcomes in prior research on human-human (Davis & Perkowski, 1979; De Ruyter & Wetzels, 2000) and human-robot interactions (Birnbbaum et al., 2016) but has hardly been studied in chatbots, as shown in a recent systematic review by Van Pinxteren et al. (2020). We thus consider it important to examine the effects of chat agents' responsiveness on user responses. In three online vignette experiments, we aim to answer the research question: "To what extent do agent type (chatbot vs. human) and responsiveness influence users' responses to chat advisors?" Additionally, we look at the underlying mechanisms from three levels relating to different aspects of the interaction (i.e., the interaction in general, the dialogic nature, and the content of the conversation).

## Related Research and Theoretical Background

### User Responses to Chat Agents

Before turning to prior work on agent type and responsiveness, we want to introduce the user responses that form our dependent variables. Following previous work in human-agent interaction (e.g., Diers, 2020; S. Lee & Choi, 2017; Lou et al., 2021) and academic advising (e.g., Mottarella et al., 2004), we examine users' general attitude toward the way of communicating with an organization. To gain a deeper understanding, we look at specific cognitive components; for example, the extent to which the advisor is perceived as likable (b), intelligent (c), warm (d), and competent (e). These perceptions are basic dimensions in evaluating new actors (Bartneck et al., 2009; Fiske, 2018). Likability and perceived intelligence are concepts stemming from human-robot interaction research. Warmth (i.e., good social intentions) and competence (i.e., the ability or expertise of the advisor) play a crucial role in evaluating human study advisors (Lou et al., 2021; Mottarella et al., 2004). Likability and warmth cover the social aspect of agent perception, while intelligence and competence are rather task-related (Bartneck et al., 2009; Fiske, 2018). We also capture users' affective and behavioral ratings of chat advisors to provide a comprehensive picture. Perceived

---



enjoyment (f) constitutes the affective component. It refers to the extent to which interacting with a system is perceived as pleasurable and fun (Diers, 2020). We conceptualize satisfaction (g) as participants' perceived performance of the advisor they see in the vignette. In line with the Technology Acceptance Model, a prominent model to explain users' acceptance and usage of emerging technologies (Venkatesh, 2000), attitude, perceived enjoyment, and satisfaction are considered to be antecedents of user acceptance in terms of intention to use the way of communication with the organization (Diers, 2020; S. Lee & Choi, 2017). We also include the intention to use the communication medium (h), which, in turn, is considered to predict actual usage behavior (Venkatesh, 2000), but we were unable to measure this in our vignette studies.

We use the Media Are Social Actors (MASA) paradigm as an overarching theoretical framework for examining the impact of social cues on user responses to media (Lombard & Xu, 2021). The MASA paradigm extends the Computers Are Social Actors (CASA) paradigm, which states that people apply social rules to their interactions with computers (Nass & Moon, 2000) by considering a medium's social cues and the social signals these elicit to users as crucial for activating user responses. Lombard and Xu adopt Fiore et al.'s definition of social cues as "features salient to observers because of their potential as channels of useful information" (2013, p. 2). In contrast, social signals refer to the interpretation of a sender medium's social cues by the receiver (Fiore et al., 2013). Examples of a medium's social cues include gestures, motion, and language use. These can send out social signals of social identity, interactivity, and responsiveness to users (Lombard & Xu, 2021). This research focuses on cues signaling identity (agent type: human vs. chatbot) and responsiveness (i.e., the use or nonuse of verbal backchanneling cues).

A helpful model for investigating the impact of such cues is the Modality-Agency-Interactivity-Navigability (MAIN) model (Sundar, 2008). The MAIN model postulates that cognitive heuristics about an interaction's character and content are triggered when visual and identity cues are used in the interface (Sundar, 2008). The features of an interface can thus shape users' interaction experience (Sundar, 2020). The model identifies four affordances, which are present in most media: Modality, Agency, Interactivity, and Navigability. In this research, we focus on agency, which refers to the information source whose identity is communicated by an interface, in our case, an agent, to the user (Sundar, 2008). Visual or verbal cues (e.g., a human-like picture or backchannel utterances) can communicate an interface's identity (Sundar, 2008). These cues can trigger cognitive heuristics, like the machine heuristic and the social presence heuristic, which likely affect users' responses to chat advisors (Sundar, 2008).

### **Agent Type: Chatbot vs. Human**

According to the MAIN model, machines are expected to lack emotions and, thus, be objective, rule-governed, and invariant on the one hand (Sundar, 2008). Machine-like cues can positively influence credibility perceptions by triggering the machine heuristic (Sundar, 2008). For instance, Sundar and Nass (2001) found that news displayed by machine-like systems is perceived as more objective than news displayed by humans. On the other hand, machines are stereotyped as unemotional and cold (Sundar, 2020).

If the positive or negative impact of the machine heuristic prevails (i.e., whether people favor humans or machine agents) strongly depends on the task at hand (M. K. Lee, 2018). We argue that in developmental study advising, people might expect to talk to a human because they consider machines as “unfit for ‘human tasks’ that involve subjective judgments and emotional capabilities” (Sundar, 2020, p. 80). Advising prospective students about degree programs requires not only rational data processing but also intuitive judgments and interpersonal competencies (Mottarella et al., 2004). The outcome of developmental study advising can have decisive consequences for a person’s life. Thus, in this context, people might prefer communicating with a human because they consider them more flexible, adaptable, and sympathetic than a machine—in our case, a chatbot. Previous studies found that machine-like cues hinder positive agent evaluations, while human-like cues promote positive assessments; for example, users rated human agents (vs. chatbots) higher regarding expected likability and social presence (Spence et al., 2014) and social attraction (Lew & Walther, 2023). Given theory and prior research, we expect participants to have a more positive attitude toward a human advisor. They should also consider a human more likable, intelligent, warm, and competent than the chatbot (Lou et al., 2021). Moreover, we expect perceived enjoyment, satisfaction, and use intention to be higher for the human advisor (Prahl & Van Swol, 2021):

**H1:** Human identity cues have a positive effect on (a) attitude, (b) likability, (c) perceived intelligence, (d) warmth, (e) competence, (f) perceived enjoyment, (g) satisfaction, and (h) intention to use ( $\mu_{\text{human identity cues}} > \mu_{\text{chatbot identity cues}}$ ).

## Responsiveness

The general concept of responsiveness stems from the interpersonal relationship literature, where it refers to the likelihood of each partner responding to the other and the proportions of relevant and adequate responses (Davis, 1982). In close interpersonal relationships, responsiveness refers to “the processes through which relationship partners attend to and respond supportively to each other’s needs, wishes, concerns, and goals, thereby promoting each other’s welfare” (Reis & Clark, 2013, p. 400). It is considered pivotal for human attachment processes. Responsiveness depends on how partners perceive and respond to each other’s needs (Reis & Clark, 2013). When a partner feels that their needs are being met, feelings of closeness and mutual sympathy emerge. Partners who are responsive (i.e., psychologically empathetic, attentive, and supportive of one another) benefit in terms of liking, well-being, and satisfaction (Birnbaum et al., 2016; Davis & Perkowski, 1979; Reis & Clark, 2013). Responsive behavior manifests itself in asking questions, paralinguistic behavior in the form of backchannel utterances, summarizing and paraphrasing what has been said, and expressing understanding (Maisel et al., 2008). Backchannel cues signal “attention to, support or encouragement for, or even acceptance of the speaker’s message” (Mulac et al., 1998, p. 647). In service encounters, responsiveness has been shown to increase outcomes such as customer satisfaction and trust (De Ruyter & Wetzels, 2000). In human-machine communication (HMC), positive responsiveness effects on perceived competence, sociability, and willingness to use have been found for social robots (Birnbaum et al., 2016). There is first evidence that backchanneling increases the intention to use a chatbot (S. Lee et al.,

2020). Taken together, we expect that responsive (vs. not responsive) advisors increase cognitive, affective, and behavioral user responses (Birnbaum et al., 2016; De Ruyter & Wetzels, 2000):

**H2:** Responsive verbal cues have a positive effect on (a) attitude, (b) likability, (c) perceived intelligence, (d) warmth, (e) competence, (f) perceived enjoyment, (g) satisfaction, and (h) intention to use ( $\mu_{\text{responsive verbal cues}} > \mu_{\text{no responsive verbal cues}}$ ).

## The Interplay Between Agent Type and Responsiveness

We do not know yet whether the effect of responsiveness is the same across both types of agents or not. According to expectation violation theory (Burgoon & Hale, 1988; Grimes et al., 2021), initial user expectations of an agent's performance and whether they are confirmed or violated matter. Users attribute more attention to violated expectations than confirmed ones (Burgoon & Hale, 1988). Expectation violations can be positive or negative: A positive violation is seen as beneficial (e.g., when the user perceives a conversational agent as better than expected). In contrast, a negative violation indicates that the user expected more of the agent than they received (Burgoon & Hale, 1988). By telling users they are about to chat with a chatbot, compared to a human, expectations of the agent are reduced (Grimes et al., 2021). When the chatbot employs responsive cues and thus strongly resembles the human agent in its conversational characteristics, a responsive chatbot might receive more positive user responses than a chatbot without responsive verbal cues. As overly anthropomorphic chatbots are often perceived as eerie (Mori et al., 2012), users might feel uncomfortable talking to a responsive chatbot. Therefore, one could also expect the responsive chatbot to receive more negative user responses than the chatbot without responsive verbal cues. Given these conflicting lines of reasoning and the lack of previous research (except, e.g., Beattie et al., 2020; Sundar et al., 2016), we formulate an exploratory research question:

**RQ1:** Is there an interaction effect between agent type (human vs. chatbot identity cues) and responsiveness (responsive vs. no responsive verbal cues) on (a) attitude, (b) likability, (c) perceived intelligence, (d) warmth, (e) competence, (f) perceived enjoyment, (g) satisfaction, and (h) intention to use?

## Underlying Processes of Agent Type and Responsiveness Effects

To contribute to a more comprehensive understanding of the effects of agent type and responsiveness on user responses, we consider the processes underlying this relationship from three levels of analysis that relate to different aspects of interaction believed to be important in virtual interactions with social actors (Go & Sundar, 2019; Van der Goot & Etzrodt, 2023): the interaction in general, the dialogic nature, and the conversation content.

When looking at the interaction in general, we examine the mediating role of social presence. Social presence is defined as the perception of "being with another" (Biocca et al., 2003, p. 468). In HMC, it refers to the user's perception of interacting with a social entity rather than a machine (Sundar, 2008). The concept has been shown to positively impact attitudinal and behavioral outcomes in virtual interactions (Gefen & Straub, 2004;

Oh et al., 2018). Social presence is a fleeting judgment of an interaction influenced by the medium (Biocca et al., 2003). For instance, agents that provide human-like visual and verbal cues lead to stronger perceptions of social presence than agents that do not (S. Lee et al., 2020; Sundar, 2008). We thus expect human and responsive advisors to elicit higher levels of social presence, resulting in higher cognitive, affective, and behavioral user-related outcomes (Biocca et al., 2003; Go & Sundar, 2019).

**H3:** The effects of agent type on (a) attitude, (b) likability, (c) perceived intelligence, (d) warmth, (e) competence, (f) perceived enjoyment, (g) satisfaction, and (h) intention to use are mediated by social presence.

**H4:** The effects of responsiveness on (a) attitude, (b) likability, (c) perceived intelligence, (d) warmth, (e) competence, (f) perceived enjoyment, (g) satisfaction, and (h) intention to use are mediated by social presence.

Next, we look at the dialogic nature of the conversation. Dialogue is a fundamental feature of human conversations and provides the interlocutors with a sense of reciprocity, cooperation, and support (Kent & Taylor, 2002), also attributed to responsiveness (Reis & Clark, 2013). Conversations with responsively communicating agents should be perceived more as a dialogue than ones with an agent not using responsive verbal cues. Similar effects have been found for verbal cues signaling message contingency (Go & Sundar, 2019; Sundar et al., 2016). Perceived dialogue has been shown to increase advisor perceptions and usage intention (Go & Sundar, 2019). We argue that responsive verbal cues positively affect user responses via perceived dialogue:

**H5:** The effects of responsiveness on (a) attitude, (b) likability, (c) perceived intelligence, (d) warmth, (e) competence, (f) perceived enjoyment, (g) satisfaction, and (h) intention to use are mediated by perceived dialogue.

Closely related to the concept of perceived dialogue but focused more on the actual content of the conversation is the concept of feeling heard, defined as “the feeling that one’s communication is received with attention, empathy, respect, and in a spirit of mutual understanding” (Roos et al., 2023, p. 5). Responsive verbal cues could reinforce the user’s feeling of being heard, leading to more positive evaluations of the advisor and the interaction. As feeling heard is a new and under-researched concept, we want to answer the following research question:

**RQ2:** Does feeling heard mediate the effects of responsiveness on (a) attitude, (b) likability, (c) perceived intelligence, (d) warmth, (e) competence, (f) perceived enjoyment, (g) satisfaction, and (h) intention to use?

All three potential mediating mechanisms are considered to play important roles in virtual interactions with social actors (Go & Sundar, 2019), which is why we believe they could operate in parallel (see Ischen et al., 2020 for a similar approach). Feelings of social presence and being heard as well as the perception of dialogue in an interaction all involve

---

notions of reciprocity, responsiveness, and mutual understanding (Roos et al., 2023; Sundar et al., 2016). Therefore, we will simultaneously investigate whether social presence, perceived dialogue, and feeling heard mediate the responsiveness effects on our outcomes of interest. Entering all three mediators in the same model controls for shared variance and provides stronger evidence for conclusions about the underlying processes (Hayes, 2022).

## Overview of the Current Studies

In three online vignette experiments conducted in 2021, we investigated the extent to which agent type and responsiveness influence users' responses to chat advisors. Vignette designs are common in HMC research (Greussing et al., 2022; e.g., Abendschein et al., 2021; Beattie et al., 2020). The local ethics committee of the Leibniz-Institut für Wissensmedien, Tübingen, approved the studies. Informed consent was obtained from participants before their participation. Preregistrations, materials, data, and additional results are freely accessible on OSF: <https://osf.io/w8dzv>.

## Study 1

### Method

We conducted a 2 (agent type: chatbot vs. human)  $\times$  2 (responsiveness: absence vs. presence of responsive verbal cues) between-subjects experiment. Participants were recruited via the online sampling platform Prolific. Of the 280 participants who completed the study, 253 passed the agent type manipulation check and were retained ( $n_{\text{female}} = 101$ , age:  $M = 28.18$ ,  $SD = 9.25$ , range = 18–69) (power analysis in Appendix A). Participants were randomly assigned to one of four conditions. After providing informed consent, all participants saw a vignette in the form of a pre-recorded animated chat conversation between the study advisor, Sophie, and Marc, a prospective student at Sophie's university. The advisor asked the user several questions during the conversation to find out his interests. Based on Marc's answers, Sophie recommended suitable degree programs. Finally, we asked participants to complete a survey about their perceptions of the interaction and key demographics. The design of the chat interface resembled the design of contemporary messenger interfaces used in practice (Appendix B). A robot icon was used to represent the chatbot, and it introduced itself as "Sophie, the chatbot of the student advisory service." The human advisor was represented with the portrait of a businesswoman, and she introduced herself as "Sophie, a student advisor" (see Go & Sundar, 2019 for a similar approach). All conversations were equal in content and without disruptions. Responsiveness was manipulated using short backchanneling cues and tokens like questions, paraphrases, and expressions signaling thinking processes (Maisel et al., 2008). Specifically, the responsive agents responded to the user's inputs with utterances like "Mhm," "Got it," and "Hmm, let me think." The responsive agent also asked the user for his name and repeated it in the following input.

To measure the *attitude* toward the means of communication, we used a scale by Diers (2020). Specific cognitive user responses were assessed using the *likability* and *perceived intelligence* scales from the Godspeed Questionnaire (Bartneck et al., 2009) and the *warmth* and *competence* scales from Fiske (2018). Affective user response was assessed using a

*perceived enjoyment* scale, and the behavioral aspect was reflected with an *intention to use* scale, both by Diers (2020). We asked participants to rate whether they would have found the advisor's behavior satisfactory if they had been in the student's position using Lagace et al.'s *satisfaction* scale (1991). We adopted the scales for *social presence*, *perceived dialogue*, and *feeling heard* from Gefen and Straub (2004), Sundar et al. (2016), and Roos et al. (2023), respectively. We included manipulation checks for agent type (adapted from Go & Sundar, 2019) and responsiveness (adapted from De Ruyter & Wetzels, 2000) to ensure effective manipulations. We assessed the same variables in all studies (Appendix C, descriptive statistics in Appendix D). Bivariate correlations were rather strong (Appendix E), ranging from  $r = .39$  to  $r = .89$  ( $p < .001$ ). Attitude and intention to use, both drawn from the Technology Acceptance Model literature (Venkatesh, 2000) and relating to the communication with the organization, were strongly correlated ( $r > .86$ ). Perceived dialogue and feeling heard also correlated strongly, potentially due to similarity in item content. Internal consistency of all constructs was satisfactory (Cronbach's  $\alpha > .80$ ).

## Results

As intended, participants in the responsive conditions perceived the agent as significantly more responsive ( $M = 5.89$ ,  $SD = 1.03$ ) than those in the conditions without responsive verbal cues ( $M = 5.15$ ,  $SD = 1.36$ ), as a Welch two-sample  $t$ -test showed ( $t(237.59) = 4.88$ ,  $p < .001$ ,  $d = 0.61$ ). To test H1 and H2 and to answer RQ1, we carried out a two-way multivariate analysis of variance (MANOVA). Using Pillai's trace, there was a significant main effect of agent type on the outcomes ( $V = 0.15$ ,  $F(1, 248) = 5.40$ ,  $p < .001$ ). Separate univariate analyses of variance (ANOVA) only revealed a significant agent type effect for satisfaction (g) ( $F(1, 248) = 6.72$ ,  $p = .010$ ): Participants in the chatbot conditions tended to be more satisfied ( $M = 5.56$ ,  $SD = 1.24$ ) than participants in the human conditions (Satisfaction:  $M = 5.13$ ,  $SD = 1.48$ ;  $t(239.74) = -2.51$ ,  $p = .013$ ,  $d = -0.32$ ). Neither the responsiveness effect ( $V = 0.04$ ,  $F(1, 248) = 1.16$ ,  $p = .324$ ) nor the interaction between agent type and responsiveness ( $V = 0.02$ ,  $F(1, 248) = 0.74$ ,  $p = .657$ ) were significant. We rejected H1 and H2. We did not perform mediation analyses to test H3–H5 and to answer RQ2 because neither agent type nor responsiveness positively impacted the outcomes. We thus rejected H3–H5.

## Discussion

Contrary to the hypotheses, neither human identity cues nor responsive behavior positively affected user responses. The zero effects of responsiveness are striking, given the significant responsiveness effects on the manipulation check. Moderation analyses did not yield significant results. The sample size was slightly below the target size due to our exclusion criterion. As we had based our power considerations on a small interaction effect that Go and Sundar (2019) found in a study where participants directly interacted with chat agents, the effect sizes in our vignette design could be even smaller. Hence, we decided to replicate our study with a larger sample.



## Study 2

### Method

The experimental design and measures were equal to those in Study 1. We aimed to recruit 403 participants via university mailing lists. A total of 520 participants completed the study. We excluded 118 participants because they failed the agent type manipulation check and one who admitted to not having answered the questionnaire reliably, which led to a final sample of  $N = 401$  ( $n_{\text{female}} = 287$ , age:  $M = 24.26$ ,  $SD = 6.11$ , range = 18–69).

### Results

As intended, participants in the responsive conditions scored significantly higher on the responsiveness manipulation check ( $M = 5.86$ ,  $SD = 1.20$ ) than participants in the conditions without responsive verbal cues ( $M = 4.67$ ,  $SD = 1.54$ ) as a Welch two-sample  $t$ -test showed ( $t(357.33) = 8.51$ ,  $p < .001$ ,  $d = 0.86$ ). We conducted a two-way MANOVA to test H1 and H2 and answer RQ1. Using Pillai's trace, there was a significant main effect of agent type ( $V = 0.06$ ,  $F(1, 394) = 3.36$ ,  $p = .001$ ). The responsiveness effect ( $V = 0.03$ ,  $F(1, 394) = 1.58$ ,  $p = .129$ ) and the interaction between agent type and responsiveness ( $V = 0.02$ ,  $F(1, 394) = 0.95$ ,  $p = .388$ ) were not significant in the multivariate model. Separate ANOVAs revealed a significant agent type effect for likability (b) ( $F(1, 394) = 4.92$ ,  $p = .036$ ). Participants in the chatbot conditions rated the agent more likable ( $M = 5.63$ ,  $SD = 1.05$ ) than participants in the human conditions ( $M = 5.41$ ,  $SD = 1.07$ ;  $t(397) = -2.21$ ,  $p = .035$ ,  $d = -0.21$ ). Although the responsiveness effect was not significant, we computed separate ANOVAs, revealing positive responsiveness effects on warmth (d) ( $F(1, 394) = 6.32$ ,  $p = .012$ ) and satisfaction (g) ( $F(1, 394) = 6.53$ ,  $p = .011$ ). We rejected H1 and accepted H2d, g). We computed parallel multiple mediator models (Hayes, 2022) predicting warmth and satisfaction using the R package *lavaan* (Rosseel et al., 2021) but did not find significant indirect effects (OSF).

### Discussion

Like in Study 1, human identity cues did not significantly improve user responses in Study 2. However, we found significant effects of responsive conversational cues on warmth and satisfaction. We conducted a third study to clarify our findings from Studies 1 and 2.

## Study 3

### Method

We collected data from 418 participants via the crowdsourcing platform Clickworker. Three hundred fifty-one participants passed the agent type manipulation check and were retained ( $n_{\text{female}} = 127$ , age:  $M = 38.53$ ,  $SD = 12.28$ , range = 18–73). The experimental design and measures were equal to those we used in Studies 1 and 2.<sup>1</sup>

1. Study 3 was designed to additionally explore the impact of agent response time on users' responses, so response time (immediate or dynamically delayed) was included as a third experimental factor. However, because the results were not vital to answering our research question, we decided to move them to OSF.

## Results

Participants in the responsive conditions scored significantly higher on the responsiveness manipulation check ( $M = 6.08$ ,  $SD = 0.88$ ) than participants in the conditions without responsive verbal cues ( $M = 5.54$ ,  $SD = 0.96$ ), as shown in a two-sample  $t$ -test ( $t(349) = 5.54$ ,  $p < .001$ ,  $d = 0.59$ ). To test H1–H2 and to answer RQ1, a two-way MANOVA was carried out. Using Pillai's trace, there was a significant main effect of agent type ( $V = 0.08$ ,  $F(1, 347) = 4.11$ ,  $p < .001$ ) on the outcome variables. The responsiveness effect ( $V = 0.04$ ,  $F(1, 347) = 1.66$ ,  $p = .108$ ) and the interaction between agent type and responsiveness ( $V = 0.02$ ,  $F(1, 347) = 1.04$ ,  $p = .403$ ) were not significant in the multivariate model. Separate ANOVAs only revealed a significant agent type effect on competence (e) ( $F(1, 347) = 12.49$ ,  $p < .001$ ) and a marginally significant responsiveness effect on likability (b) ( $F(1, 347) = 3.85$ ,  $p = .051$ ). Follow-up tests yielded that participants in the human conditions rated the agent more competent ( $M = 5.76$ ,  $SD = 0.93$ ) than participants in the chatbot conditions ( $M = 5.36$ ,  $SD = 1.12$ ;  $t(349) = 3.54$ ,  $p < .001$ ,  $d = 0.01$ ). We, therefore, accepted H2e) and rejected H2. As we did not find significant positive agent type or responsiveness effects (H2), we did not compute parallel multiple mediator models.

## Discussion

Participants perceived the human agent as more competent than the chatbot, but no significant responsiveness effects were found. Still, a considerable proportion of participants did not recognize the alleged human, reducing the sample and resulting in a power loss for estimating interaction effects (posthoc power = 74.40%,  $N = 351$ ,  $\alpha = .05$ ,  $f = .14$ ). To mitigate the potential power issues of Studies 1 and 3, we conducted analyses based on the pooled data from all studies.

## Additional Analyses: Pooled Data

Using the pooled data ( $N = 1,005$ ) and controlling for study number, we performed an exploratory MANCOVA to clarify the main effects on the outcomes. Using Pillai's trace, the inclusion of study number as a control variable indicated differences between studies ( $V = 0.11$ ,  $F(2, 995) = 7.19$ ,  $p < .001$ ). Specifically, participants in Study 2 showed lower values on all outcomes than those in Studies 1 and 3, pointing to a generational effect. Study 2 comprised university students who were younger on average than participants in Studies 1 and 3 and, thus, may have had more experience with chat advisors. In contrast to the individual study results, using Pillai's trace, we found significant main effects of agent type (H1;  $V = 0.07$ ,  $F(1, 995) = 9.10$ ,  $p < .001$ ) and responsiveness (H2;  $V = 0.03$ ,  $F(1, 995) = 3.25$ ,  $p = .001$ ). The interaction between agent type and responsiveness (RQ1) was not significant ( $V < .01$ ,  $F(1, 995) = 0.59$ ,  $p = .783$ ). Univariate ANCOVAs and pairwise comparisons yielded significant positive effects of human identity cues on competence (e) and intention to use (h). Participants perceived the interaction with the human agent as less enjoyable than the interaction with the chatbot. In addition, significant positive responsiveness effects emerged for likability (b), warmth (d), and satisfaction (g), strengthening the individual study findings (Table 1, with adjusted means in Table 2).

**TABLE 1 Two-Way ANCOVA Statistics and Effect Sizes for Study Variables (Pooled Data)**

Variable	Effect	F ratio	p	$\eta^2_{\text{partial}}$
Attitude	Study	15.80	< .001	.03
	AT	3.27	.071	.00
	R	0.64	.422	.00
	AT × R	0.44	.510	.00
Likability	Study	3.23	.040	.01
	AT	3.79	.052	.00
	R	6.12	.014	.01
	AT × R	0.10	.755	.00
Perceived intelligence	Study	16.43	< .001	.03
	AT	2.41	.121	.00
	R	1.97	.161	.00
	AT × R	0.34	.558	.00
Warmth	Study	10.45	< .001	.02
	AT	0.51	.475	.00
	R	8.93	.003	.01
	AT × R	0.71	.400	.00
Competence	Study	7.02	< .001	.01
	AT	7.32	.007	.01
	R	0.04	.847	.00
	AT × R	0.25	.616	.00
Perceived enjoyment	Study	26.27	< .001	.05
	AT	4.95	.026	.01
	R	2.67	.102	.00
	AT × R	0.01	.927	.00
Satisfaction	Study	14.29	< .001	.03
	AT	3.28	.070	.00
	R	6.09	.013	.01
	AT × R	1.11	.293	.00
Intention to use	Study	21.83	< .001	.04
	AT	6.90	.009	.01
	R	0.48	.487	.00
	AT × R	0.15	.694	.00

Note.  $N = 1,005$ . ANCOVA = analysis of covariance. Study = study number, AT = agent type, R = responsiveness.  $df = 1,995$ , except  $df_{\text{Study}} = 2,995$ .

**TABLE 2** Adjusted Means and Effect Sizes for Study Variables (Pooled Data)

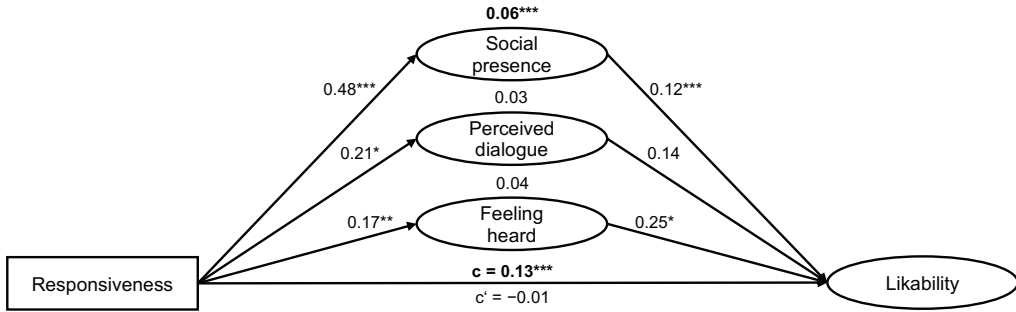
Variable	Agent Type				Responsiveness			
	Chatbot	Human	$p$	$\eta^2_{\text{partial}}$	Absent	Present	$p$	$\eta^2_{\text{partial}}$
Attitude	4.45	4.64	.068	.00	4.50	4.58	.435	.00
Likability	5.70	5.57	.050	.00	5.55	5.71	.013	.01
Perceived intelligence	5.58	5.68	.125	.00	5.58	5.67	.175	.00
Warmth	5.21	5.17	.481	.00	5.09	5.29	.003	.01
Competence	5.47	5.64	.006	.01	5.55	5.54	.844	.00
Perceived enjoyment	4.27	4.06	.029	.00	4.10	4.26	.089	.00
Satisfaction	5.20	5.04	.076	.00	5.01	5.24	.012	.01
Intention to use	4.35	4.65	.009	.01	4.45	4.52	.517	.00

Note.  $N = 1,005$ . One-way ANCOVAs controlled for study number.

To analyze the underlying relationship processes between agent type and responsiveness and the outcomes, we computed three parallel multiple mediator models predicting the outcomes significantly impacted by responsiveness; for example, likability (b), warmth (d), and satisfaction (g), using the R package *lavaan* (Rosseel et al., 2021). Including social presence, perceived dialogue, and feeling heard allowed us to model our three potential mediation levels simultaneously. We operationalized the latent constructs using reflective measurement models composed of the corresponding items.<sup>2</sup> All standardized factor loadings were sufficiently strong ( $\lambda > .50$ ) and significant ( $p < .001$ ). Knowing the mediators to be strongly correlated, we specified their covariances. We controlled the models for study number. Model fit was acceptable (Westland, 2015). As expected, we found high correlations between social presence and perceived dialogue ( $r = .67$  in the likability,  $r = .66$  in the warmth and satisfaction models), social presence and feeling heard ( $r = .61$ ), and perceived dialogue and feeling heard ( $r = .95$ ) throughout the models. Figures 1–3 display the results of the mediation models predicting likability, warmth, and satisfaction. Significant positive indirect effects emerged for likability via social presence, for warmth via social presence and feeling heard, and for satisfaction via social presence and perceived dialogue. The direct effects (i.e., the effects of responsiveness on the outcomes when the mediators were included in the model) were not significant, suggesting full mediations.

2. The standardized factor loadings of the two inversely coded feeling heard items were  $< .50$ . We thus followed Roos et al. (2023) in specifying the covariance between the residuals accounting for different response behaviors for inversely coded items. The error terms correlated moderately ( $r = .42$ ,  $p < .001$ ).

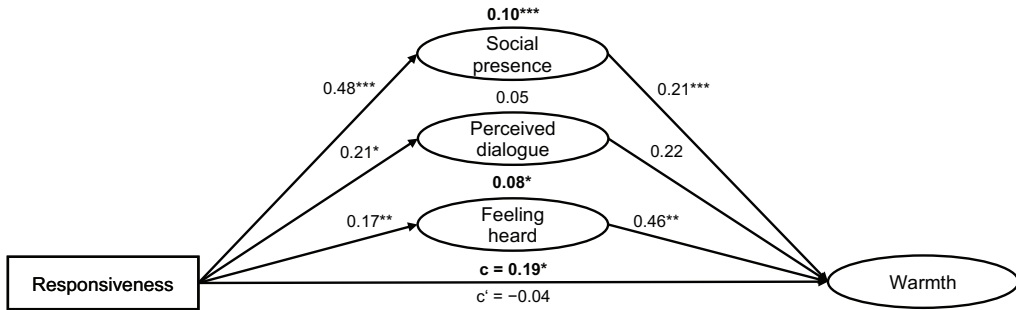
**FIGURE 1 Parallel Multiple Mediator Model for the Effect of Responsiveness on Likability (Pooled Data)**



Note.  $df = 261$ .  $X^2 = 1592.500$ ,  $p < .001$ ,  $CFI = .914$ ,  $RMSEA = .071$ ,  $CI_{RMSEA} (.068, .075)$ ,  $SRMR = .063$ . Unstandardized coefficients. Controlled for study number.  $R^2_{Likability} = .473$ ,  $R^2_{Social\ presence} = .054$ ,  $R^2_{Perceived\ dialogue} = .014$ ,  $R^2_{Feeling\ heard} = .009$ .

\*  $p < .05$ , \*\*  $p < .01$ , \*\*\*  $p < .001$ .

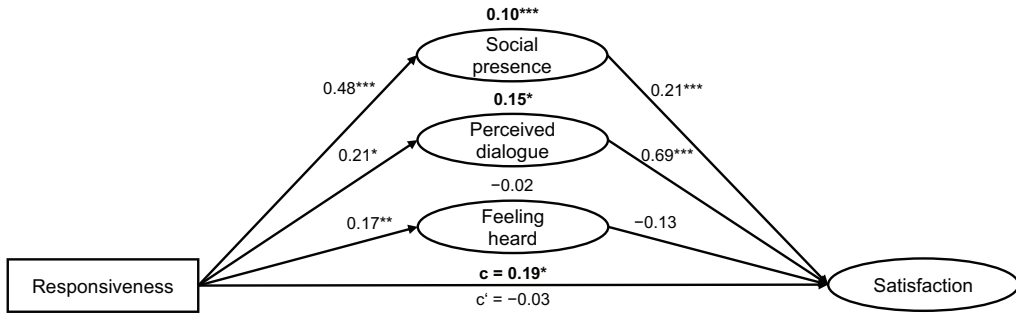
**FIGURE 2 Parallel Multiple Mediator Model for the Effect of Responsiveness on Warmth (Pooled Data)**



Note.  $df = 285$ .  $\chi^2 = 2406.803$ ,  $p < .001$ ,  $CFI = .875$ ,  $RMSEA = .086$ ,  $CI_{RMSEA} (.083, .089)$ ,  $SRMR = .071$ . Unstandardized coefficients. Controlled for study number.  $R^2_{Warmth} = .734$ ,  $R^2_{Social\ presence} = .054$ ,  $R^2_{Perceived\ dialogue} = .015$ ,  $R^2_{Feeling\ heard} = .009$ .

\*  $p < .05$ , \*\*  $p < .01$ , \*\*\*  $p < .001$ .

**FIGURE 3 Parallel Multiple Mediator Model for the Effect of Responsiveness on Satisfaction (Pooled Data)**



Note.  $df = 238$ ,  $\chi^2 = 1552.062$ ,  $p < .001$ ,  $CFI = .918$ ,  $RMSEA = .074$ ,  $CI_{RMSEA} (.071, .078)$ ,  $SRMR = .063$ . Unstandardized coefficients. Controlled for study number.  $R^2_{Satisfaction} = .592$ ,  $R^2_{Social\ presence} = .054$ ,  $R^2_{Perceived\ dialogue} = .015$ ,  $R^2_{Feeling\ heard} = .010$ . \*  $p < .05$ , \*\*  $p < .01$ , \*\*\*  $p < .001$ .

### General Discussion

We conducted three vignette experiments to answer the question: To what extent do agent type and responsiveness influence users’ responses to chat agents? Three key findings and subsequent implications emerged.

First, no study clearly supported our hypothesis on the positive impact of human identity cues on user responses (H1). The pooled analyses, however, suggested that participants considered the human agent more competent and indicated a higher intention to use it compared to the chatbot. This aligns with our hypothesis and suggests the machine heuristic does not hold for *human* tasks like developmental study advising (Sundar, 2020). Still, participants preferred the chatbot in terms of perceived enjoyment. There were no effects on the other outcomes, aligning with prior research suggesting little overall difference in the perception of humans and anthropomorphic chatbots (Beattie et al., 2020; Nass & Moon, 2000). Only the pooled analyses yielded small effects of human identity cues, although statistical power was high. The agents’ error-free answers and suitable program recommendations to the user might be the reason for the small effects. Data were collected before ChatGPT was launched, so the high quality of the answers allegedly stemming from a chatbot might have been surprising for the participants. Considering the rapid improvement of generative AI, the question of whether and how small identity cues affect chatbot evaluations gets even more important. We can conclude that regardless of how well a chatbot performs and how much people enjoy it, human agents seem to be preferred as study advisors of a university.

Second, we found significant positive responsiveness effects on warmth and satisfaction in Study 2 (H2). The pooled analyses confirmed these findings and yielded an additional significant positive responsiveness effect on likability. We showed that a responsive communication style elicits positive responses in contexts where agents must provide support and understanding. Positive responses were elicited regarding the agent’s social traits like likability, warmth, and satisfaction, corresponding to earlier findings from interpersonal



communication research (Davis & Perkowski, 1979). Responsive cues alone might be too subtle to tell whether people will want to be advised by an agent and not as decisive for perceptions relating to the successful completion of a task. Once high efficiency, the primary reason for using chatbots (Følstad & Skjuve, 2019), is reached, these softer cues might become more important.

Third, perceptions of the agent mediated the relationship between responsiveness and likability, warmth, and satisfaction (H4-H5, RQ2). Social presence consistently mediated the effects on likability, warmth, and satisfaction, whereas perceived dialogue and feeling heard only mediated the effect on satisfaction and warmth, respectively. Overall, this is consistent with our theoretical assumption that responsive verbal cues from an interlocutor signal understanding and support, thereby leading to more positive user responses (Reis & Clark, 2013). However, because perceived dialogue and feeling heard were highly correlated and had similarities in item content (Appendix C), controlling for one construct eliminated the respective other's effect. This raises the question of whether the variables represent different constructs. We showed that verbal cues have the potential to make users feel more socially present and heard (Lombard & Xu, 2021). How the interaction is perceived appears to be more critical to perceptions of the advisor's social attributes and satisfaction than individual aspects (e.g., its dialogic nature and conversation content).

Interestingly, responsiveness did not interact with agent type. Responsive cues seem equally important to peoples' perception of chatbots and humans. Future research could employ other interindividual moderators that could affect the effect of agent type on user outcomes (e.g., affinity for technological interaction; Franke et al., 2019). Context-specific differences could also be explored; for example, responsive cues might matter more when a chatbot serves as an emotional support tool (e.g., Birnbaum et al., 2016 for social robots).

Agent type influenced certain outcomes, while responsiveness influenced others. The machine heuristic suggests machine actors are viewed as more objective, rule-based, and competent than humans (Sundar, 2008). Our results challenge this, as competence and intention to use were higher in the human conditions. The task of study advising, which we consider a *human task* at its core, might be the reason (Sundar, 2020). Additionally, the items used to assess intention to use referred to the way of communication with an organization. Thus, even if chatbots perform just as well as humans, a preference for talking to humans and an aversion to the use of automation and algorithms in universities' communication remain (Dietvorst et al., 2015).

We aimed to investigate users' perceptions of the agent and the interaction as well as the classical technology acceptance variables attitude and intention to use (Venkatesh, 2000). Agent and interaction perceptions are established antecedents of attitude and use intention, whereas the latter can predict actual usage (i.e., adoption; Diers, 2020; S. Lee & Choi, 2017; Venkatesh, 2000). Responsiveness, in contrast, is more likely to impact social perceptions and seems relevant to users' satisfaction. For researchers more interested in the processes underlying chatbot adoption, the variables affected by responsiveness become relevant as they might mediate responsiveness effects on user satisfaction, which in turn might influence intention to use (Lou et al., 2021).

Our research contributes to the emerging research field of HMC regarding the impact of social cues on the perception and evaluation of machine agents (Gambino et al., 2020; Lombard & Xu, 2021). The different user responses to humans and chatbots suggest that

the media equation does not apply to all social interactions with machines (i.e., not all machines are always perceived as social actors; Van der Goot & Etzrodt, 2023). There is reason to believe that users mindfully evaluate the source depending on situational factors (e.g., the interaction context) and dispositional factors (e.g., personality) (E.-J. Lee, 2023). Van der Goot and Etzrodt (2023) recommend conducting more qualitative research to unravel these processes and to understand “how users negotiate the blurring boundaries between humans and machines” (p. 27). This question will become more important as human-like chatbots based on generative AI continue to gain traction.

Previous research on study advising has shown that a warm and supportive advising style is an influential factor for satisfaction and is even more important than the advising approach (Mottarella et al., 2004). A warm communication style is often associated with developmental advising. Although prescriptive advising is task-oriented and focuses on explaining requirements and procedures, a more responsive style could improve student acceptance. We suggest that researchers investigating the differences between various advising approaches should pay more attention to the advisor’s communication style, whether human or chatbot.

The results have implications for university practitioners considering using chatbots in developmental advising. While perceived competence and intention to use were higher for the human advisor, chatbot scores for these variables were above the scale means. So, chatbot support for student advising might be an efficient addition when financial resources or staff shortages are an issue. Leveraging the benefits of automated communication can thus be feasible without eliciting negative user responses. Yet, developers must ensure that the chatbots work well (e.g., by adequately exploiting the advantages of LLMs; Kasneci et al., 2023). But the way the chatbot presents the information is also critical. Study advising chatbots should be designed to evoke feelings of warmth and support, which have been shown to facilitate successful advising (Mottarella et al., 2004). Integrating responsive features into chat interactions may help universities and schools build and maintain warm and supportive relationships with their (potential) students. A well-thought-out dialogue design can help integrate responsive verbal cues without too much financial or human effort.

## **Limitations and Future Research**

Although vignette designs have high internal validity and give participants a unique perspective (Abendschein et al., 2021), they cannot offer as much ecological validity as experiments where participants directly interact with an agent. In our studies, participants were mere observers of the interaction, which could have increased their distance from the interaction, decreasing their involvement and identification with the user. The high nonrandom dropout rates due to failed agent type manipulation checks could have been related to the study design. To ensure experimental control, we kept the layout and content across all conditions constant. We thus manipulated agent type only in terms of the agent’s introduction and avatar, which may have led participants to perceive the human agent as an anthropomorphic chatbot. Future studies could examine participants’ direct interactions with chat agents to increase ecological validity. To ensure the comparability of our results, we used the same stimulus materials in all studies, which might have affected the validity of our results in case the stimuli did not optimally manipulate our independent variables. Future studies

---

could use stimulus sampling (i.e., employ a variety of user-agent conversations) to reduce the impact of the unique features of a particular stimulus on the results and strengthen the conclusions (Jackson & Jacobs, 1983).

Our studies focused on investigating the effects of agent type and responsiveness on a wide range of dependent variables, including cognitive, affective, and behavioral user responses, but not on the relationships between the outcomes. As there is a plethora of scales from different disciplines that measure similar constructs (e.g., human-likeness perceptions; Ischen et al., 2023), researchers call for common conceptualizations and measurement scales for key outcomes (Følstad et al., 2021; Greussing et al., 2022). A systematic assessment and confirmatory factor analysis of common scales used in HMC research could shed light on what makes each construct unique, how the constructs are empirically related, and how they contribute to chatbot adoption.

## Conclusion

In three experiments, we investigated the impact of agent type and responsiveness on a wide range of user-related outcomes in the context of study advising. Our results suggest that human agents are favored in terms of competence and intention to use but not in terms of perceived enjoyment. Further, the results indicate that responsiveness positively impacts users' perceptions of agent likability, warmth, and satisfaction, mainly by increasing perceptions of the interaction. Our studies add novel insights to the literature on human-machine communication and offer two practical implications: First, our findings may encourage educational organizations to support their study advising departments with chatbots. Second, the use of responsive language by human agents and chatbots could help organizations build and maintain healthy and sustainable relationships with their (potential) students. Due to significant advances in generative AI, we can expect that people will increasingly be unable to distinguish whether they are interacting with a human or a chatbot in the future. Therefore, it will continue to be crucial to systematically investigate the role of relatively small social cues in the perception and evaluation of AI-based chatbots.

## Author Biographies

**Stefanie H. Klein** (MA, University of Stuttgart) is a PhD student in the Everyday Media Lab at Leibniz-Institut für Wissensmedien in Tübingen, Germany. Her research focuses on human-machine communication, particularly the impact of chatbots' conversational characteristics on user acceptance.

 <https://orcid.org/0000-0002-7563-4548>

**Sonja Utz** (PhD, Catholic University of Eichstätt) is the head of the Everyday Media lab at Leibniz-Institut für Wissensmedien in Tübingen and a full professor for communication via social media at the University of Tübingen. Her research focuses on the effects of social and mobile media use, especially in knowledge-related contexts, and on human-machine interaction.

 <https://orcid.org/0000-0002-7979-3554>

## Center for Open Science



This article has earned the Center for Open Science badges for Open Data and Open Materials through Open Practices Disclosure. The authors have made their data and materials freely accessible at <https://osf.io/w8dzv/>. The article also earned a Preregistered badge for having a preregistered design available at <https://osf.io/w8dzv/>.

## References

- Abendschein, B., Edwards, C., & Edwards, A. (2021). The influence of agent and message type on perceptions of social support in human-machine communication. *Communication Research Reports*, 38(5), 304–314. <https://doi.org/10.1080/08824096.2021.1966405>
- Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1(1), 71–81. <https://doi.org/10.1007/s12369-008-0001-3>
- Beattie, A., Edwards, A. P., & Edwards, C. (2020). A bot and a smile: Interpersonal impressions of chatbots and humans using emoji in computer-mediated communication. *Communication Studies*, 71(3), 409–427. <https://doi.org/10.1080/10510974.2020.1725082>
- Biocca, F., Harms, C., & Burgoon, J. K. (2003). Toward a more robust theory and measure of social presence: Review and suggested criteria. *Presence: Teleoperators and Virtual Environments*, 12(5), 456–480. <https://doi.org/10.1162/105474603322761270>
- Birnbaum, G. E., Mizrahi, M., Hoffman, G., Reis, H. T., Finkel, E. J., & Sass, O. (2016). What robots can teach us about intimacy: The reassuring effects of robot responsiveness to human disclosure. *Computers in Human Behavior*, 63, 416–423. <https://doi.org/10.1016/j.chb.2016.05.064>
- Burgoon, J. K., & Hale, J. L. (1988). Nonverbal expectancy violations: Model elaboration and application to immediacy behaviors. *Communication Monographs*, 55(1), 58–79. <https://doi.org/10.1080/03637758809376158>
- Davis, D. (1982). Determinants of responsiveness in dyadic interaction. In W. Ickes & E. S. Knowles (Eds.), *Personality, Roles, and Social Behavior* (pp. 85–139). Springer. [https://doi.org/10.1007/978-1-4613-9469-3\\_4](https://doi.org/10.1007/978-1-4613-9469-3_4)
- Davis, D., & Perkowski, W. T. (1979). Consequences of responsiveness in dyadic interaction: Effects of probability of response and proportion of content-related responses on interpersonal attraction. *Journal of Personality and Social Psychology*, 37(4), 534–550. <https://doi.org/10.1037/0022-3514.37.4.534>
- De Ruyter, K., & Wetzels, M. G. M. (2000). The impact of perceived listening behavior in voice-to-voice service encounters. *Journal of Service Research*, 2(3), 276–284. <https://doi.org/10.1177/109467050023005>

- Diers, T. (2020). *Akzeptanz von Chatbots im Consumer-Marketing: Erfolgsfaktoren zwischen Konsumenten und künstlicher Intelligenz* [Acceptance of chatbots in consumer marketing: Success factors between consumers and artificial intelligence]. Springer. <https://doi.org/10.1007/978-3-658-29317-8>
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. <https://doi.org/10.1037/xge0000033>
- Fiore, S. M., Wiltshire, T. J., Lobato, E. J. C., Jentsch, F. G., Huang, W. H., & Axelrod, B. (2013). Toward understanding social cues and signals in human-robot interaction: Effects of robot gaze and proxemic behavior. *Frontiers in Psychology*, *4*(859), 2–15. <https://doi.org/10.3389/fpsyg.2013.00859>
- Fiske, S. T. (2018). Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science*, *27*(2), 67–73. <https://doi.org/10.1177/0963721417738825>
- Følstad, A., Araujo, T., Law, E. L.-C., Brandtzaeg, P. B., Papadopoulos, S., Reis, L., Baez, M., Laban, G., McAllister, P., Ischen, C., Wald, R., Catania, F., Meyer von Wolff, R., Hobert, S., & Luger, E. (2021). Future directions for chatbot research: An interdisciplinary research agenda. *Computing*, *103*, 2915–2942. <https://doi.org/10.1007/s00607-021-01016-7>
- Følstad, A., & Skjuve, M. (2019). Chatbots for customer service: User experience and motivation. In B. R. Cowan & L. Clark (Eds.), *Proceedings of the 1st International Conference on Conversational User Interfaces* (pp. 1–9). Association for Computing Machinery. <https://doi.org/10.1145/3342775.3342784>
- Franke, T., Attig, C., & Wessel, D. (2019). A personal resource for technology interaction: Development and validation of the affinity for technology interaction (ATI) Scale. *International Journal of Human-Computer Interaction*, *35*(6), 456–467. <https://doi.org/10.1080/10447318.2018.1456150>
- Gambino, A., Fox, J., & Ratan, R. (2020). Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication*, *1*, 71–86. <https://doi.org/10.30658/hmc.1.5>
- Gefen, D., & Straub, D. W. (2004). Consumer trust in B2C e-Commerce and the importance of social presence: Experiments in e-Products and e-Services. *Omega*, *32*(6), 407–424. <https://doi.org/10.1016/j.omega.2004.01.006>
- Go, E., & Sundar, S. S. (2019). Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior*, *97*, 304–316. <https://doi.org/10.1016/j.chb.2019.01.020>
- Gordon, V. N. (1994). Developmental advising: The Elusive ideal. *NACADA Journal*, *14*(2), 72–76. <https://doi.org/10.12930/NACADA-19-201>
- Greussing, E., Gaiser, F., Klein, S. H., Straßmann, C., Ischen, C., Eimler, S., Frehmann, K., Gieselmann, M., Knorr, C., Lermann Henestrosa, A., Räder, A., & Utz, S. (2022). Researching interactions between humans and machines: Methodological challenges. *Publizistik*, *67*(4), 531–554. <https://doi.org/10.1007/s11616-022-00759-3>
- Grimes, G. M., Schuetzler, R. M., & Giboney, J. S. (2021). Mental models and expectation violations in conversational AI interactions. *Decision Support Systems*, *144*, 113515. <https://doi.org/10.1016/j.dss.2021.113515>



- Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (3rd ed.). Guilford Press.
- Ischen, C., Araujo, T., van Noort, G., Voorveld, H., & Smit, E. (2020). "I am here to assist you today": The role of entity, interactivity and experiential perceptions in chatbot persuasion. *Journal of Broadcasting & Electronic Media*, 64(4), 615–639. <https://doi.org/10.1080/08838151.2020.1834297>
- Ischen, C., Smit, E., & Wang, E. (2023, November 22–23). *Assessing human-likeness perceptions: Measurement scales of conversational agents* [Paper presentation]. Conversations 2023 –7th International Workshop on Chatbot Research and Design, Oslo, Norway. Retrieved from [https://web.archive.org/web/20240418110620/https://2023.conversations.ws/wp-content/uploads/2023/11/conversations\\_2023\\_positionpaper\\_13\\_ischen.pdf](https://web.archive.org/web/20240418110620/https://2023.conversations.ws/wp-content/uploads/2023/11/conversations_2023_positionpaper_13_ischen.pdf)
- Jackson, S., & Jacobs, S. (1983). Generalizing about messages: Suggestions for design and analysis of experiments. *Human Communication Research*, 9(2), 169–191. <https://doi.org/10.1111/j.1468-2958.1983.tb00691.x>
- Karrenbauer, C., König, C. M., & Breitenner, M. H. (2021). Individual digital study assistant for higher education institutions: Status quo analysis and further research agenda. In F. Ahlemann, R. Schütte, & S. Stieglitz (Eds.), *Innovation Through Information Systems. WI 2021. Lecture Notes in Information Systems and Organisation* (vol. 48, pp. 108–124). Springer. [https://doi.org/10.1007/978-3-030-86800-0\\_8](https://doi.org/10.1007/978-3-030-86800-0_8)
- Kasneji, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneji, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kent, M. L., & Taylor, M. (2002). Toward a dialogic theory of public relations. *Public Relations Review*, 28(1), 21–37. [https://doi.org/10.1016/S0363-8111\(02\)00108-X](https://doi.org/10.1016/S0363-8111(02)00108-X)
- Lagace, R. R., Dahlstrom, R., & Gassenheimer, J. B. (1991). The relevance of ethical salesperson behavior on relationship quality: The pharmaceutical industry. *Journal of Personal Selling & Sales Management*, 11(4), 39–47. <https://doi.org/10.1080/08853134.1991.10753888>
- Lee, E.-J. (2023). Minding the source: Toward an integrative theory of human-machine communication. *Human Communication Research*, hqad034. <https://doi.org/10.1093/hcr/hqad034>
- Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1), 205395171875668. <https://doi.org/10.1177/2053951718756684>
- Lee, S., & Choi, J. (2017). Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. *International Journal of Human-Computer Studies*, 103, 95–105. <https://doi.org/10.1016/j.ijhcs.2017.02.005>
- Lee, S., Lee, N., & Sah, Y. J. (2020). Perceiving a mind in a chatbot: Effect of mind perception and social cues on co-presence, closeness, and intention to use. *International Journal of Human-Computer Interaction*, 36(10), 930–940. <https://doi.org/10.1080/10447318.2019.1699748>
-



- Lew, Z., & Walther, J. B. (2023). Social scripts and expectancy violations: Evaluating communication with human or AI chatbot interactants. *Media Psychology, 26*(1), 1–16. <https://doi.org/10.1080/15213269.2022.2084111>
- Lombard, M., & Xu, K. (2021). Social responses to media technologies in the 21st century: The media are social actors paradigm. *Human-Machine Communication, 2*, 29–55. <https://doi.org/10.30658/hmc.2.2>
- Lou, C., Kang, H., & Tse, C. H. (2021). Bots vs. humans: How schema congruity, contingency-based interactivity, and sympathy influence consumer perceptions and patronage intentions. *International Journal of Advertising, 41*(4), 1–30. <https://doi.org/10.1080/02650487.2021.1951510>
- Maisel, N., Gable, S. L., & Strachman, A. (2008). Responsive behaviors in good times and in bad. *Personal Relationships, 15*(3), 317–338. <https://doi.org/10.1111/j.1475-6811.2008.00201.x>
- Meyer von Wolff, R., Nörtemann, J., Hobert, S., & Schumann, M. (2020). Chatbots for the information acquisition at universities—A student’s view on the application area. In A. Følstad, T. Araujo, S. Papadopoulos, E. L.-C. Law, O.-C. Granmo, E. Luger, & P. B. Brandtzaeg (Chairs), *Chatbot Research and Design. CONVERSATIONS 2019, Lecture Notes in Computer Science* (vol. 11970, pp. 231–244). Springer. [https://doi.org/10.1007/978-3-030-39540-7\\_16](https://doi.org/10.1007/978-3-030-39540-7_16)
- Mori, M., MacDorman, K., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine, 19*(2), 98–100. <https://doi.org/10.1109/MRA.2012.2192811>
- Mottarella, K. E., Fritzsche, B. A., & Cerabino, K. C. (2004). What do students want in advising? A policy capturing study. *NACADA Journal, 24*(1 & 2), 48–61. <https://doi.org/10.12930/0271-9517-24.1-2.48>
- Mulac, A., Erlandson, K. T., Farrar, W. J., Hallett, J. S., Molloy, J. L., & Prescott, M. E. (1998). “Uh-huh. What’s that all about?” *Communication Research, 25*(6), 641–668. <https://doi.org/10.1177/009365098025006004>
- Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81–103. <https://doi.org/10.1111/0022-4537.00153>
- Neururer, M., Schlögl, S., Brinkschulte, L., & Groth, A. (2018). Perceptions on authenticity in chat bots. *Multimodal Technologies and Interaction, 2*(3), 60. <https://doi.org/10.3390/mti2030060>
- Oh, C. S., Bailenson, J. N., & Welch, G. F. (2018). A systematic review of social presence: Definition, antecedents, and implications. *Frontiers in Robotics and AI, 5*, 114. <https://doi.org/10.3389/frobt.2018.00114>
- Prahl, A., & Van Swol, L. (2021). Out with the humans, in with the machines? Investigating the behavioral and psychological effects of replacing human advisors with a machine. *Human-Machine Communication, 2*, 209–234. <https://doi.org/10.30658/hmc.2.11>
- Reis, H. T., & Clark, M. S. (2013). Responsiveness. In J. A. Simpson & L. Campbell (Eds.), *Oxford library of psychology. The Oxford handbook of close relationships* (pp. 400–423). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780195398694.013.0018>
- Roos, C. A., Postmes, T., & Koudenburg, N. (2023). Feeling heard: Operationalizing a key concept for social relations. *PLOS ONE, 18*(11), e0292865. <https://doi.org/10.1371/journal.pone.0292865>

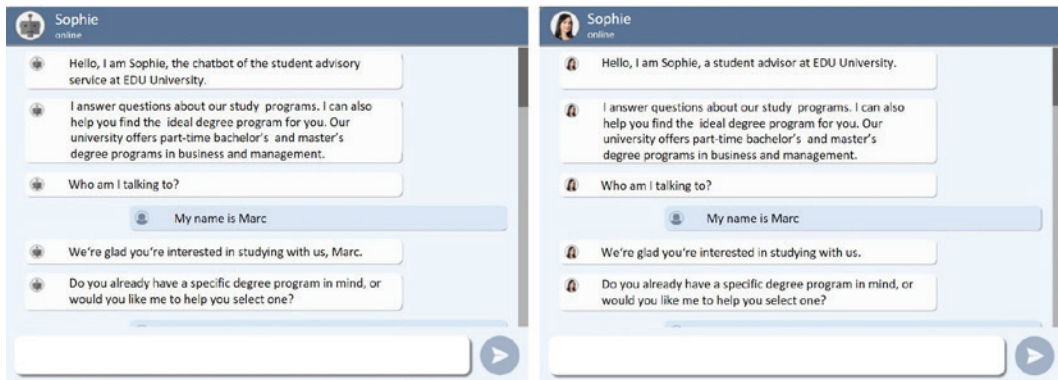
- Rossee, Y., Jorgensen, Terrence, D., & Rockwood, N. (2021). *Package "lavaan": Latent variable analysis* [computer software]. <https://doi.org/10.18637/jss.v048.i02>
- Shawar, B., & Atwell, E. (2007). Chatbots: Are they really useful? *Journal for Language Technology and Computational Linguistics*, 22(1), 29–49. <https://doi.org/10.21248/jlcl.22.2007.88>
- Spence, P. R., Westerman, D., Edwards, C., & Edwards, A. (2014). Welcoming our robot overlords: Initial expectations about interaction with a robot. *Communication Research Reports*, 31(3), 272–280. <https://doi.org/10.1080/08824096.2014.924337>
- Sundar, S. S. (2008). The MAIN Model: A heuristic approach to understanding technology effects on credibility. In M. J. Metzger & A. J. Flanagin (Eds.), *Digital media, youth, and credibility* (pp. 73–100). MIT Press. <https://betterlegalinfo.ca/wp-content/uploads/2019/12/Sundar-paper.pdf>
- Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human-AI interaction (HAI). *Journal of Computer-Mediated Communication*, 25(1), 74–88. <https://doi.org/10.1093/jcmc/zmz026>
- Sundar, S. S., Bellur, S., Oh, J., Jia, H., & Kim, H.-S. (2016). Theoretical importance of contingency in human-computer interaction. *Communication Research*, 43(5), 595–625. <https://doi.org/10.1177/0093650214534962>
- Sundar, S. S., & Nass, C. (2001). Conceptualizing sources in online news. *Journal of Communication*, 51(1), 52–72. <https://doi.org/10.1111/j.1460-2466.2001.tb02872.x>
- Van der Goot, M., & Etzrodt, K. (2023). Disentangling two fundamental paradigms in human-machine communication research: Media equation and media evocation. *Human-Machine Communication*, 6, 17–30. <https://doi.org/10.30658/hmc.6.2>
- Van Pinxteren, M. M., Pluymaekers, M., & Lemmink, J. G. (2020). Human-like communication in conversational agents: A literature review and research agenda. *Journal of Service Management*, 31(2), 203–225. <https://doi.org/10.1108/JOSM-06-2019-0175>
- Venkatesh, V. (2000). Determinants of perceived ease of use: Integrating control, intrinsic motivation, and emotion into the technology acceptance model. *Information Systems Research*, 11(4), 342–365. <https://doi.org/10.1287/isre.11.4.342.11872>
- Westland, J. C. (2015). *Structural equation models: From paths to networks*. Springer. <https://doi.org/10.1007/978-3-319-16507-3>
-

## Appendices

### Appendix A—Power Analyses

The sample size of Study 1 ( $N = 256$ ) was determined by an a priori power analysis ( $f = .18$ ,  $\alpha = .05$ ,  $1-\beta = .80$ ) for an interaction effect resulting from an ANOVA. The power analysis was based on the small interaction effect ( $\eta^2_{\text{partial}} = .03$ ) found by Go & Sundar (2019) in an experiment on the effects of identity and conversational cues on attitude toward the website where the chatbot is placed, using real interactions with chat agents. As we expected effects in Studies 2 and 3 to be smaller than Go and Sundar's interaction effect, we determined a sample size of  $N = 403$  a priori for a significant interaction effect of  $f = .14$  ( $\alpha = .05$ ,  $1-\beta = .80$ ).

### Appendix B—Screenshots of Exemplary, Translated Chat Conversations in Chatbot and Human Conditions



## Appendix C—English Translations of Items and Alpha Values for Variables Across Studies

Variable and Items	Cronbach's $\alpha$ in Study		
	1	2	3
<b>Attitude</b>	.95	.95	.91
1. I find it attractive to communicate with an organization in this way.			
2. I find it useful to communicate with an organization in this way.			
3. I find this way of communicating with an organization interesting.			
4. I find it helpful to communicate in this way with an organization.			
5. Communicating in this way with an organization helps me to meet my needs.			
<b>Likability</b>	.92	.87	.90
Please rate your impression of Sophie on these scales: [dislike–like, unfriendly–friendly, unkind–kind, unpleasant–pleasant, awful–nice].			
<b>Perceived intelligence</b>	.88	.90	.92
Please rate your impression of Sophie on these scales: [incompetent–competent, ignorant–knowledgeable, irresponsible–responsible, unintelligent–intelligent, foolish–sensible].			
<b>Warmth</b>	.87	.88	.91
How [warm, trustworthy, friendly, honest, likable, sincere] do you think Sophie was?			
<b>Competence</b>	.88	.86	.91
How [competent, intelligent, skilled, efficient, assertive, confident] do you think Sophie was?			
<b>Perceived enjoyment</b>	.90	.89	.91
1. The conversation evokes positive feelings in me.			
2. I found the conversation entertaining.			
3. I enjoyed reading the conversation.			
<b>Satisfaction</b>	.92	.91	.94
1. I would be happy with Sophie's recommendations for courses of study.			
2. I would be satisfied with the way Sophie spoke to Marc.			
3. I would be satisfied with the information Sophie gave Marc.			
4. I would be satisfied with the conversation Marc had with Sophie.			
<b>Intention to use</b>	.96	.96	.97
1. If an organization offers this possibility of communication, I will use it.			
2. If I have the opportunity to communicate with an organization in this way, I will.			
3. I am very likely to use this way of communicating with an organization.			
4. Once this way of communicating with an organization is established, it will be my preferred method.			

Variable and Items	Cronbach's $\alpha$ in Study		
	1	2	3
<b>Social presence</b>	.94	.94	.96
1. There was a sense of human contact in the interaction.			
2. There was a sense of personalness in the interaction.			
3. There was a feeling of sociability in the interaction.			
4. There was a feeling of human warmth in the interaction.			
5. There was a feeling of human sensitivity in the interaction.			
<b>Perceived dialogue</b>	.85	.82	.86
1. I had the feeling that Sophie was in an active dialogue with Marc.			
2. Marc's interactions with Sophie felt like a back-and-forth conversation.			
3. I felt that Sophie and Marc were involved in a joint task when choosing a program.			
4. Sophie was quick to respond to Marc's input and requests.			
5. I felt that Sophie took Marc's individual wishes into account.			
<b>Feeling heard</b>	.84	.82	.86
1. Marc felt heard.			
2. Marc was able to say what he really wanted to say.			
3. Sophie seemed to care more about something else than what Marc said.			
4. Sophie listened to Marc.			
5. Sophie tried to put herself in Marc's shoes.			
6. Sophie seemed insensitive to Marc's thoughts and feelings.			
7. Sophie treated Marc with respect.			
8. Sophie and Marc understood each other.			
<b>Agent type manipulation check</b>			
If you think back to the chat interaction you just saw: Who was Marc talking to?	—	—	—
1 = the study advisor Sophie, 2 = the professor Sophie, 3 = the chatbot Sophie, 4 = the doctor Sophie, 5 = don't know			
<b>Responsiveness manipulation check</b>	.77	.65	.59
1. Study 1: Sophie used affirmative expressions to indicate that she was really listening to Marc.			
2. Studies 2, 3: Sophie used affirmative expressions to indicate that she was listening to Marc.			
3. Study 1: Sophie appropriately picked up on what Marc said in her response.			
4. Studies 2, 3: Sophie picked up on what Marc said in her response.			

Note. 7-point Likert-type rating scales (1 = do not agree at all, 7 = fully agree), except likability, perceived intelligence (7-point semantic differentials) and the agent type manipulation check.

**Appendix D—Means and Standard Deviations for Variables Across Studies**

Variable	Study 1		Study 2		Study 3		Pooled Data	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
<b>Dependent Variables</b>								
Attitude	4.79	1.58	4.18	1.67	4.76	1.61	4.53	1.65
Likability	5.68	1.08	5.53	1.06	5.72	0.96	5.63	1.04
Perceived intelligence	5.90	0.88	5.43	1.13	5.66	1.00	5.63	1.04
Warmth	5.35	0.99	5.01	1.11	5.29	1.11	5.19	1.09
Competence	5.73	0.90	5.43	0.98	5.54	1.06	5.54	1.00
Perceived enjoyment	4.29	1.42	3.78	1.51	4.55	1.50	4.18	1.52
Satisfaction	5.34	1.38	4.83	1.52	5.30	1.39	5.12	1.46
Intention to use	4.74	1.76	4.04	1.84	4.80	1.67	4.49	1.80
<b>Mediators</b>								
Social presence	3.80	1.52	3.58	1.51	4.35	1.52	3.91	1.55
Perceived dialogue	5.28	1.16	5.13	1.19	5.49	1.10	5.29	1.16
Feeling heard	5.43	0.92	5.30	0.95	5.48	0.97	5.39	0.95

Note.  $N_1 = 253$ ,  $N_2 = 401$ ,  $N_3 = 351$ ,  $N_{\text{total}} = 1,005$ .



## Appendix E—Bivariate Correlations Between Variables Across Studies

Variable	1	2	3	4	5	6	7	8	9	10	11
1. Attitude	—										
2. Likability	.39	—									
	.43										
	.49										
3. Perceived intelligence	.47	.70	—								
	.38	.59									
	.54	.73									
4. Warmth	.57	.69	.63	—							
	.56	.68	.52								
	.60	.74	.71								
5. Competence	.48	.44	.67	.66	—						
	.50	.49	.54	.67							
	.59	.64	.76	.80							
6. Perceived enjoyment	.61	.57	.49	.65	.45	—					
	.57	.46	.31	.61	.49						
	.65	.59	.56	.73	.60						
7. Satisfaction	.61	.58	.66	.72	.68	.61	—				
	.65	.58	.52	.72	.68	.63					
	.70	.63	.64	.76	.73	.71					
8. Intention to use	.89	.36	.41	.48	.44	.56	.53	—			
	.86	.38	.32	.49	.42	.54	.57				
	.91	.48	.50	.57	.54	.59	.63				
9. Social presence	.54	.51	.48	.68	.50	.71	.49	.57	—		
	.51	.53	.42	.63	.50	.66	.46	.61			
	.59	.52	.50	.67	.55	.70	.57	.59			
10. Perceived dialogue	.52	.56	.51	.73	.61	.55	.50	.71	.57	—	
	.49	.56	.46	.69	.59	.54	.43	.64	.63		
	.51	.61	.64	.72	.73	.59	.48	.68	.57		
11. Feeling heard	.52	.56	.53	.71	.59	.46	.45	.67	.53	.74	—
	.43	.55	.42	.67	.58	.47	.36	.59	.41	.79	
	.46	.66	.62	.74	.72	.53	.42	.66	.51	.79	

Note. Pearson's correlations  $r$ . Grey shaded cells: 1st line = Study 1 ( $N = 253$ ), 2nd line = Study 2 ( $N = 401$ ), 3rd line = Study 3 ( $N = 351$ ). All correlations are significant at  $p < .001$ .



# The Impact of Human-AI Relationship Perception on Voice Shopping Intentions

Marisa Tschopp<sup>1,2</sup> , and Kai Sassenberg<sup>2,3</sup> 

1 scip AG

2 Leibniz-Institut für Wissensmedien (IWM), Tübingen, Germany

2 Leibniz-Institut für Psychologie (ZPID), Trier, Germany

## Abstract

In the emerging field of voice shopping with quasi-sales agents like Amazon's Alexa, we investigated the influence of perceived human-AI relationships (i.e., authority ranking, market pricing, peer bonding) on (voice-)shopping intentions. In our cross-sectional survey among experienced voice shoppers ( $N = 423$ ), we tested hypotheses specifically differentiating voice shopping for low- and high-involvement products. The results emphasized the importance of socio-emotional elements (i.e., peer bonding) for voice shopping for high-involvement products. While calculative decision-making (i.e., market pricing) was less relevant, the master-servant relationship perception (i.e., authority ranking) was important in low-involvement shopping. An exploratory analysis of users' desired benefits of voice shopping reinforces our claims. The outcomes are relevant for conversation designers, business developers, and policymakers.

**Keywords:** voice shopping; human-AI relationship; conversational AI; high- and low-involvement; perceived benefits

## Introduction

With the introduction of online shopping, people could purchase almost anything with a few clicks. Three decades later, people can just *tell* a computer to place an order. Although voice shopping is a form of e-commerce, it substantially differs from traditional online shopping (Klaus & Zaichkowsky, 2022). We argue that voice shopping with a conversational

**CONTACT** Marisa Tschopp  • [mats@scip.ch](mailto:mats@scip.ch) • scip AG • Badenerstrasse 623 • 8048 Zurich, Switzerland

ISSN 2638-602X (print)/ISSN 2638-6038 (online)  
[www.hmcjournal.com](http://www.hmcjournal.com)



Copyright 2024 Authors. Published under a Creative Commons Attribution 4.0 International (CC BY-NC-ND 4.0) license.

artificial intelligence (AI) is conceptually more similar to decision-making in a brick-and-mortar store involving in-person interactions with human salespeople and should be investigated as such.

Research on relationships between the consumer and (human) seller is popular in the marketing literature (Alvarez & Fournier, 2016). For example, studies have shown that a positive seller-buyer relationship leads to greater brand trust and more positive affect by consumers (Carroll & Ahuvia, 2006; Chaudhuri & Holbrook, 2002). But the relationship perspective has yet not been translated into human-AI interaction, investigating the perception of conversational AI as quasi-sales agents whom consumers form some sort of relationship with (e.g., Lim et al., 2022; Ramadan, 2021; Rhee & Choi, 2020). In fact, research precisely on human-AI relationships is, in general, still nascent (Pentina et al., 2023), and the few existing findings paint a complex picture.

Hu et al. (2022) found that people who see their conversational AI mostly as assisting them have stronger voice shopping intentions, motivated by a hierarchical power experience over their voice assistants, a claim supported by Tassiello et al. (2021). While Hu et al. (2022) did not differentiate what users bought, Tassiello et al. as well as Rhee and Choi (2020) did. Both used the concept of low- and high-involvement: Low-involvement products are characterized as low-cost items consumers tend to consider without extensive deliberation, in contrast to high-involvement products, which are typically pricier and necessitate thorough evaluation (Mari & Algesheimer, 2021; Rhee & Choi, 2020; Tassiello et al., 2021). In the development of their hypotheses, they argue that shoppers think differently about the products, requiring different persuasive messages to facilitate voice shopping. Contrasting Hu et al.'s (2021) and Tassiello et al.'s (2020) findings, Rhee and Choi (2020) found that a friend-like voice shopping user interface increased voice shopping intentions for low-involvement products.

These partly inconsistent findings call for further research, including the nature of the perceived relationship and the purchase. Therefore, we apply a multidimensional human-AI relationship model while differentiating between low- and high-involvement products. Assuming that users perceive their relationship to their conversational AI not just along a friend or servant dimension but along several dimensions, as suggested by Tschopp et al. (2023), holds promise in gaining differentiated insights into users' voice shopping behavior and addressing the contradictions in the current landscape. Thus, the focal question of this study is: Does the way users relate to their conversational AI influence what kind of products they buy?

## **How Users Perceive Their Relationship to Conversational AI**

The remarkable progress in AI in the past decades has steadily stretched the boundaries of human-AI interaction and communication, demonstrated by the developments of language models. These advancements have rendered users' interactions not only social in the sense of being imbued with meaning or emotion but have also expanded the potential for the establishment of what might be considered relationships with AI agents, as asserted by Pentina et al. (2023).

To examine the human-AI relationship perception from a multidimensional perspective, we are building upon Tschopp et al.'s (2023) adaptation of the Relational Models

---

**TABLE 1** Description of the Three Modes of Human-AI Relationships  
(based on Tschopp et al., 2023)

Peer Bonding	Market Pricing	Authority Ranking
<ul style="list-style-type: none"> <li>▶ Most human-like dimension where the user treats the conversational AI as an equivalent peer.</li> <li>▶ Best characterized as a communal relationship.</li> </ul>	<ul style="list-style-type: none"> <li>▶ The user perception is guided by cost-benefit analyses with no hierarchies.</li> <li>▶ Best characterized as an exchange relationship on <i>eye level</i>.</li> </ul>	<ul style="list-style-type: none"> <li>▶ A hierarchical order is perceived between users and the conversational AI.</li> <li>▶ Best characterized as a master-servant relationship.</li> </ul>
The user tends to feel emotionally closer to the system.	The user tends to care about competence and rational trust in the system.	The user tends to use the system for a greater variety of purposes.

Theory (RMT) by Alan P. Fiske (Haslam & Fiske, 1999) to human-AI relationships. RMT is a theory on how humans construe their relationships with other humans. RMT describes four dimensions and has received mighty empirical support in the past decades. These four dimensions are (1) communal sharing (i.e., a kinship-like relationship as it is with families based on mutual trust), (2) equality matching (i.e., a tit-for-tat-like relationship as with roommates in a shared flat where equal give-and-take is key), (3) authority ranking (i.e., a hierarchical relationship characterized by a clear chain of command like soldiers and their superiors), and (4) market pricing (i.e., a currency-based relationship characterized by cost-benefit analyses as it is with employers and their bosses in a workplace).

Applying RMT, Tschopp et al. (2023), found that human-AI relationships are perceived along three dimensions varying in emotional breadth and perceived agency. Communal sharing and equality matching merged into one emotional dimension named peer bonding (see Table 1). They found that conversational AI users characterized their relationship mostly by authority ranking (i.e., a hierarchical owner-assistant relationship) and market pricing (i.e., a nonhierarchical exchange relationship) and least by peer bonding (i.e., a peer-like relationship). Notably, authority ranking was not informative for variables concerning system perception (e.g., trust, perceived intelligence, or affinity to technology). The two rather interactive dimensions (i.e., market pricing and peer bonding) had stronger predictive values, especially regarding anthropomorphism (Tschopp et al., 2023), which drives the development of our hypotheses and research questions.

While the initial work by Tschopp et al. (2023) remained exploratory, we aim to further investigate their assumptions in an applied context, namely voice shopping. This context presents an intriguing opportunity because multiturn dialogues are necessary to make a purchase decision. In other words, you have to actually communicate with the conversational AI and not only give orders, such as turning off the lights, where other relational dynamics may be involved.

Peer bonding, often regarded as the most emotionally charged connection, involves regarding the partner as an equal and companion-like figure while also upholding a sense of responsibility for one's conduct (Tschopp et al., 2023). Arguably, for people who see their

device through this relationship mode, the voice shopping experience would be more like shopping with a peer.

The newly introduced perception of conversational AI as a rational exchange partner, called market pricing, was found to be rather popular (Tschopp et al., 2023). Its core characteristic lies in the reliance on ratio values, devoid of hierarchies, thus resembling an equal-other, granted some sort of agency. Arguably, for people who see their device through this relationship mode, the voice shopping experience would be more like having a professional sales agent making the shopping decision together with the consumer.

The majority of respondents perceived their devices as authority ranking. The key characteristic of this arrangement is the creation of a linear hierarchy between humans and the conversational AI. For people who see their device through this relationship mode, the shopping experience would be more like shopping with a subservient helper or concierge. However, before making such assumptions, a better understanding of voice commerce is necessary.

## Shopping via Conversational AI

Voice shopping, or voice commerce, is an emerging commercial trading system where, for instance, Alexa users (Amazon's conversational AI) can search, purchase, and track products on Amazon solely through a voice user interface (VUI) (Halbauer & Klarmann, 2022; Ramadan, 2021). Alexa shoppers predominantly purchase entertainment products (such as music or books), household essentials (like batteries or toilet paper), and clothing, whereby re-purchases and new orders occur with equal frequency (for a comprehensive breakdown of product categories, see Kinsella, 2018). Practitioners are eager to leverage this new sales channel. However, research in the field is in its infancy, with limited empirical data on what promotes or hinders voice shopping scattered across disciplines (Klaus & Zaichkowsky, 2022; Lim et al., 2022).

From a psychological perspective, initial studies have investigated what drives voice shopping intentions. Trust (Huh et al., 2023; Mari & Algesheimer, 2021), perceived human-likeness/anthropomorphism (Han, 2021; Huh et al., 2023), perceptions of social presence, emotional bonding, and para-social interaction and dialogue (Ramadan, 2021), were found to have a positive influence on voice shopping intentions and continuance. These studies stress the importance of the social dimension in voice purchasing behavior. Especially with regard to the voice shopping process, the increasing interactive verbal decision-making processes and two-way interaction render "voice assistants partners in the decision-making dialogue rather than mere order takers" (de Bellis & Venkataramani Johar, 2020; Dellaert et al., 2020)

Furthermore, only a limited number of empirical studies have distinguished voice shopping intentions based on the specific products individuals purchase, which likely engage distinct processes as comprehensively laid out by Rhee and Choi (2020). In simpler terms, it is highly likely that there is a notable distinction between buying batteries and purchasing a laptop through voice commands, where there is limited access to information and a varying necessity to rely on the AI as a sales agent.

When using conversational AI for product selection, Klaus and Zaichkowsky (2022) suggest that the algorithm serves distinct purposes based on the complexity and functionality

---



of the product. In their model, they differentiate high- and low-involvement situations, where the algorithm serves different functions depending on whether the product is simpler and more functional (i.e., low-involvement). This entails a more utilitarian approach, where users allow the conversational AI to handle the purchase. This concept was also applied in a study by Mari and Algesheimer (2021), who selected batteries as a low-involvement product, invoking the “yeah, whatever” heuristic. In contrast, the decision-making process for intricate, costly, and/or high-risk products, as outlined in Klaus and Zaichkowsky’s model, appears quite different. When acquiring items like a \$500 vacuum cleaner, more information and guidance are necessary, making them high-involvement purchases that demand greater time and effort for decision-making. In this framework, an algorithm aids the buyer in making the most informed shopping decision collaboratively.

Against this background and given the inclination of people to respond to technological systems in social ways (Nass & Moon, 2000) and the empirical importance of the social dimensions as antecedents of (voice) shopping decisions, it is rather surprising that only a few studies have looked at the impact of perceived relationship to the conversational AI on home shopping behavior. Much research has focused on relational proxies, assessing constructs such as perceived warmth, psychological distance, or anthropomorphism (e.g., Gong, 2008; Pitardi & Marriott, 2021) or role ascriptions (e.g., Sundar et al., 2017). Furthermore, and as mentioned above, inconsistent results raise further questions: Hu et al. (2022) have found that presenting conversational AI as servants enabled a power experience for users as masters and increased voice shopping intentions (given that they had a desire for power). Similarly, an experimental study by Tassiello et al. (2021) found that the subservient assistant role facilitated voice shopping. On the other hand, Rhee and Choi (2020) found that a friend-like social design had a positive influence on voice shopping intentions. Notably, this was particularly important for buying low-involvement products. These findings underscore the need for further research to carefully examine and dissect voice shopping intentions, particularly by distinguishing between different types of products that involve varying levels of involvement in the purchase decision-making process.

## Hypotheses Development

### *Does the Perceived Human-AI Relationship Influence Voice Shopping Intentions?*

Dellaert et al.’s (2020) argument that virtual assistants serve as partners in decision-making suggests that peer bonding and market pricing are highly relevant for voice shopping, more so than authority ranking. To reiterate, a large amount of research suggests that human-like system perception variables such as perceived human-likeness (Huh et al., 2023) or emotional bonding (Ramadan, 2021) are promoting shopping intentions. We thus predict:

**H1:** Higher values in peer bonding predict a stronger intention to use voice shopping.

Market pricing, the non-hierarchical relationship dimension characterized by exchange and interaction, is emotionally less pronounced. However, market pricing still constitutes a human-like relationship, in the sense that it requires that users attribute agency to the system and see their conversational AI rather as an exchange partner whom they meet on “eye

level” than as a tool. Relying on the fact that human-like perceptions of conversational AI go hand in hand with voice shopping intentions (Huh et al., 2023), we also expect:

**H2:** Higher values in market pricing predict a stronger intention to use voice shopping.

Based on the rationale that the conversational AI functions as a sales agent rather than a simple order processor, and considering the absence of predictive information regarding authority ranking as per Tschopp et al.’s study (2023), we posed the influence of authority ranking as an exploratory research question in our preregistration. The results were analyzed in an equitable manner within our results section.

**RQ1:** How does authority ranking associate with general voice shopping intentions?

### **Different Predictors for Different Products?**

We argue that different relationship dimensions will predict shopping intentions for different products because people evaluate products differently. Inspired by Rhee and Choi’s (2020) arguments, this rationale is based on the elaboration likelihood model (ELM, Petty & Cacioppo, 1986), which distinguishes two routes. The *peripheral route* is characterized by a low amount of effort taken to process product information, but it could also be based on evaluating characteristics of the seller (see also Rhee & Choi, 2020). The peripheral route is typically used for low-involvement items, which are often cheap and interchangeable products (e.g., toilet paper or chewing gum; see Rhee & Choi, 2020). In other words, when a shopping decision bears no real risk, people do not think a lot but follow intuitions and emotions. This focus on intuition and emotions resonates with peer bonding, which is characterized by emotions and similar to a relationship with human peers whom people follow intuitively without much thought. This is in line with the study by Rhee and Choi that demonstrated the positive effect of a friend-like social design on shopping for low-involvement products but not for high-involvement products.

The *central route* is used for more cognitively demanding products. This form of information processing is characterized by careful elaboration of the quality of arguments, facts, or figures (Petty & Cacioppo, 1986). This cognitive effort is typically only invested when the motivation to process the information is high, in other words, in a shopping context in which more is at stake—financially or personally. This should apply in the case of high-involvement products. When voice shopping for high-involvement products, the decision-making process resembles the central route. Voice shoppers should be highly motivated to evaluate product characteristics and rationality should dominate in a “cost-benefits-analysis style.” This style fits a market pricing relationship based on cost-benefit analysis. Taken together, the intuitive and emotional processing style applied when shopping low-involvement products resonates with peer bonding, whereas the cost-benefit-analysis style applied when buying high-involvement products resonates with market pricing (see Table 1). We thus predict:

---

**H3:** The intention to buy low-involvement products via voice shopping is predicted to a stronger extent by peer bonding than by market pricing.

**H4:** The intention to buy high-involvement products via voice shopping is predicted to a stronger extent by market pricing than by peer bonding.

As before, we posed an exploratory question regarding the role of authority ranking:

**RQ2:** How does authority ranking associate with voice shopping intentions for low- and high-involvement products?

To situate the relational approach into common customer value frameworks, we assessed what people care about in voice shopping. We looked at desired hedonic, utilitarian, symbolic, and social benefits (inspired by McLean & Osei-Frimpong, 2019) and how they associate with voice shopping intentions and human-AI relationships. We anticipate that the exploratory analysis will provide conceptual reinforcement for our findings. Given the early stage of the field, it is premature to make definitive predictions and thus commit to the exploration of our research question.

**RQ3:** How do desired shopping benefits associate with the human-AI relationship perception and voice shopping intentions?

## Methods

### Design and Participants

We conducted a preregistered cross-sectional study to test our hypotheses <https://aspredicted.org/2pg28.pdf>. The study was run online via Prolific in July 2022. We aimed at a sample of 450 based on the assumption that  $N = 250$  is required for stable correlations (Schönbrodt & Perugini, 2013). We added 200 participants to definitely end up with  $N > 250$ , even in case of substantial exclusions. We preregistered the following exclusion criteria: no experience in voice shopping, failing at least one attention check, and too short ( $< 150$  seconds) or too long ( $> 80,000$  seconds) duration of the survey. In a prescreening, we surveyed people ( $N_{total} = 800$ ) to identify potential participants engaging regularly in voice shopping with conversational AIs such as Alexa. We collected data from 451 participants fulfilling this criterion in exchange for £1.10. Twenty-eight participants were excluded based on the criteria mentioned above or because they were outliers with an absolute studentized deleted residual  $> 2.59$  in the regression testing (H1 and 2), another preregistered exclusion criterion. The remaining respondents  $N = 423$  (57% female, 42%, male, 1% other; age  $M = 41$ ,  $SD = 11.4$ , age range 19–84 years) responded to the questionnaire regarding their use of Alexa (78%), Google Assistant (16%), Siri (5%), or other conversational AI (1%). More information about users' voice shopping preferences can be found in the supplement. A sensitivity analysis for a single predictor in multiple regression analysis with three predictors (the analysis for the main predictions) indicated that the sample size was sufficient to detect an effect of  $f^2 = .018$  at  $\alpha = .05$  and  $1 - \beta = .8$ .

## Procedure

We invited participants to take part in a study on users' perceptions of voice shopping. After providing consent, participants had to choose which conversational AI their answers referred to and then respond to the human-AI relationship questionnaire (adapted from Haslam & Fiske, 1999; see Tschopp et al., 2023). The instructions for the measure require people to focus on a specific device when reporting their relationship. Afterward, we surveyed users about their shopping intentions to test the predictions. Variables were presented in a fixed order. All items were randomized. Next, exploratory variables were assessed. Perceived and desired benefits, trust, and user characteristics (device specifics, frequency of and experience in voice shopping, estimated voice shopping spending per year). We placed questions for demographic information and a final opportunity to withdraw their data at the end. Analyses have been conducted using SPSS 25.0 unless reported otherwise. Supplemental data, code, data, and pre-registration are available at <https://researchbox.org/1029>.

## Measures

**Human-AI relationship** was assessed using the questionnaire by Tschopp et al. (2023). Administering the questionnaire involves a specific mandatory procedure. Responding to the Human-AI relationship questionnaire necessitates first choosing a voice assistant their answers refer to (e.g., Alexa or Google Assistant). After selecting their preferred assistant, participants were directed to reflect on past shopping experiences and rate the extent to which items described their relationship with the chosen assistant in mind. The questionnaire consisted of 17 items using a 7-point Likert scale (1 = *not at all true for this relationship*, 7 = *very true for this relationship*): nine items for peer bonding, four items for authority ranking, and four items for market pricing. A principal component analysis (PCA) with varimax rotation was conducted (see Table 2). The three factor solution (based on the Kaiser criterion) explained 56.42% of the variance. As in prior studies, the first component represents peer bonding, the second component authority ranking, and the third component market pricing. Due to high loadings ( $> .4$ ) on a factor they were not intended to correlate with, we omitted items 6 and 17. The final scales presented sufficient reliabilities: *Cronbach's Alpha* = .91 for peer bonding, *Alpha* = .71 for authority ranking, and *Alpha* = .66 for market pricing. Market pricing was positively correlated with peer bonding ( $r = .47$ ,  $N = 423$ ,  $p < .001$ ) and authority ranking ( $r = .27$ ,  $N = 423$ ,  $p < .001$ ). No significant correlation was found between authority ranking and peer bonding ( $r = -.09$ ,  $N = 423$ ,  $p = .073$ ).

**Voice shopping.** We measured the *general intention to continue voice shopping* with three items adapted from McLean and Osei-Frimpong (2019). Respondents indicated their agreement on a 7-point Likert scale (3 items, 1 = *strongly disagree* to 7 = *strongly agree*). For instance, "I plan to continue to use the conversational AI for shopping in the future." An index was formed by averaging the responses (*Cronbach's Alpha* = .98).

**TABLE 2 Results From a Factor Analysis of the Human-AI Relationship Questionnaire (N = 423)**

	Item	Factor Loading		
		1	2	3
<b>Peer Bonding</b>				
1	There is a moral obligation to act kindly to each other	<b>.550</b>		.387
2	Decisions are made together by consensus	<b>.771</b>		
3	You tend to develop similar attitudes and behaviors	<b>.756</b>		
4	It seems you have something unique in common	<b>.839</b>		
5	You two are like a unit: you belong together	<b>.784</b>		
6	You are like tit for tat: you do something and expect something similar in return	<b>.487</b>		.425
7	Everyone has an equal say when a decision is made	<b>.780</b>		
8	You take turns doing what the other wants	<b>.786</b>		
9	You are like peers or fellow co-partners	<b>.783</b>		
<b>Authority Ranking</b>				
10	One of us is entitled to more than the other		<b>.701</b>	
11	One directs the work, the other pretty much follows		<b>.675</b>	
12	You are like leader and follower		<b>.691</b>	
13	One is above the other in a kind of hierarchy		<b>.745</b>	
<b>Market Pricing</b>				
14	What you get from this interaction is directly proportional to how much you give			<b>.661</b>
15	You have a right to a fair rate of return for what you put into this interaction			<b>.733</b>
16	You expect the same return on your effort other people get			<b>.740</b>
17	Your interaction is a strictly rational cost-benefit analysis		<b>.536</b>	

Note. Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization. Rotation converged in five iterations. The highest factor loadings are in bold, factor loadings below .30 are not displayed.

*Intention to continue voice shopping for low-involvement products* and the *intention to continue voice shopping for high-involvement products* were assessed with a single item each on a 7-point Likert scale (1 = *strongly disagree* to 7 = *strongly agree*). Participants read the description, see Table 3, and rated their agreement “I predict I would continue to use the conversational AI for shopping in the future.” General voice shopping intentions were highly correlated with low-involvement shopping intentions ( $r = .81, N = 423, p < .001$ ) and moderately with high-involvement shopping intentions ( $r = .41, N = 423, p < .001$ ). Using the three indicators is supported by a principal component analysis (see supplement).

**TABLE 3** Description of Low- and High-Involvement Shopping Intentions

<b>Intention to continue voice shopping for low-involvement products</b>	<b>Intention to continue voice shopping for high-involvement products</b>
Think about your future voice shopping experiences. Would you use the voice assistant to shop for products, which are rather convenience products, that require no effort to buy, and there are no emotional values or risks attached? For example, products such as paper towels, chewing gum, cereals, or a specific book. Please rate the extent to which these statements describe your intention to continue purchasing these types of products with your voice assistant in the future.	Think about your future voice shopping experiences. Would you use the voice assistant to shop for products which are rather complicated and require some effort to make a decision, with higher emotional values or risks attached? For example, a laptop, a smartphone, a vehicle, or a tablet. Please rate the extent to which these statements describe your intention to continue purchasing these types of products with your voice assistant in the future.

We tested these instructions in a pretest. In response to the high-involvement product description, people bought items such as laptops or smartphones, jewelry, or clothes. In response to the low-involvement product description, people reported household items such as toilet paper or soap, books, or groceries. Thus, the instructions seem to work as intended (see supplement).

**Desired benefits** were assessed with a 10 items scale measuring hedonic, utilitarian, symbolic, and social benefits inspired by McLean and Osei-Frimpong (2019). Respondents rated their agreement on a 7-point Likert scale (1 = *strongly disagree* to 7 = *strongly agree*). Because the original questionnaire assessed *actual* rather than *desired* benefits, we performed a factor analysis supporting the intended four-factor structure (see supplement). Two items assessed hedonic benefits (e.g., “It is important to me to have fun while shopping with my voice assistant”,  $r = .60$ ,  $N = 423$ ,  $p < .001$ ), four items utilitarian benefits (e.g., “It is important to me that the voice assistant makes shopping more efficient,” *Cronbach’s Alpha* = .84), two items symbolic benefits (e.g., “It is important to me that shopping with my voice assistant enhances my image among my peers,”  $r = .82$ ,  $N = 423$ ,  $p < .001$ ), and two items measured social benefits (e.g., “I care that shopping with a voice assistant is like dealing with a real person,”  $r = .77$ ,  $N = 423$ ,  $p < .001$ ).

**User characteristics.** We assessed participants’ use of smart speaker or tablet, screen use, and voice shopping spendings (see Table in the supplement). We measured *frequency of use* (“How often do you use voice assistant for shopping?”) on a single-item 6-point scale from 1 = *almost daily* to 5 = *1–2 times per year* (including an option 6 = *not at all*, ensuring to only survey experienced voice shoppers). *Experience of use* (“Since when do you use voice assistant for shopping purposes?”) was measured on a scale ranging from 1 = *5 years or more* to 6 = *less than 12 months*.



## Results

### Preliminary Analysis

We conducted an ANOVA with repeated measures and post-hoc comparison using Bonferroni correction to test for differences between the dimensions of the relationship perception. Participants saw their relationship with the conversational AI as more strongly characterized by authority ranking ( $M = 4.85$ ,  $SD = 1.38$ ,  $N = 423$ ) than by market pricing ( $M = 4.42$ ,  $SD = 1.43$ ,  $N = 423$ ) and peer bonding ( $M = 2.61$ ,  $SD = 1.31$ ,  $N = 423$ ), all  $ps < .001$ ,  $F(1.76, 422.00) = 402.28$ ,  $p < .001$ ,  $\eta^2_{part} = .488$  (with Huyn-Feldt correction). For all descriptive results, see Table 4 below.

**TABLE 4 Means, Standard Deviations, and Bivariate Correlations ( $N = 423$ )**

Scale	<i>M</i>	<i>SD</i>	Human-AI Relationship			Voice Shopping			Desired Benefits		
			PB	AR	MP	GI	LI	HI	HB	UB	SyB
<b>Human-AI Relationship</b>											
Peer Pondering (PB)	2.61	1.31									
Authority Ranking (AR)	4.85	1.38	-.09								
Market Pricing (MP)	4.42	1.43	.47**	.27**							
<b>Voice Shopping</b>											
General Continuance Intention (GI)	5.31	1.3	.20**	.10*	.18**						
Continuance Intention Low-Involvement (LI)	5.40	1.38	.15**	.12**	.14**	.81**					
Continuance Intention High-Involvement (HI)	3.60	1.93	.42**	-.01	.23**	.41**	.38**				
<b>Desired Benefits</b>											
Hedonic Benefits (HB)	4.69	1.21	.34**	.10*	.34**	.30**	.27**	.31**			
Utilitarian Benefits (UB)	5.17	1.09	.20**	.27**	.44**	.42**	.40**	.24**	.51**		
Symbolic Benefits (SyB)	2.30	1.5	.45**	.01	.14**	.14**	.14**	.35**	.35**	.17**	
Social Benefits (SoB)	3.20	1.51	.50**	-.01	.23**	.19**	.19**	.35**	.47**	.33**	.62**

Note. \*\*Bivariate correlation is significant at the .01 level. Correlation is significant at the .05 level.

## Main Analyses

### General Voice Shopping Intentions (H1 and H2, RQ1)

We tested the predictions that higher values in peer bonding (H1) and market pricing (H2) would predict a stronger general intention to use voice shopping by regressing general voice

shopping intentions on the human-AI relationship dimensions. Supporting H1, the regression analysis showed that higher values in peer bonding were associated with a stronger intention to continue voice shopping in general ( $\beta = 0.18, p = .001, 95\%-CI[0.73,0.29]$ ). H2 was not supported as market pricing did not predict a higher intention to engage in voice shopping ( $\beta = 0.07, p = .255, 95\%-CI[-0.04,0.16]$ ). The same was true for authority ranking, which was included in the regression for exploratory reasons ( $\beta = 0.10, p = .055, 95\%-CI[-0.002,0.19]$ ).

### **Intention to Engage in Low-Involvement Voice Shopping (H3, RQ2)**

We hypothesized that the intention to buy low-involvement products is predicted to a stronger extent by peer bonding than by market pricing. Voice shopping intentions for low-involvement products were regressed on the dimensions of human-AI relationship perception. We found that peer bonding predicts intentions to engage in low-involvement shopping ( $\beta = 0.15, p = .006, 95\%-CI[0.05,0.28]$ ). Market pricing was not associated with low-involvement voice shopping intentions ( $\beta = 0.02, p = .684, 95\%-CI[-0.09,0.13]$ ). Evidence for H3 was provided by the fact that the CIs for both standardized regression coefficients did not include the respective other regression coefficient. Notably, authority ranking positively predicted intentions to voice shop for low-involvement products ( $\beta = 0.15, p = .004, 95\%-CI[0.05,0.25]$ ).

### **Intention to Engage in High-Involvement Voice Shopping (H4, RQ2)**

We hypothesized that the intention to buy high-involvement products is predicted to a stronger extent by market pricing than by peer bonding. Voice shopping intentions for high-involvement products were regressed on the dimensions of relationship perception. We found no significant association of market pricing with intentions to engage in voice shopping for high-involvement products ( $\beta = 0.04, p = .410, 95\%-CI[-0.08,0.20]$ ). However, peer bonding predicted high-involvement shopping intentions ( $\beta = 0.40, p < .001, 95\%-CI[0.43,0.73]$ ). Thus, we did not find evidence for H4. The intention to buy high-involvement products via voice shopping was not predicted by the market pricing but by the perception of peer bonding relationship (Table 5). The reported correlations did not substantially change when shopping spendings or screen use were included as covariates in the regressions reported so far (for details, see supplement).

**TABLE 5 Regression Coefficients of Relational Modes and Shopping Intentions on Desired Benefits (N = 423)**

Variable	General Voice Shopping	Low-Involvement Voice Shopping	High-Involvement Voice Shopping
	$\beta$	$\beta$	$\beta$
Authority Ranking	.10	<b>.15*</b>	.02
Market Pricing	.07	.02	.04
Peer Bonding	<b>.18**</b>	.15*	<b>.40**</b>

Note. \* $p < .05$ . \*\* $p < .01$ . Significant values in bold.

## Relation Between Human-AI Relationships, Desired Benefits, and Voice Shopping Intentions (RQ3)

We regressed the relationship dimensions on the desired benefits (see Table 6). Higher values of desired utilitarian benefits were associated with higher values in authority ranking,  $\beta = 0.31$ ,  $t(418) = 5.53$ ,  $p < .001$ , and market pricing,  $\beta = 0.35$ ,  $t(418) = 6.90$ ,  $p < .001$ . Market pricing was also predicted by desired hedonic benefits,  $\beta = 0.13$ ,  $t(418) = 2.37$ ,  $p = .018$ . Higher values in hedonic benefits,  $\beta = 0.11$ ,  $t(418) = 2.01$ ,  $p = .037$ , desired symbolic,  $\beta = 0.22$ ,  $t(418) = 4.17$ ,  $p < .001$ , and social benefits,  $\beta = 0.31$ ,  $t(418) = 5.37$ ,  $p < .001$ , significantly predicted higher values in peer bonding. The other relations were not significant. Then, we regressed the two voice shopping dimensions on the desired benefits, showing that low-involvement shopping was predicted by desired utilitarian benefits ( $\beta = 0.35$ ,  $t(418) = 6.65$ ,  $p < .001$ ). High-involvement shopping, on the other hand, was significantly associated with desired hedonic ( $\beta = 0.13$ ,  $t(418) = 2.27$ ,  $p = .024$ ), symbolic ( $\beta = 0.21$ ,  $t(418) = 3.64$ ,  $p < .001$ ), and social benefits ( $\beta = 0.14$ ,  $t(418) = 2.21$ ,  $p = .028$ ).

**TABLE 6** Regression Coefficients of Relational Modes and Shopping Intentions on Desired Benefits ( $N = 423$ )

Variable	Authority Ranking	Market Pricing	Peer Bonding	Low-Involvement Voice Shopping	High-Involvement Voice Shopping
Desired Benefits	$\beta$	$\beta$	$\beta$	$\beta$	$\beta$
Utilitarian Benefits	<b>.31**</b>	<b>.35**</b>	.01	<b>.35**</b>	.10
Hedonic Benefits	-.02	<b>.13*</b>	<b>.11*</b>	.06	<b>.13*</b>
Symbolic Benefits	.04	.00	<b>.22**</b>	.05	<b>.21**</b>
Social Benefits	-.12	.06	<b>.31**</b>	.01	<b>.14*</b>

Note. \* $p < .05$ . \*\* $p < .01$ .

In sum, utilitarian benefits are the primary predictor of authority ranking, market pricing, and low-involvement shopping, whereas hedonic, symbolic, and social benefits are related to peer bonding and high-involvement shopping.

## Discussion

The primary goal of this study was to investigate whether voice shopping intentions for low- and high-involvement products depend on how users perceive the human-AI relationships (i.e., peer bonding, market pricing, and authority ranking, based on Tschopp et al., 2023).

Supporting H1, we found that general shopping intentions were predicted by peer bonding, in line with prior research highlighting social dimensions in voice shopping (e.g., Mari & Algesheimer, 2021). Peer bonding showed stronger predictive values for low- and high-involvement shopping than market pricing, supporting H3 but contradicting H4. Peer bonding may not only be relevant for low-involvement shopping but, as indicated by a

strong regression coefficient, even more in high-involvement shopping. This is interesting because it contrasts Rhee and Choi's results (2020) with regard to high-involvement shopping yet supports the findings regarding low-involvement shopping. Against our prediction, market pricing was unrelated to shopping intentions (contradicting H2 and H4). Market pricing may not relate to voice shopping, as the rational calculations inherent in market pricing may not be conducive to the presumably swift decision-making process involved in voice shopping. Thus, one could posit that voice shopping appears to be associated more with rapid decision-making than deliberative, slow thinking (cf. Kahneman, 2012). The difference in results compared to Rhee and Choi (2020) could be due to the different study approaches. They conducted an experiment with undergraduates potentially lacking voice shopping experience and confronted them with a shopping scenario—yielding high internal validity, whereas we recruited experienced voice shoppers and asked about their shopping intentions—yielding high external validity.

Our complementary analysis (RQ3) on the desired benefits sheds light on reasons for the strong predictive power of peer bonding. High-involvement shopping (not low-involvement shopping) was related to perceived hedonic, social, and symbolic benefits, which are more socio-emotional in nature. The importance of the socio-emotional dimensions in all facets of voice shopping supports Dellaert et al.'s (2020) claim that AI assistants are more partners in an interactive decision-making process than subservient assistants. Notably, low-involvement shopping was also related to authority ranking (RQ1 and 2), products traditionally associated with utilitarian purposes, where interaction focuses on efficiency.

In sum, people tend to use voice shopping either in a utilitarian manner, by giving orders to their AI assistant, and/or in a more socio-emotional fashion, immersed in a rather emotional shopping experience. No evidence was found for market pricing we assumed to predict high-involvement shopping, invalidating the concept of low- and high-involvement decision-making. Maybe the technology is simply “not there yet,” or high-involvement products might be bought via voice shopping after the calculative decision process has been performed.

## **Implications for Theory**

The proposed differentiation of perceived human-AI relationships proved to be helpful to disentangle the consequences of different social perceptions on behavioral intentions. Researchers can use the framework to further explore voice shopping or other functionalities (e.g., smart home) and other applications in the broader AI field (e.g., automated driving). Our study focused on voice shopping intentions, yet if our findings also hold for actual behavior, outcomes have strong practical implications.

## **Implications for Practice and Policy**

System designers may have to rethink effective conversational design strategies tailored to different shoppers as well as shopping scenarios. However, more research is needed to draw safe conclusions. Implications may also arise for business developers choosing the sales channel. For selling low-involvement products, Alexa as a channel might work well despite

---

the lack of control over the conversational design. For high-involvement items, control over the social design might be critical due to the found importance of socio-emotional elements. Thus, with limited control over the social design, Alexa as a sales channel for high-involvement products might not work well. Last but not least, the results may also be relevant for policymakers who further aim to investigate the manipulation and addiction potential of human-AI relationships and the potential facilitation thereof through emotional or personalized social designs (Véliz, 2023). In other words, more evidence is needed on whether these relationship dynamics can be exploited.

## Strengths and Limitations

The study enriches the comprehension of the emerging field of voice shopping by investigating experienced voice shoppers and amplifies the value of the perceived human-AI relationships (Tschopp et al., 2023) as predictors thereof. Thereby, this research allows for recommending differentiated voice user interface design strategies and may guide strategic sales channel decisions. A limitation of our findings is the reliance on self-reported shopping intentions instead of actual shopping behavior as well as the lack of cultural variation. Caution is advised regarding the market pricing predictions due to lower scale reliability. The internal consistency was low and could, unfortunately, not be improved by dropping single items. Future research should use longitudinal and/or experimental designs.

## Conclusion

We have investigated the influence of differently perceived human-AI relationships on general, high- and low-involvement shopping intentions. The results emphasized the importance of socio-emotional elements (i.e., peer bonding) for voice shopping, in particular for high-involvement products. For low-involvement products, however, the traditional master-servant relationship (i.e., authority ranking) was still found to be relevant. Understanding the impact of multidimensional human-AI relationship perception is relevant for researchers, system designers, and business developers—presumably not only in voice shopping. Additionally, it holds relevance for policymakers, given recent studies pointed out potential negative impacts like user manipulation or addiction through humanized design (Ramadan, 2021).

## Author Biographies

**Marisa Tschopp** (Dr. des., University of Tübingen) is a corporate researcher at scip AG (Zurich, Switzerland) and associated researcher at the Social Processes Lab at the IWM (Leibniz-Institut für Wissensmedien, Tübingen, Germany). Her research investigates users' perception of human-like characteristics of conversational artificial intelligence (AI) (e.g., the human-AI relationship, trust in AI, and the influence on consumer behavior, such as voice shopping).

 <https://orcid.org/0000-0001-5221-5327>

**Kai Sassenberg** (PhD, University of Göttingen) was head of the Social Processes Lab at the Leibniz-Institut für Wissensmedien and professor at the University of Tübingen (Germany). Since 2023, he is professor at Trier University and Director of the Leibniz Institute for Psychology. His research interests are social influence, conspiracy beliefs, technology acceptance, and metascience.

 <https://orcid.org/0000-0001-6579-8250>

## References

- Alvarez, C., & Fournier, S. (2016). Consumers' relationships with brands. *Current Opinion in Psychology*, *10*, 129–135. <https://doi.org/10.1016/j.copsyc.2015.12.017>
- Carroll, B. A., & Ahuvia, A. C. (2006). Some antecedents and outcomes of brand love. *Marketing Letters*, *17*(2), 79–89. <https://doi.org/10.1007/s11002-006-4219-2>
- Chaudhuri, A., & Holbrook, M. B. (2002). Product-class effects on brand commitment and brand outcomes: The role of brand trust and brand affect. *Journal of Brand Management*, *10*(1), 33–58. <https://doi.org/10.1057/palgrave.bm.2540100>
- de Bellis, E., & Venkataramani Johar, G. (2020). Autonomous shopping systems: Identifying and overcoming barriers to consumer adoption. *Journal of Retailing*, *96*(1), 74–87. <https://doi.org/10.1016/j.jretai.2019.12.004>
- Dellaert, B. G. C., Shu, S. B., Arentze, T. A., Baker, T., Diehl, K., Donkers, B., Fast, N. J., Häubl, G., Johnson, H., Karmarkar, U. R., Oppewal, H., Schmitt, B. H., Schroeder, J., Spiller, S. A., & Steffel, M. (2020). Consumer decisions with artificially intelligent voice assistants. *Marketing Letters*, *31*(4), 335–347. <https://doi.org/10.1007/s11002-020-09537-5>
- Gong, L. (2008). How social is social responses to computers? The function of the degree of anthropomorphism in computer representations. *Computers in Human Behavior*, *24*(4), 1494–1509. <https://doi.org/10.1016/j.chb.2007.05.007>
- Halbauer, I., & Klarmann, M. (2022). How voice retailers can predict customer mood and how they can use that information. *International Journal of Research in Marketing*, *39*(1), 77–95. <https://doi.org/10.1016/j.ijresmar.2021.09.008>
- Han, M. C. (2021). The impact of anthropomorphism on consumers' purchase decision in chatbot commerce. *Journal of Internet Commerce*, *20*(1), 46–65. <https://doi.org/10.1080/15332861.2020.1863022>
- Haslam, N., & Fiske, A. P. (1999). Relational models theory: A confirmatory factor analysis. *Personal Relationships*, *6*(2), 241–250. <https://doi.org/10.1111/j.1475-6811.1999.tb00190.x>
- Hu, P., Lu, Y., & Wang, B. (2022). Experiencing power over AI: The fit effect of perceived power and desire for power on consumers' choice for voice shopping. *Computers in Human Behavior*, *128*(6), 107091. <https://doi.org/10.1016/j.chb.2021.107091>
- Huh, J., Whang, C., & Kim, H.-Y. (2023). Building trust with voice assistants for apparel shopping: The effects of social role and user autonomy. *Journal of Global Fashion Marketing*, *14*(1), 5–19. <https://doi.org/10.1080/20932685.2022.2085603>
- Kahneman, D. (2012). *Thinking, fast and slow*. Penguin psychology. Penguin Books.
- Kinsella, B. (2018). Voice shopping to reach \$40 billion in U.S. and \$5 billion in UK by 2022. <https://voicebot.ai/2018/03/05/voice-shopping-reach-40-billion-u-s-5-billion-uk-2022/>



- Klaus, P., & Zaichkowsky, J. L. (2022). The convenience of shopping via voice AI: Introducing AIDM. *Journal of Retailing and Consumer Services*, 65(3), 102490. <https://doi.org/10.1016/j.jretconser.2021.102490>
- Lim, W. M., Kumar, S., Verma, S., & Chaturvedi, R. (2022). Alexa, what do we know about conversational commerce? Insights from a systematic literature review. *Psychology & Marketing*, 39(6), 1129–1155. <https://doi.org/10.1002/mar.21654>
- Mari, A., & Algesheimer, R. (2021). The role of trusting beliefs in voice assistants during voice shopping. In T. Bui (Ed.), *Proceedings of the Annual Hawaii International Conference on System Sciences*. Hawaii International Conference on System Sciences. <https://doi.org/10.24251/HICSS.2021.495>
- McLean, G., & Osei-Frimpong, K. (2019). Hey Alexa . . . examine the variables influencing the use of artificial intelligent in-home voice assistants. *Computers in Human Behavior*, 99(4), 28–37. <https://doi.org/10.1016/j.chb.2019.05.009>
- Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1), 81–103. <https://doi.org/10.1111/0022-4537.00153>
- Pentina, I., Xie, T., Hancock, T., & Bailey, A. (2023). Consumer–machine relationships in the age of artificial intelligence: Systematic literature review and research directions. *Psychology & Marketing*, 40(8), 1593–1614. <https://doi.org/10.1002/mar.21853>
- Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In R. E. Petty & J. T. Cacioppo (Eds.), *Communication and Persuasion* (pp. 1–24). Springer New York. [https://doi.org/10.1007/978-1-4612-4964-1\\_1](https://doi.org/10.1007/978-1-4612-4964-1_1)
- Pitardi, V., & Marriott, H. R. (2021). Alexa, she’s not human but . . . Unveiling the drivers of consumers’ trust in voice-based artificial intelligence. *Psychology & Marketing*, 38(4), 626–642. <https://doi.org/10.1002/mar.21457>
- Ramadan, Z. B. (2021). “Alexafying” shoppers: The examination of Amazon’s captive relationship strategy. *Journal of Retailing and Consumer Services*, 62(August), 102610. <https://doi.org/10.1016/j.jretconser.2021.102610>
- Rhee, C. E., & Choi, J. (2020). Effects of personalization and social role in voice shopping: An experimental study on product recommendation by a conversational voice agent. *Computers in Human Behavior*, 109(1), 106359. <https://doi.org/10.1016/j.chb.2020.106359>
- Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, 47(5), 609–612. <https://doi.org/10.1016/j.jrp.2013.05.009>
- Sundar, S. S., Jung, E. H., Waddell, T. F., & Kim, K. J. (2017). Cheery companions or serious assistants? Role and demeanor congruity as predictors of robot attraction and use intentions among senior citizens. *International Journal of Human-Computer Studies*, 97, 88–97. <https://doi.org/10.1016/j.ijhcs.2016.08.006>
- Tassiello, V., Tillotson, J. S., & Rome, A. S. (2021). “Alexa, order me a pizza!”: The mediating role of psychological power in the consumer–voice assistant interaction. *Psychology & Marketing*, 38(7), 1069–1080. <https://doi.org/10.1002/mar.21488>
- Tschopp, M., Gieselmann, M., & Sassenberg, K. (2023). Servant by default? How humans perceive their relationship with conversational AI. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 17(3). <https://doi.org/10.5817/CP2023-3-9>
- Véliz, C. (2023). Chatbots shouldn’t use emojis. *Nature*, 615(7952), 375. <https://doi.org/10.1038/d41586-023-00758-y>



# External and Internal Attribution in Human-Agent Interaction: Insights From Neuroscience and Virtual Reality

Nina Lauharatanahirun<sup>1,2</sup> , Andrea Stevenson Won<sup>3</sup> , and Angel Hsing-Chi Hwang<sup>4</sup> 

1 Department of Biomedical Engineering, Pennsylvania State University, University Park, Pennsylvania, USA


2 Department of Biobehavioral Health, Pennsylvania State University, University Park, Pennsylvania, USA

3 Department of Communication, Cornell University, Ithaca, New York, USA

4 Ann S. Bowers College of Computing and Information Science, Cornell University, Ithaca, New York, USA

## Abstract

Agents are designed in the image of humans, both internally and externally. The internal systems of agents imitate the human brain, both at the levels of hardware (i.e., neuromorphic computing) and software (i.e., neural networks). Furthermore, the external appearance and behaviors of agents are designed by people and based on human data. Sometimes, these humanlike qualities of agents are purposely selected to increase their social influence over human users, and sometimes the human factors that influence perceptions of agents are hidden. Inspired by Blascovich's "threshold of social influence" (Blascovich et al., 2002), a model designed to explain the effects of different methods of anthropomorphizing embodied agents in virtual environments, we propose a novel framework for understanding how humans' attributions of human qualities to agents affects their social influence in human-agent interaction. The External and Internal Attributions model of social influence (EIA) builds on previous work on agent-avatars in immersive virtual reality and provides a framework to link previous social science theories to neuroscience. EIA connects external and internal attributions of agents to two brain networks related to social influence: the external perception system, and the mentalizing system. Focusing human-agent interaction research along each of the attributional

**CONTACT** Nina Lauharatanahirun  • [nina.lauhara@psu.edu](mailto:nina.lauhara@psu.edu) • Pennsylvania State University • 531 Chemical and Biomedical Engineering Building • University Park, PA 16802

ISSN 2638-602X (print)/ISSN 2638-6038 (online)  
[www.hmcjournal.com](http://www.hmcjournal.com)



Copyright 2024 Authors. Published under a Creative Commons Attribution 4.0 International (CC BY-NC-ND 4.0) license.

dimensions of the EIA model, or at the functional integration of the two, may lead to a better understanding of the thresholds of social influence necessary for optimal human-agent interaction.

**Keywords:** human-AI, human-agent, neuroscience, virtual reality, social influence

## Introduction

Communicating and interacting with nonhuman agents is becoming increasingly prevalent. In this paper, we define agents as computer programs designed to take actions and/or have specific goals. Such agents can range from virtual assistant technologies to fully autonomous robots. While the technological capability and sophistication of artificially intelligent systems continues to advance, our understanding of how humans process interactions with artificial agents is incomplete. Recently, it has even been suggested that the field of human-robot interaction is approaching a “social robotics winter,” referencing the mismatch between the promise of social robots and the outcome of failed human-robot interactions (Henschel et al., 2020). One source of this mismatch between unrealistic human expectations and social robotics reality comes from attempts to leverage human social reflexes to enhance trust and liking toward agents. However, such interactions can be problematic. Unrealistic expectations and incorrect grounding of human-agent interactions may set humans up for unsuccessful, disappointing, or disingenuous interactions with agents. In such cases, people may be reluctant to adopt or interact with these agents in the future. Thus, it becomes paramount to understand human expectations and perceptions of agent systems with the goal of managing such beliefs in pursuit of more authentic and realistic interactions with these technologies.

We integrate selected research from human-machine communication, human-computer interaction, human-robot interaction, psychology, virtual reality, and social cognitive neuroscience to inform a conceptual framework of humans’ perceptions of agents. We propose a novel adaptation of a key model for human-agent interactions in virtual reality—Blascovich’s Model of Social Influence (Blascovich et al., 2002). We build on this model to define two dimensions of agent characteristics as perceived by humans. Our proposed dimensions are (1) *external* attributions: the tendency to ascribe *anthropomorphic embodiment*, humanlike appearance and/or behavior, to nonhuman agents; and (2) *internal* attributions: the tendency to ascribe agentic humanlike internal states (e.g., mental states, motivations, intentions, and autonomy) to nonhuman agents. We explain how these two dimensions map onto two dissociable neural processing systems—the external perception system and the mentalizing system—that serve as the basis for social cognition and behavior. Finally, we review relevant human-agent and human-computer interaction theories and empirical support for these dimensions. Our aim is for this integrated framework to provide a useful scaffold for research on *understanding* and *predicting* human perceptions of agents, with the broader goal of facilitating transparent and authentic human-agent interactions.

Below, we will first discuss Blascovich’s Model of Social Influence by agent-avatars in virtual reality (Blascovich et al., 2002). In a selective review of the neuroscience literature,

we will relate human-computer interaction theory broadly and Blascovich's model specifically to these two pathways through which our brains process social information. We aim to contribute a better understanding of the neural basis of these social perceptions. We hope that by using neuroscience as a basis for understanding the pathways by which human users become socially influenced by nonhuman agents, will lead to more authentic and more useful social interactions with agents in the future.

## Blascovich's Model of Social Influence

In 2002, Blascovich and colleagues (2002) published a key paper on the experimental potential of agent-avatars (virtual representations that could look and behave like people but were controlled by a computer system; Fox et al., 2015). Specifically, they described how agent-avatars in immersive virtual environments could be provided with human-like appearances and behaviors with the goal of using such agent-avatars for experiments in social psychology. This paper introduced two intersecting dimensions: (1) *behavioral realism* (humanoid appearance and behavior) and (2) *social presence* (whether an entity is believed to be controlled by another person, or by a computer program) as part of a framework that explains under what circumstances such embodied agents (human-appearing social actors controlled by a computer) would be socially influential.

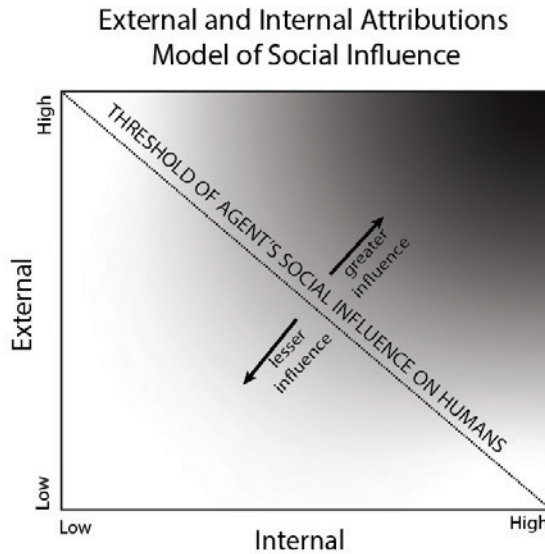
While this framework was proposed as a way to justify immersive virtual reality as a tool for social psychology, the authors made several propositions relevant to artificial agents more broadly considered. First, they proposed that the level of "behavioral realism" exhibited by a virtual human, which included both avatar appearance and behavior (speech, gestures, etc.) could influence human users socially *even if they were aware that the virtual human was an agent* (i.e., controlled by a computer rather than human). Second, they recognized that the influence of human agency was still important—that an agent that a participant believed was controlled by a human (rather than a computer) would be socially influential even if its level of behavioral realism was low. While this model was specific to the field of embodied agents in virtual reality, it can be usefully applied to a much broader context of social agents. The conceptual framework of the intersection between behavioral realism (which we will characterize as *external attributions*) and social presence (which we will expand to *internal attributions*) can be adapted to guide experimental work on identifying the features that create more authentic human-agent interactions. This model allows us to conceptually understand not only group-level effects, but also how individuals may differ in how they experience social influence. Figure 1 shows these relationships, below.

Blascovich et al.'s (2002) model of social influence has been highly influential in work on embodied agents. Considerable work has examined the effects of anthropomorphic external cues, generally finding that greater anthropomorphism leads to greater trust (De Visser et al., 2016) although a meta-analysis from 2007 found that overall effects of anthropomorphism from embodied agents were small (Yee et al., 2007) and more recent analyses have found mixed effects of different aspects of anthropomorphism (for example, appearance versus behavior) on measures of social presence (Oh et al., 2018). The attribution of agency has also commonly been manipulated. A meta-analysis by Fox et al. (2015, p. 1) found support for the importance of internal attributions of agency specifically, identifying an interaction effect such that "studies conducted on a desktop that used objective

measures showed a stronger effect for agency than those that were conducted on a desktop but used subjective measures.” However, a more recent meta-analysis (Felnhofer et al., 2023, p. 1) found that “while deliberate social responses like social presence and evaluation depend on perceived agency, automatic behaviors do not.”

Here, we make an important distinction. In this paper, we are not discussing agent-mediated interactions between humans in which an intelligent agent mediates or otherwise serves as an assistant to human communication (Hancock et al., 2020; Hohenstein & Jung, 2018). Instead, we are examining “stand-alone” agents which present as entities with which individual humans can engage 1:1 (Nass & Moon, 2000; Nass et al., 1994).

**FIGURE 1** An adapted version of the threshold model of social influence in virtual environments, as applied to agents. *External* attribution replaces “behavioral realism” to indicate that the agent is behaving and/or appearing in a human fashion. *Internal* attribution replaces “social presence” to indicate the extent to which the human user attributes internal states, especially intentional agency, to the agent’s actions.



The tendency to anthropomorphize objects and nonhuman agents has been reliably demonstrated in contexts ranging from geometric shapes (Heider & Simmel, 1944) to computer-animated blobs (Guthrie, 1995; Morewedge et al., 2007). Anthropomorphism is broadly defined as the “tendency to imbue the real or imagined behavior of nonhuman agents with humanlike characteristics, motivation, intentions, or emotions” (Epley et al., 2007, p. 864). According to psychological research, humans are motivated to engage in anthropomorphic behavior based on two primary factors (Epley et al., 2007). First, humans are driven to effectively manage uncertainty and need to predict and understand their interaction partners for effective communication and interaction. Second, humans are driven



to form social connections with other humans, a desire that may extend to nonhuman artificial agents. These motivations may lead humans to seek relevant cues in human-agent interaction.

In the context of human-agent interaction, Gambino and colleagues (2020) suggest that people may be consciously assessing the “humanness” of agents and then behaving accordingly. This aligns with work by Sundar (1998), proposing that human users’ information processing styles interact with agent characteristics following Petty and Cacioppo’s Elaboration Likelihood Model (ELM; Petty et al., 1986). The ELM suggests individuals can take either a central or peripheral route for decision-making. If people take the central path, they adopt logical, systematic approaches to processing information; with the peripheral route, people make “fast and frugal” decisions based on heuristic cues. For example, Sundar (1998) posited that highly engaged users would take the central route and evaluate computer-mediated content more systematically, considering the source of online news and the credibility of the author(s) who wrote the news article. On the other hand, casual viewers who take the peripheral route would be more affected by tangential factors, such as visual layout and design of the content. Following on this, Sundar and Nass (2001) proposed that varying the perceived source of presented news (visible, technological, audience, or self) also changed ratings of the story itself. More recently, Sundar and colleagues (2015; 2020) adopted a dual-path framework to conceptualize users’ perception of machine agency. In this work, Sundar uses the Theory of Interactive Media Effects (TIME; Sundar et al., 2015) framework. The TIME model suggests that users evaluate applications of emerging technology either through an action route or a cue route: Through the action route, users determine how to interact with an application based on its actual functions, such as system performance, technical capability, and interacting behaviors demonstrated on an user interface; through the cue route, users evaluate novel applications based on peripheral features (e.g., appearances and content presentation) that are not necessarily related to their technical performance and capabilities per se. Based on the TIME framework, Sundar and Kim propose that the affordances of a given system can lead users to deploy different cognitive heuristics (Sundar & Kim, 2019; Lee, 2018). These include machine heuristics and social heuristics. The former refer to users’ common expectations and even stereotypical impressions for mechanical/computational systems, such as they could perform complex computation tasks accurately and efficiently. By contrast, social heuristics point to humans’ tendencies to treat nonhuman subjects as social entities, such as interacting with them through natural language and verbal communication. This assertion implies that there may be multiple pathways to influence how users make attributions about agents. If users rely on cues and are prompted to use a more social heuristic rather than a “machine heuristic,” for example, through external, anthropomorphic embodiment cues, then this could affect which associated brain networks become active. Alternatively, the “action route” could lead users to actively assess an agent’s source attribution and internal states during interaction, which could also lead to less “mindless” assessments of machine agency. However, while these external and internal factors can be manipulated independently, their effects on attribution are likely intertwined; for example, a person interacting with a very humanlike agent may not be able to avoid attributing internal states to that agent.

## The External-Internal Attribution Model and Neuroscience

Paralleling the dimensions of external and internal attributions in our proposed EIA model, research from social cognitive neuroscience has identified brain networks that are involved in the representation and processing of *external* and *internal* information of others during social interactions. Multiple brain networks are involved in processing external features, such as appearance and movement, of human or nonhuman others. While the social robotics and human neuroscience literatures use slightly different terminology, the networks are analogous. For instance, the action-observation network described in social robotics research (Henschel et al., 2020) is analogous to what is called the mirror neuron network in cognitive neuroscience (Sperduti et al., 2014; Spunt & Lieberman, 2014; Spunt et al., 2015). Similarly, the person-perception brain network (Henschel et al., 2020) from social robotics is equivalent to the face-body perception network (Downing et al., 2001; Kanwisher et al., 1997) from neuroscience. We will refer to brain networks that support the human brain's processing of embodiment cues such as perceptions of movement and appearance as the *external perception system*. Another brain system that is equally important in guiding social influence during social interactions is the *mentalizing system* which is also referred to in the literature as theory of mind. The mentalizing system is involved in processing the *internal* states of another (Alcalá-Lopez et al., 2019; Frith & Frith, 2006; Sperduti et al., 2014; Spunt & Lieberman, 2014; Spunt et al., 2015). Social neuroscience has primarily been focused on understanding the brain systems that support social information processing between humans, but we propose that this line of research may complement existing research in the human-agent interaction field. Below, we operationalize *external* attributions as anthropomorphic embodiment, and *internal* attributions as focusing on intentional agency, where agency refers to an agent's ability to have internal states guiding decision-making and potentially autonomous actions. We integrate both behavioral and neuroscience findings and discuss how our proposed dimensions relate to these brain networks as they are currently understood.

### External Attributions of Anthropomorphic Embodiment

External attribution cues have been much leveraged by designers of human-agent interactions, and these methods of anthropomorphically embodying agents are closely related to Blasovich et al.'s (2002) concept of "behavioral realism," in which an entity's physical form appears and/or behaves like a human being. Embodiment in agents is most clearly illustrated by robots, as the robots necessarily are physically embodied (Breazeal, 2003; Duffy, 2003). However, embodiment can also be a component of "embodied agents"; for example, virtual representations of humans that exist only digitally, such as in virtual or augmented reality applications, or even in AI assistants such as Siri and Alexa which can evoke anthropomorphic embodiment concepts such as gender or age (i.e., the voices used by these devices imply the source of an adult female).

In our framework, we operationalize these external attribution cues as a continuum in which human-like characteristics or cues (e.g., speech, cadence, tone of voice, physical appearance, movement, or other behaviors or features) are applied to nonhuman agents. For example, providing an agent with a female voice, giving it the body of an older adult,

---

or having it raise “eyebrows” as a means of nonverbal expression are all ways to embody nonhuman agents by leveraging human appearance or human behavioral cues. This definition is in line with current research showing that altering artificial agents to appear more human-like in terms of their appearance and behavior can lead to smoother human-agent communication and enhanced engagement (Waytz et al., 2010).

Embodiment features can trigger and enhance anthropomorphism providing more channels for communication (Deng et al., 2019) leading to enhanced human-agent communication and performance (Wainer et al., 2007). Previous research that examined the effect of the physical appearance and behavior of agents on users’ perception and behaviors (von der Pütten et al., 2010; De Visser et al., 2016) supports the effectiveness of anthropomorphic embodiment on evoking social responses in humans. For instance, it is well documented that the fusiform face area/fusiform gyrus (FFA/FFG) responds selectively to faces (Kanwisher et al., 1997) and that the extrastriate body area (EBA) responds selectively to bodies and body parts (Downing et al., 2001), which are key to the fundamental detection and recognition of other people. This recruitment of the FFA represents a fundamental low-level process that is often integrated with higher order cognitive and emotional attributions/appraisals. In the social robotics literature, activation of such brain areas as the FFA/FFG is referred to as the person perception network (PPN; Henschel et al., 2020). Importantly, evidence from brain imaging studies indicates that humans activate the PPN when observing robots express humanlike emotions (Hortensius & Cross, 2018) and when observing other humans interact with robots (Wang & Quadflieg, 2015), although this is moderated by what Blascovich’s model would identify as the factors leading to social influence. Specifically, the right FFA and bilateral posterior superior temporal sulcus showed higher levels of activation in response to human-human interactions relative to human-robot interaction (Wang & Quadflieg, 2015). Moreover, another study found that FFA/FFG activity corresponded with subjective ratings of human likeness ratings, where decreasing activity was observed for artificial agents (Rosenthal-von der Pütten et al., 2019).

Evidence from social robotics research has shown that changing robot appearance (e.g., giving robots faces and human shapes) and robot motor behavior (e.g., hand gestures when communicating) can activate similar brain areas typically recruited during human-human social interactions (Chaminade et al., 2010; Cross et al., 2012), brain regions known as the *mirror neuron network* or the *action-observation network*. While the promise and broad application of the mirror neuron network to higher levels of social cognitive function may have been overstated, its involvement in linking perceptions and actions of others has been replicated in many empirical studies (for reviews see Bonini et al., 2022; Heyes & Catmur, 2022). Perception of agents is an active and automatic process that involves identifying and extracting features of an interaction partner (e.g., speech, appearance, gestures) from the influx of sensory information to help the human observer understand *what* the agent is and *what its function* might be (often indicated through motor movements). This process of identification activates the mirror neuron network in the brain, which includes but is not restricted to the dorsal and ventral regions of the premotor cortex, anterior inferior parietal lobule, anterior temporal cortex, and the temporal parietal junction/superior temporal sulcus (Rizzolatti & Craighero, 2004; Spunt & Lieberman, 2014). When we observe others, this network of brain areas communicates sensory information about another’s motor actions into a representation of a goal-directed action (Iacoboni et al., 2005;

Zacks et al., 2001). For instance, the superior temporal sulcus has been linked to the perception of faces (Haxby et al., 2000; Puce et al., 1998), biological motion (Grossman et al., 2000; Herrington et al., 2011), understanding other's actions (Vander Wyk et al., 2009), and voice perception (Deen et al., 2015). Thus, the human brain synthesizes incoming sensory information regarding the anthropomorphically embodied features of an agent, which in turn can lead to the formation of perceptions that guide our attributional inferences.

Mirror neurons are brain cells distributed across motor, sensory, and motivational brain areas that have been proposed to play a role in social cognition, supporting social interaction (Bonini et al., 2022). Mirror neurons were first discovered in the ventral premotor region F5 of the macaque (di Pellegrino et al., 1992; Gallese et al., 1996; Rizzolatti et al., 1996) and have been identified in a number of species, including humans (Molenberghs et al., 2012; Mukamel et al., 2010). Activation of mirror neurons occurs both when a person performs an action and when a similar action is performed by another individual, thus providing a neural basis for linking perceptions (observations) with motor movements. This key *mirroring* feature of neurons is thought to subserve people's ability to learn new behaviors through imitation and understand the actions of others (for review, see Bonini et al., 2022; Heyes & Catmur, 2022). In nonhuman animal studies, mirror neurons were thought to exist primarily in the ventral premotor cortex and inferior parietal lobule (e.g., di Pellegrino et al., 1992; Rizzolatti et al., 1996); however, human experimental studies have shown that this mirroring feature allowing the mapping of other's actions onto self-related brain regions is not limited to these two brain structures alone. Perhaps one of the most influential studies regarding mirror neurons within the human brain comes from Mukamel and colleagues (2010) who recorded electrophysiological signals from neurons in the medial frontal and temporal cortices while human participants both executed and observed grasping motor movements. The results from their study provide evidence that human neurons in the medial frontal lobe (supplementary motor area), hippocampus, parahippocampal gyrus, and entorhinal cortex fired in response to both performing and observing grasping motor actions. These results not only provide direct evidence of the existence of mirror neurons in the human brain, but indicate that the mirror neuron property exists in brain structures beyond what was previously observed in animal studies.

With regard to social interactions, being able to recognize and perceive the actions of others is key for planning or predicting how we should behave in future situations. While this is a core social cognitive function and the initial starting point for better understanding how humans perceive agents during social interactions, we acknowledge that social interactions are complex and involve the simultaneous processing of multisensory information in response to another's expressions, behaviors, movements, and intentions. In the last decade, social robotics researchers have leveraged neuroimaging technologies to advance our understanding of the neurocognitive mechanisms subserving social behavior during human-robot interactions (for reviews see Cross et al., 2019; Henschel et al., 2020; Hortensius & Cross, 2018). In these studies, the mirror neuron network is referred to as the action-observation network (AON) which includes areas of the parietal, premotor, and middle temporal cortices. Research studies show that the action-observation network is active not only when humans observed other humans, but also when robots grasp and handle objects (Cross et al., 2012; Cross et al., 2019; Henschel et al., 2020). For instance, one study found that AON activation was stronger when human participants were observing

unfamiliar robotic movements (regardless of whether humans or robots performed the action; Cross et al., 2012). This result suggests that humans may engage the mirror neuron network during uncertain social interactions. As previous studies investigating the mirror neuron network suggest (Molenberghs et al., 2012; Mukamel et al., 2010), engagement of this system helps humans learn about their interaction partners. In sum, the action observation/mirror neuron network plays a reflexive and automatic role in understanding the actions of others (humans or artificial agents), and suggests that this system permits the connection of self to other through the simulation of other's actions at the motor level. Together with the face-body/person-perception network, anthropomorphic embodiment cues are processed by an *external perception system* in the brain that ultimately shapes the extent of social influence an agent can have based on whether the human observer's mind determines whether actors exhibit social or nonsocial features.

Neuroscience findings can also help address aspects of embodiment that may be problematic. According to the uncanny valley hypothesis (Mori, 1970; Mori et al., 2012), human perceptions of artificial agents are nonlinear such that likability increases with anthropomorphized agents but precipitously decreases if these agents are perceived to be too humanlike. This has been partially addressed in the neuroscience literature, in that previous work has aimed to uncover the neurocognitive mechanisms associated with human responses to unknown artificial agents. For example, one study identified that nonlinear responses in the ventromedial prefrontal cortex (vmPFC) similarly aligned with the subjective likability and human likeness ratings of artificial agents (Rosenthal-von der Pütten et al., 2019). Responses of the vmPFC scaled with human ratings such that higher ratings of likability and human likeness were associated with greater vmPFC activity, and this association decreased for highly humanlike agents (Rosenthal-von der Pütten et al., 2019). The study also found that amygdala responses predicted when human participants would reject gifts from artificial agents, which is in line with other reports implicating the amygdala's involvement in the processing of social information (Phelps & LeDoux, 2005) such as face processing (Adolphs, 2009) and anthropomorphism (Heberlein & Adolphs, 2004). The role of the amygdala in anthropomorphic perceptions and behavior is not new: Researchers examining patients with basolateral amygdala lesions found that they exhibited decreased anthropomorphic behavior for inanimate stimuli relative to healthy controls (Waytz et al., 2019).

These findings elucidate the neural infrastructure that enables anthropomorphic behavior in guiding humans to process signals and information as social or nonsocial.

## Internal Attributions and Intentional Agency

Early communication theories have suggested that when humans interacted with agents, including text-based interactions with a computer, people were unable to avoid applying human-human social scripts to their interactions (Reeves & Nass, 1996). Conversational agents such as chatbots, virtual agents, and social robots were designed based on the influential "computers-as-social-actors" or CASA theory, which states that humans interact with computers as if they are human (Nass & Moon, 2000; Nass et al., 1994). In these studies, even though human users were consciously aware that computers were not sentient agents, they attributed intentional agency to the devices rather than, for example, to the human programmers of the devices (Nass & Moon, 2000; Nass et al., 1994). However, more recent



work suggests that as people gain experience with computers and incorporate agents into other aspects of their life, they may no longer attribute agency in the same way (Gambino et al., 2020; Heyselaar, 2023).

We operationalize *internal attribution cues* as a continuum in which attributions of humanlike mental states, motivations, intentions, and autonomy are applied to nonhuman agents. One example is the extent to which an artificial agent is perceived to have internal states indicating that it has internal agency; that it is “alive” and “in control” of its own expressions and behaviors. When humans interact with others (humans or artificial agents), we attempt to understand who we are interacting with and will often make attributional inferences about another’s internal states (e.g., beliefs, values) to both explain and predict another’s actions (Frith & Frith, 2006). Even though machines, robots, and artificial agents lack a mind per se, they are programmed with existing policies for actions, movements, and expressions, and thus these internal attributions remain useful and relevant.

The internal attribution dimension in the proposed model maps onto an inferential social cognitive process that involves attributing mental states, intentions, and internal states known as “*mentalizing*” (Frith & Frith, 2006, p. 531). It has been argued that humans and primates alike have evolved to develop larger brain volumes (Dunbar, 1998) as well as specialized brain networks that support social cognition (Adolphs, 2009; Fareri & Delgado, 2014; Kliemann & Adolphs, 2018; Lockwood et al., 2020; Spunt et al., 2015). Being able to engage in social interactions involves a diverse suite of social cognitive abilities that range from low-level sensory processes such as recognizing faces (discussed above as a component of external attributions) to high-level cognitive functions such as making inferences about the intentions of others.

Neuroimaging evidence over the last decade suggests that a network of brain areas is recruited and reliably activated to support higher-level social cognitive processes such as mentalizing. The mentalizing brain network includes key brain regions such as the superior temporal sulcus (STS), temporal parietal junction (rTPJ, lTPJ), posterior cingulate cortex (PCC), and the ventromedial prefrontal cortex (vmPFC). Perhaps one of the most consistently reported brain areas subserving social cognition is the superior temporal sulcus (Deen et al., 2015; Pelphrey et al., 2004; Saxe et al., 2004; Yamada et al., 2022; Zilbovicius et al., 2006). The medial prefrontal cortex has been suggested to play a general role in representing social or emotionally relevant information about oneself (Frith & Frith, 2006; Northoff & Bermpohl, 2004) or another person (Saxe & Powell, 2006). Finally, the brain area that is most notably associated with theory of mind or mentalizing is the temporal parietal junction (TPJ). The TPJ is theorized to play a role in synthesizing lower-level processing streams into higher-order social-cognitive functions. Research has demonstrated that the anterior TPJ is recruited for regulating attentional processes and mentalizing in social situations (Krall et al., 2015; Saxe, 2006; Saxe & Powell, 2006; Van Overwalle, 2009). These neurobiological correlates are important for linking human brain processes with the human mind and, thus, behavior during social interactions.

Recently, likely due to the advancement of technology, researchers have started to examine whether social cognition and mentalizing of humans recruits similar neural circuitry when compared to nonhuman artificial agents. For instance, one study found that social cognitive brain areas such as the TPJ and mPFC selectively responded to humans only relative to humanoid robots (Chaminade et al., 2012). This finding suggests that while

there may be some similarity in how humans perceive appearance and motor features of humans and nonhuman agents, humans still distinguish between intentional agents and entities that may have humanlike internal states (e.g., desires, beliefs) guiding their behavior. Another line of evidence from social neuroscience research has used economic games to understand the neural bases of social interactions (Chang et al., 2023; Fareri et al., 2012; McCabe et al., 2001; Rilling et al., 2004). In these studies, humans engage in social exchange games with humans and computers. Behaviorally, studies have shown that humans entrust resources similarly to humans and agent partners (Schniter et al., 2020). However, future studies examining how the human brain processes these exchanges with agents relative to other people are needed to better understand the neural mechanisms that give rise to social cognition and perception within social interactions. Research in this area would increase our understanding of under what circumstances AI and other nonhuman entities may be perceived as intentional social beings with internal states.

One aspect of internal attribution that has been less explored, at least in quantitative social science, is the fact that most agents are designed and created by an organization (e.g., technology companies) or groups of people and, therefore, their creation cannot be attributed to a single person (Luria, 2020). For example, Apple's Siri voice agent has the modified voice of a human woman, but Siri's design is indebted to hundreds or perhaps thousands of researchers and designers, and the data that built it and refines its output arises from millions of individual human users (Hwang & Won, 2022).

Addressing this gap, one school of researchers proposed that the perceived agency of an agent can be attributed to a single "source" (e.g., Apple), which can then be distributed to various entities through embodiment in different devices (e.g., Siri on your phone, on your tablet, etc.; Luria et al., 2019). This allows a single source of agency (Apple) to be deployed and become omnipresent across different Apple devices, and even re-embodied when a physical artifact is renewed or replaced (e.g., when one gets a new iPhone, and hears Siri's voice coming out of the new speaker). This again suggests that the users can conceive of agency as distinct from embodiment, and hints at a more accurate view of attribution, since most agents are the product of many, many human minds contributing to the overall goals of a business or other entity. This more complex view of the relationship between external attribution (embodiment in a given device) with an internal attribution (a central source of agency such as a company) hints to how users can associate internal states such as intentions with a corporate entity rather than an individual device.

## Perceptions During Social Interactions and the Importance of Social Context

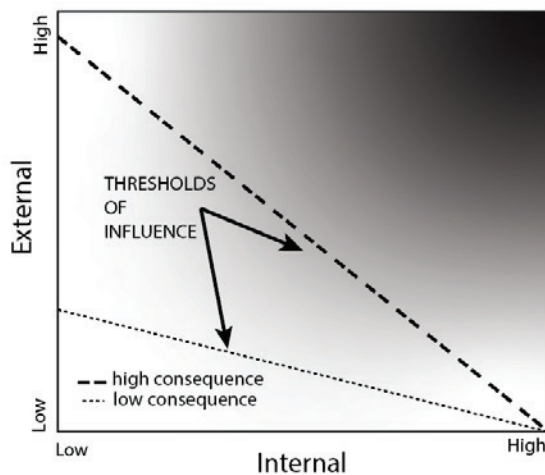
Social interactions are complex. People have to represent their own intentions, beliefs, and values, but also must engage in perspective-taking to understand others' motives, beliefs, and values. Moreover, social interactions require human brains to integrate low-level sensory information that relies on external attributions (e.g., visual, auditory, somatosensory) with higher-level social cognitive processes requiring internal attribution, such as mental state reasoning. Understanding how the human brain integrates both low-level sensory features and higher-level social information for understanding others is not only an interesting area in its own right, but it is also an area ripe for interdisciplinary insights.

---



Blascovich and colleagues (2002) proposed two additional factors that could moderate the threshold of social influence and that are relevant to current communication theories. A reflexive response could be evoked by any agent, but a socially significant situation (for example, taking romantic advice from an agent) would have a higher bar of social influence. In addition, the value or meaning of the interaction to the human user was important. For trivial tasks, Blascovich and colleagues proposed that behavioral realism was *less* likely to be influential, while consequential tasks would retain the higher threshold of social influence. Figure 2 shows these proposed dual thresholds of social influence—a testable proposition that contrasts interestingly with other predictions that different cues will have different weight depending on context and on the importance of the situation to the human interactant.

**FIGURE 2** Is anthropomorphism more important to social influence in high-consequence situations? The original Theory of Social Influence proposed the answer was “yes,” as shown above, but other communication theories predict that low-consequence situations might lead people to rely even more on cues such as anthropomorphism.



Integrating research across internal and external attributions suggests that on the one hand, there may be similarity in how humans perceive appearance and motor features across humans and artificial agents (Chaminade et al., 2012; Frith, 2008; Johnson, 2003; Scholl & Tremoulet, 2000; Thompson et al., 2011). On the other hand, the results also indicate that humans distinguish between intentional agents that have internal states and agents that do not. We argue that, in addition to considering “surface” aspects of human characteristics (i.e., appearance, behavior), designers of artificial agents should also consider how humans perceive the “deeper” social goals and intentions of artificial agents. We believe that studying how people perceive agents within *social contexts* provides an ideal testbed to identify the intersection of external and internal attribution features that reliably recruit

brain systems involved in social cognitive processes (i.e., external perception system and mentalizing system).

Our existing methods for studying human agent interaction are often unidirectional and static to enable controlled testing of experimental manipulations. However, our perceptions are dynamic and continuously updated as we process and integrate incoming information during interactions with agents. Equally important is the fact that these human-agent interactions do not occur within a vacuum. We often interact with agents when other humans are present, and our perceptions may be moderated by how other humans perceive and respond to the agents involved in the interaction. We suggest that we can complement existing behavioral paradigms with neuroimaging and physiological measures to objectively measure how the human brain and mind responds to agents, how humans perform tasks with agents, and how they develop mutual understanding and social engagement over time.

## Next Steps

Further, we ask how understanding the roles that humans play in creating artificial agents might enhance the perception of *intentional agency attributed to humans* who design, build, and provide data to create artificial agents. Such an improved understanding will have at least two potentially useful effects. First, it will make more transparent the influence of the groups of people whose data, opinions, or technical skills inform the creation of AI agents. This will make discussions of bias in AI more intelligible and more salient. For example, many people are still not aware that conversational agents are built using specific datasets that over-represent some humans (people publishing in academic journals, people posting on the programming site Stack Overflow) and under-represent others (people without access to the internet; people who are not literate). While this will not necessarily increase trust in agents, it will allow people to calibrate their trust in these agents based on their real social knowledge of other humans' abilities and biases. We note again that the CASA paradigm described above found that people did not naturally make attributions to, for example, the programmer behind the computer agent. However, we are now living in different times. For instance, a recent replication of the original CASA study found that participants do not treat desktop computers as social actors (Heyselaar, 2023) highlighting the need to conduct new research studies with emergent technologies. Given people's increased experience with agents and the different cultural context in which human-agent interactions occur, it is now time to ask again whether providing more information about the humans and human organizations behind the agents can lead people to make such attributions. Below, we list some research questions that can shed light on whether such conscious reflection on the human element can predict, and improve, the outcomes of human-agent interaction.

## Suggested Research Questions

**RQ1.** When humans are interacting with a group of humans or a group of artificial agents, is intentional agency ascribed to the group as a singular unit? Are similar social cognitive brain networks recruited during interactions with a group of humans versus a group of artificial agents?

**RQ2a.** Does the combination of agency and embodiment mutually enhance activation of the social brain? or:

**RQ2b.** Do the multiple sources of human agency that contribute to artificial agents conflict with anthropomorphic cues, which are necessarily single?

**RQ3a.** Does the *type* of task (consequential and/or social, following Blascovich's proposed moderators of the threshold of social influence) moderate the degree to which mentalization is linked to social influence and/or task success?

**RQ3b.** Does the *type* of task (consequential and/or social, following Blascovich's proposed moderators of the threshold of social influence) moderate the degree to which anthropomorphic cues are linked to social influence and/or task success?

## Conclusion

The modernization and technological advancement occurring within our society necessitates a deeper understanding of how humans perceive agents during human-agent interactions, which may benefit from interdisciplinary perspectives. The broad goal of our proposed framework is to integrate research across disciplines to support the mechanistic understanding of human social cognition during social interactions. Specifically, the intersection of external and internal attributions as described in the EIA model may provide an accessible framework for understanding the social influence agents may have on humans. The framework also provides researchers across disciplines a guide to experimentally test which features activate human social cognitive processing (at the level of the brain or mind) when interacting with artificial agents. It may also help researchers gain insights regarding the conditions under which human perceptions may lead to unrealistic expectations and inaccurate predictions of an agent's actions. Considering social influence as a product of both external and internal attribution cues can also provide a framework for better understanding how neuroscience can be used to enhance our understanding of human-agent interaction and integrate it into more recent work from communication examining AI-mediated communication (Hancock et al., 2020). In turn, we believe this lens can lead to design recommendations for AI that are both more effective and truer to the actual AI ecosystem.

## Author Biographies

**Dr. Nina Lauharatanahirun** (PhD, Virginia Tech) is an Assistant Professor of Biomedical Engineering and Biobehavioral Health at Pennsylvania State University, and the director of the Decision Neuroscience Laboratory. The lab's work is focused on understanding the neurobehavioral mechanisms of social decision-making with the goal of leveraging theoretically grounded neurobehavioral signals for the design of algorithmic solutions that improves human-human and human-agent team decisions.

 <https://orcid.org/0000-0001-8229-1099>

---

**Dr. Andrea Stevenson Won** (PhD, Stanford University) is an Associate Professor of Communication at Cornell University, and the director of the Virtual Embodiment Lab. The lab's work examines tracking and transforming aspects of embodiment, including appearance and behavior, with a focus on virtual reality's clinical, collaborative, and educational capabilities.

 <https://orcid.org/0000-0001-5240-6166>

**Angel Hsing-Chi Hwang** (PhD, Cornell University) is a Post-Doctoral Associate at the Ann S. Bowers College of Computing and Information Science at Cornell University, whose work focuses on researching and designing human-AI interaction at large scales in various applied settings (e.g., Future of Work, mental health care ecosystem, and policy sandbox and prototyping).

 <https://orcid.org/0000-0002-0951-7845>

## References

- Adolphs, R. (2009). The social brain: Neural basis of social knowledge. *Annual Review of Psychology*, 60, 693–716. <https://doi.org/10.1146/annurev.psych.60.110707.163514>
- Alcalá-López, D., Vogeley, K., Binkofski, F., & Bzdok, D. (2019). Building blocks of social cognition: Mirror, mentalize, share?. *Cortex*, 118, 4–18. <https://doi.org/10.1016/j.cortex.2018.05.006>
- Blascovich, J., Loomis, J., Beall, A. C., Swinth, K. R., Hoyt, C. L., & Bailenson, J. N. (2002). Immersive virtual environment technology as a methodological tool for social psychology. *Psychological Inquiry*, 13(2), 103–124. [https://doi.org/10.1207/S15327965PLI1302\\_01](https://doi.org/10.1207/S15327965PLI1302_01)
- Bonini, L., Rotunno, C., Arcuri, E., & Gallese, V. (2022). Mirror neurons 30 years later: Implications and applications. *Trends in cognitive sciences*, 26(9), 767–781.
- Breazeal, C. (2003). Toward sociable robots. *Robotics and Autonomous Systems*, 42(3–4), 167–175. [https://doi.org/10.1016/S0921-8890\(02\)00373-1](https://doi.org/10.1016/S0921-8890(02)00373-1)
- Chaminade, T., Rosset, D., Da Fonseca, D., Nazarian, B., Lutchter, E., Cheng, G., & Deruelle, C. (2012). How do we think machines think? An fMRI study of alleged competition with an artificial intelligence. *Frontiers in Human Neuroscience*, 6, 103. <https://doi.org/10.3389/fnhum.2012.00103>
- Chaminade, T., Zecca, M., Blakemore, S. J., Takanishi, A., Frith, C. D., Micera, S., Dario, P., Rizzolatti, G., Gallese, V., & Umiltà, M. A. (2010). Brain response to a humanoid robot in areas implicated in the perception of human emotional gestures. *PLoS ONE*, 5(7), e11577. <https://doi.org/10.1371/journal.pone.0011577>
- Chang, L. A., Armaos, K., Warns, L., Ma de Sousa, A. Q., Paauwe, F., Scholz, C., & Engelmann, J. B. (2023). Mentalizing in an economic games context is associated with enhanced activation and connectivity in the left temporoparietal junction. *Social Cognitive and Affective Neuroscience*, 18(1), nsad023. <https://doi.org/10.1093/scan/nsad023>
- Cross, E. S., Hortensius, R., & Wykowska, A. (2019). From social brains to social robots: Applying neurocognitive insights to human–robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1771), 20180024. <https://doi.org/10.1098/rstb.2018.0024>

- Cross, E. S., Liepelt, R., de C. Hamilton, A. F., Parkinson, J., Ramsey, R., Stadler, W., & Prinz, W. (2012). Robotic movement preferentially engages the action observation network. *Human Brain Mapping*, 33(9), 2238–2254. <https://doi.org/10.1002/hbm.21361>
- Deen, B., Koldewyn, K., Kanwisher, N., & Saxe, R. (2015). Functional organization of social perception and cognition in the superior temporal sulcus. *Cerebral Cortex*, 25(11), 4596–4609. <https://doi.org/10.1093/cercor/bhv111>
- Deng, E., Mutlu, B., & Mataric, M. J. (2019). Embodiment in socially interactive robots. *Foundations and Trends in Robotics*, 7(4), 251–356. <https://doi.org/10.1561/23000000056>
- De Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, 22(3), 331. <https://doi.org/10.1037/xap0000092>
- di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: A neurophysiological study. *Experimental Brain Research*, 91(1), 176–180. <https://doi.org/10.1007/BF00230027>
- Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science*, 293(5539), 2470–2473. <https://doi.org/10.1126/science.1063414>
- Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, 42(3–4), 177–190. [https://doi.org/10.1016/S0921-8890\(02\)00374-3](https://doi.org/10.1016/S0921-8890(02)00374-3)
- Dunbar, R. I. (1998). The social brain hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews*, 6(5), 178–190. [https://doi.org/10.1002/\(SICI\)1520-6505\(1998\)6:5%3C178::AID-EVAN5%3E3.0.CO;2-8](https://doi.org/10.1002/(SICI)1520-6505(1998)6:5%3C178::AID-EVAN5%3E3.0.CO;2-8)
- Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4), 864–886. <https://doi.org/10.1037/0033-295X.114.4.864>
- Fareri, D. S., Chang, L. J., & Delgado, M. R. (2012). Effects of direct social experience on trust decisions and neural reward circuitry. *Frontiers in Neuroscience*, 6, 148. <https://doi.org/10.3389/fnins.2012.00148>
- Fareri, D. S., & Delgado, M. R. (2014). Social rewards and social networks in the human brain. *The Neuroscientist*, 20(4), 387–402. <https://doi.org/10.1177/1073858414521869>
- Felnhofer, A., Knaust, T., Weiss, L., Goinska, K., Mayer, A., & Kothgassner, O. D. (2023). A virtual character's agency affects social responses in immersive virtual reality: A systematic review and meta-analysis. *International Journal of Human-Computer Interaction*, 1–16. <https://doi.org/10.1080/10447318.2023.2209979>
- Fox, J., Ahn, S. J., Janssen, J. H., Yeykelis, L., Segovia, K. Y., & Bailenson, J. N. (2015). Avatars versus agents: A meta-analysis quantifying the effect of agency on social influence. *Human-Computer Interaction*, 30(5), 401–432. <https://doi.org/10.1080/07370024.2014.921494>
- Frith, C. D. (2008). Social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1499), 2033–2039. <https://doi.org/10.1098/rstb.2008.0005>
- Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron*, 50(4), 531–534. <https://doi.org/10.1016/j.neuron.2006.05.001>
- Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, 119(2), 593–609. <https://doi.org/10.1093/brain/119.2.593>
-

- Gambino, A., Fox, J., & Ratan, R. A. (2020). Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication, 1*, 71–85. <https://doi.org/10.30658/hmc.1.5>
- Grossman, E., Donnelly, M., Price, R., Pickens, D., Morgan, V., Neighbor, G., & Blake, R. (2000). Brain areas involved in perception of biological motion. *Journal of Cognitive Neuroscience, 12*(5), 711–720. <https://doi.org/10.1162/089892900562417>
- Guthrie, S. E. (1995). *Faces in the clouds: A new theory of religion*. Oxford University Press.
- Hancock, J. T., Naaman, M., & Levy, K. (2020). AI-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication, 25*(1), 89–100. <https://doi.org/10.1093/jcmc/zmz022>
- Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences, 4*(6), 223–233. [https://doi.org/10.1016/s1364-6613\(00\)01482-0](https://doi.org/10.1016/s1364-6613(00)01482-0)
- Heberlein, A. S., & Adolphs, R. (2004). Impaired spontaneous anthropomorphizing despite intact perception and social knowledge. *Proceedings of the National Academy of Sciences, 101*(19), 7487–7491. <https://doi.org/10.1073/pnas.0308220101>
- Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology, 57*(2), 243. <https://doi.org/10.2307/1416950>
- Henschel, A., Hortensius, R., & Cross, E. S. (2020). Social cognition in the age of human–robot interaction. *Trends in Neurosciences, 43*(6), 373–384. <https://doi.org/10.1016/j.tins.2020.03.013>
- Herrington, J. D., Nymberg, C., & Schultz, R. T. (2011). Biological motion task performance predicts superior temporal sulcus activity. *Brain and Cognition, 77*(3), 372–381. <https://doi.org/10.1016/j.bandc.2011.09.001>
- Heyes, C., & Catmur, C. (2022). What happened to mirror neurons? *Perspectives on Psychological Science, 17*(1), 153–168.
- Heyselaar, E. (2023). The CASA theory no longer applies to desktop computers. *Scientific Reports, 13*(1), 19693. <https://doi.org/10.1038/s41598-023-46527-9>
- Hohenstein, J., & Jung, M. (2018, April). AI-supported messaging: An investigation of human-human text conversation with AI support. In *Extended abstracts of the 2018 CHI conference on human factors in computing systems* (pp. 1–6). <https://doi.org/10.1145/3170427.3188487>
- Hortensius, R., & Cross, E. S. (2018). From automata to animate beings: The scope and limits of attributing socialness to artificial agents. *Annals of the New York Academy of Sciences, 1426*(1), 93–110. <https://doi.org/10.1111/nyas.13727>
- Hwang, A. H. C., & Won, A. S. (2022, April). AI in your mind: Counterbalancing perceived agency and experience in human-AI interaction. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts* (pp. 1–10). <https://doi.org/10.1145/3491101.3519833>
- Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., & Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLoS Biology, 3*(3), e79. <https://doi.org/10.1371/journal.pbio.0030079>
- Johnson, S. C. (2003). Detecting agents. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 358*(1431), 549–559. <https://doi.org/10.1098/rstb.2002.1237>



- Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17(11), 4302–4311. <https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997>
- Kliemann, D., & Adolphs, R. (2018). The social neuroscience of mentalizing: Challenges and recommendations. *Current Opinion in Psychology*, 24, 1–6. <https://doi.org/10.1016/j.copsyc.2018.02.015>
- Krall, S. C., Rottschy, C., Oberwelling, E., Bzdok, D., Fox, P. T., Eickhoff, S. B., Fink, G. R., & Konrad, K. (2015). The role of the right temporoparietal junction in attention and social interaction as revealed by ALE meta-analysis. *Brain Structure and Function*, 220, 587–604. <https://doi.org/10.1007/s00429-014-0803-z>
- Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1), 205395171875668. <https://doi.org/10.1177/2053951718756684>
- Lockwood, P. L., Apps, M. A., & Chang, S. W. (2020). Is there a ‘social’ brain? Implementations and algorithms. *Trends in Cognitive Sciences*, 24(10), 802–813. <https://doi.org/10.1016/j.tics.2020.06.011>
- Luria, M. (2020). Mine, yours or Amazon’s?: Designing agent ownership and affiliation. *Companion Publication of the 2020 ACM Designing Interactive Systems Conference*, 537–542. <https://doi.org/10.1145/3393914.3395830>
- Luria, M., Reig, S., Tan, X. Z., Steinfeld, A., Forlizzi, J., & Zimmerman, J. (2019, June). Re-embodiment and co-embodiment: Exploration of social presence for robots and conversational agents. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (pp. 633–644). <https://doi.org/10.1145/3322276.3322340>
- McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences*, 98(20), 11832–11835. <https://doi.org/10.1073/pnas.211415698>
- Molenberghs, P., Cunnington, R., & Mattingley, J. B. (2012). Brain regions with mirror properties: A meta-analysis of 125 human fMRI studies. *Neuroscience & Biobehavioral Reviews*, 36(1), 341–349. <https://doi.org/10.1016/j.neubiorev.2011.07.004>
- Morewedge, C. K., Preston, J., & Wegner, D. M. (2007). Timescale bias in the attribution of mind. *Journal of Personality and Social Psychology*, 93(1), 1–11. <https://doi.org/10.1037/0022-3514.93.1.1>
- Mori, M. (1970) The uncanny valley. *Energy*, 7(4), 33–35.
- Mori, M., MacDorman, K., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, 19(2), 98–100. <https://doi.org/10.1109/MRA.2012.2192811>
- Mukamel, R., Ekstrom, A. D., Kaplan, J., Iacoboni, M., & Fried, I. (2010). Single-neuron responses in humans during execution and observation of actions. *Current Biology*, 20(8), 750–756. <https://doi.org/10.1016/j.cub.2010.02.045>
- Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1), 81–103. <https://doi.org/10.1111/0022-4537.00153>
- Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Conference Companion on Human Factors in Computing Systems*, 204. <https://doi.org/10.1145/259963.260288>
- Northoff, G., & Bermpohl, F. (2004). Cortical midline structures and the self. *Trends in Cognitive Sciences*, 8(3), 102–107. <https://doi.org/10.1016/j.tics.2004.01.004>
-

- Oh, C. S., Bailenson, J. N., & Welch, G. F. (2018). A systematic review of social presence: Definition, antecedents, and implications. *Frontiers in Robotics and AI*, 5, 409295.
- Pelphrey, K. A., Morris, J. P., & 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), 1706–1716. <https://doi.org/10.1162/0898929042947900>
- Petty, R. E., Cacioppo, J. T., Petty, R. E., & Cacioppo, J. T. (1986). *The elaboration likelihood model of persuasion* (pp. 1–24). Springer New York. [https://doi.org/10.1016/S0065-2601\(08\)60214-2](https://doi.org/10.1016/S0065-2601(08)60214-2)
- Phelps, E. A., & LeDoux, J. E. (2005). Contributions of the amygdala to emotion processing: From animal models to human behavior. *Neuron*, 48(2), 175–187. <https://doi.org/10.1016/j.neuron.2005.09.025>
- Puce, A., Allison, T., Bentin, S., Gore, J. C., & McCarthy, G. (1998). Temporal cortex activation in humans viewing eye and mouth movements. *Journal of Neuroscience*, 18(6), 2188–2199. <https://doi.org/10.1523/JNEUROSCI.18-06-02188.1998>
- Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. CSLI Publications; Cambridge University Press.
- Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2004). The neural correlates of theory of mind within interpersonal interactions. *Neuroimage*, 22(4), 1694–1703. <https://doi.org/10.1016/j.neuroimage.2004.04.015>
- Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, 27, 169–192. <https://doi.org/10.1146/annurev.neuro.27.070203.144230>
- Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3(2), 131–141. [https://doi.org/10.1016/0926-6410\(95\)00038-0](https://doi.org/10.1016/0926-6410(95)00038-0)
- Rosenthal-von der Pütten, A. M., Krämer, N. C., Maderwald, S., Brand, M., & Grabenhorst, F. (2019). Neural mechanisms for accepting and rejecting artificial social partners in the uncanny valley. *The Journal of Neuroscience*, 39(33), 6555–6570. <https://doi.org/10.1523/JNEUROSCI.2956-18.2019>
- Saxe, R. (2006). Uniquely human social cognition. *Current Opinion in Neurobiology*, 16(2), 235–239. <https://doi.org/10.1016/j.conb.2006.03.001>
- Saxe, R., & Powell, L. J. (2006). It's the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science*, 17(8), 692–699. <https://doi.org/10.1111/j.1467-9280.2006.01768.x>
- Saxe, R., Xiao, D. K., Kovacs, G., Perrett, D. I., & Kanwisher, N. (2004). A region of right posterior superior temporal sulcus responds to observed intentional actions. *Neuropsychologia*, 42(11), 1435–1446. <https://doi.org/10.1016/j.neuropsychologia.2004.04.015>
- Schniter, E., Shields, T. W., & Sznycer, D. (2020). Trust in humans and robots: Economically similar but emotionally different. *Journal of Economic Psychology*, 78, 102253. <https://doi.org/10.1016/j.joep.2020.102253>
- Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, 4(8), 299–309. [https://doi.org/10.1016/S1364-6613\(00\)01506](https://doi.org/10.1016/S1364-6613(00)01506)

- Sperduti, M., Guionnet, S., Fossati, P., & Nadel, J. (2014). Mirror neuron system and mentalizing system connect during online social interaction. *Cognitive Processing*, 15(3), 307–316. <https://doi.org/10.1007/s10339-014-0600-x>
- Spunt, R. P., & Lieberman, M. D. (2014). Automaticity, control, and the social brain. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual-process theories of the social mind* (pp. 279–296). The Guilford Press.
- Spunt, R. P., Meyer, M. L., & Lieberman, M. D. (2015). The default mode of human brain function primes the intentional stance. *Journal of Cognitive Neuroscience*, 27(6), 1116–1124. [https://doi.org/10.1162/jocn\\_a\\_00785](https://doi.org/10.1162/jocn_a_00785)
- Sundar, S. S. (1998). Effect of source attribution on perception of online news stories. *Journalism & Mass Communication Quarterly*, 75(1), 55–68. <https://doi.org/10.1177/107769909807500108>
- Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human–AI interaction (HAI). *Journal of Computer-Mediated Communication*, 25(1), 74–88. <https://doi.org/10.1093/jcmc/zmz026>
- Sundar, S. S., Jia, H., Waddell, T. F., & Huang, Y. (2015). Toward a theory of interactive media effects (TIME): Four models for explaining how interface features affect user psychology. In S. S. Sundar (Ed.), *The handbook of the psychology of communication technology* (1st ed., pp. 47–86). Wiley. <https://doi.org/10.1002/9781118426456.ch3>
- Sundar, S. S., & Kim, J. (2019). Machine heuristic: When we trust computers more than humans with our personal information. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–9. <https://doi.org/10.1145/3290605.3300768>
- Sundar, S. S., & Nass, C. (2001). Conceptualizing sources in online news. *Journal of Communication* 51, 1 (2001), 52–72. <https://doi.org/10.1111/j.1460-2466.2001.tb02872.x>
- Thompson, J. C., Trafton, J. G., & McKnight, P. (2011). The perception of humanness from the movements of synthetic agents. *Perception*, 40(6), 695–704. <https://doi.org/10.1068/p6900>
- Vander Wyk, B. C., Hudac, C. M., Carter, E. J., Sobel, D. M., & Pelphrey, K. A. (2009). Action understanding in the superior temporal sulcus region. *Psychological Science*, 20(6), 771–777. <https://doi.org/10.1111/j.1467-9280.2009.02359.x>
- Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping*, 30(3), 829–858. <https://doi.org/10.1002/hbm.20547>
- von der Pütten, A. M., Krämer, N. C., Gratch, J., & Kang, S. H. (2010). “It doesn’t matter what you are!” explaining social effects of agents and avatars. *Computers in Human Behavior*, 26(6), 1641–1650. <https://doi.org/10.1016/j.chb.2010.06.012>
- Wainer, J., Feil-Seifer, D. J., Shell, D. A., & Mataric, M. J. (2007). Embodiment and human-robot interaction: A task-based perspective. *RO-MAN 2007 The 16th IEEE International Symposium on Robot and Human Interactive Communication*, 872–877. <https://doi.org/10.1109/ROMAN.2007.4415207>
- Wang, Y., & Quadflieg, S. (2015). In our own image? Emotional and neural processing differences when observing human–human vs human–robot interactions. *Social Cognitive and Affective Neuroscience*, 10(11), 1515–1524. <https://doi.org/10.1093/scan/nsv043>
- Waytz, A., Cacioppo, J. T., Hurlmann, R., Castelli, F., Adolphs, R., & Paul, L. K. (2019). Anthropomorphizing without social cues requires the basolateral amygdala. *Journal of cognitive neuroscience*, 31(4), 482–496. [https://doi.org/10.1162/jocn\\_a\\_01365](https://doi.org/10.1162/jocn_a_01365)
-

- 
- Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences*, 14(8), 383–388. <https://doi.org/10.1016/j.tics.2010.05.006>
- Yamada, Y., Sueyoshi, K., Yokoi, Y., Inagawa, T., Hirabayashi, N., Oi, H., Shirama, A., & Sumiyoshi, T. (2022). Transcranial direct current stimulation on the left superior temporal sulcus improves social cognition in schizophrenia: An open-label study. *Frontiers in Psychiatry*, 13. <https://doi.org/10.3389/fpsy.2022.862814>
- Yee, N., Bailenson, J. N., & Rickertsen, K. (2007, April). A meta-analysis of the impact of the inclusion and realism of human-like faces on user experiences in interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1–10). <https://doi.org/10.1145/1240624.1240626>
- Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., Buckner, R. L., & Raichle, M. E. (2001). Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience*, 4(6), 651–655. <https://doi.org/10.1038/88486>
- Zilbovicius, M., Meresse, I., Chabane, N., Brunelle, F., Samson, Y., & Boddaert, N. (2006). Autism, the superior temporal sulcus and social perception. *Trends in Neurosciences*, 29(7), 359–366. <https://doi.org/10.1016/j.tins.2006.06.004>
-



# In Seal We Trust? Investigating the Effect of Certifications on Perceived Trustworthiness of AI Systems

Magdalena Wischnewski<sup>1</sup> , Nicole Krämer<sup>1,2</sup> , Christian Janiesch<sup>3</sup> ,  
Emmanuel Müller<sup>1,4</sup> , Theodor Schnitzler<sup>1</sup> , and Carina Newen<sup>1</sup> 

1 Research Center for Trustworthy Data Science and Security, Dortmund, Germany

2 Social Psychology: Media and Communication, University of Duisburg-Essen, Duisburg, Germany

3 Enterprise Computing, Technical University Dortmund, Dortmund, Germany


4 Data Science and Data Engineering, Technical University Dortmund, Dortmund, Germany

## Abstract

Trust certification through so-called trust seals is a common strategy to help users ascertain the trustworthiness of a system. In this study, we examined trust seals for AI systems from two perspectives: (1) In a pre-registered online study with  $N = 453$  participants, we asked whether trust seals can increase user trust in AI systems, and (2) qualitatively, we investigated what participants expect from such AI seals of trust. Our results indicate mixed support for the use of AI seals. While trust seals generally did not affect the participants' trust, their trust in the AI system increased if they trusted the seal-issuing institution. Moreover, although participants understood verification seals the least, they desired verifications of the AI system the most.

**Keywords:** artificial intelligence, seals of trust, epistemic trust, transparency, formal verification

**Notes:** We have no conflict of interest to report. This work has been supported by the Research Center Trustworthy Data Science and Security (<https://rc-trust.ai>), one of the Research Alliance centers within the <https://uaruhr.de>

**CONTACT** Magdalena Wischnewski  • [magdalena.wischnewski@tu-dortmund.de](mailto:magdalena.wischnewski@tu-dortmund.de) • Research Center Trustworthy Data Science and Security • Joseph-von-Fraunhofer-Straße 25 • 44227 Dortmund, Germany

ISSN 2638-602X (print)/ISSN 2638-6038 (online)  
[www.hmcjournal.com](http://www.hmcjournal.com)



Copyright 2024 Authors. Published under a Creative Commons Attribution 4.0 International (CC BY-NC-ND 4.0) license.



## Introduction

Artificial intelligence (AI) systems are ubiquitous and have become integral to everyday professional and private life. AI systems such as Open AI's ChatGPT or Google's BERT can generate meaningful text (Feuerriegel et al., 2024), other AI systems are components in safety-critical applications such as those that enable autonomous driving (Grigorescu et al., 2020), and even further, AI systems process highly sensitive health information such as echocardiograms (Madani et al., 2018). Simultaneously, these systems and their underlying building blocks, such as deep learning models, have become very complex, aggravating the so-called black box phenomenon. Consequently, knowing when (not) to trust an AI system can be challenging for different stakeholders, from users to decision-makers and even developers. While efforts to develop inherently trustworthy AI systems are much needed, approaches solely focusing on technical aspects are insufficient, as trust results from a system's perceived rather than its actual trustworthiness. Consequently, users sometimes perceive a system inappropriately, placing either too much or too little trust in an AI system.

To help users' trust calibration, different paths can be taken. One popular and well-researched example is explainable AI (XAI), which aims to increase an AI systems' intelligibility by providing explanations for the system's behavior, making internal processes visible, and increasing the overall transparency of the system (Arrieta et al., 2020). Typical methods of XAI are, for example, visual explanations such as heat maps, which highlight areas of input data that were most influential for the system's output, or textual explanations which provide written or oral statements of the explainer. However, XAI is no panacea to cure a lack of trust, and concerns have been raised in terms of users' cognitive biases (Bertrand et al., 2022) and the cognitive burden that explanations pose on users when explanations are not designed with the end-user in mind (Miller, 2019).

In this paper, we aim to counter the shortcomings of XAI and tackle the problem of trust from a different perspective. We empirically explore the effects of AI certifications, so-called *AI seals of trust*. Such seals are credentials which certify that software has been tested and validated to meet specific predefined criteria or standards in various dimensions. Theoretically grounded in works on epistemic trust, trust theory, signaling theory, and persuasion literature, we examined the effects of three different AI seals of trust in a quantitative online experiment. To do so, participants of our study either viewed an AI system with (experimental groups) or without (control group) an AI seal of trust. In addition, in a qualitative part we asked participants in an open-ended format about their preferences for AI certification.

The importance of this work is underlined by initiatives such as the EU AI Act, which suggests certification as a central mechanism to communicate to the public the compliance with industry and legislative requirements. To date, however, empirical studies investigating the effects of such certifications for AI systems are scarce.

## Theoretical Background

### From Trust in AI to Calibrated Trust in AI

To describe and define *trust in AI*, previous work builds on thoughts from various disciplines, such as philosophy, sociology, and psychology that predominantly examine trust as

---

an interpersonal judgment between two or more individuals. Moreover, choosing interpersonal trust as a starting point to examine trust in AI seems sensible as humans, at times, react socially to machines (Nass & Moon, 2000). In fact, the most widely adopted definition of trust in automation originates in Mayer et al.'s (1995) dyadic model of organizational trust, in which trust results from a person's (the trustor) perceptions of another person's (the trustee) ability, benevolence, and integrity. While the direct application of an interpersonal trust conceptualization might be appropriate for certain occasions, this is not always the case (Madhavan & Wiegmann, 2007). Hence, emanating from Mayer et al.'s ability-benevolence-integrity framework, Lee and See (2004) postulate that for a person to trust a machine, the person needs to assess the perceived reliability and functionality of an AI (ability = performance), the intentions with which it was built (benevolence = purpose), and the intelligibility of AI (integrity = process). Beyond these three trust antecedents, Lee and See (2004) define trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (p. 54).

Hence, users' trust must be appropriately calibrated to the system's actual trustworthiness (Lee & See, 2004; Madhavan & Wiegmann, 2007; Parasuraman & Riley, 1997). As described above, users' trust depends on various factors, such as the system's overall performance or the perceived integrity of the system. However, cognitive and social psychology insights suggest that users' perceptions can be distorted, possibly leading users to place too little or too much trust in a system. Such a mismatch of the perceived and actual system trustworthiness can result in either the system's disuse (i.e., resistance to use the system) or the system's misuse (over-reliance on the system). Both disuse and misuse pose serious consequences. In the context of semi-automated driving, for example, ignoring and over-relying on autopilot has led to deadly incidents.<sup>1</sup> Hence, reaching calibrated user trust is essential.

To calibrate user trust, different approaches have been taken. Wischniewski et al. (2023) offer a systematic overview of previous approaches. In their work, the authors surveyed different empirical, human-centered interventions to match perceived and actual system trustworthiness for automated systems accurately. Many of the interventions reviewed aim to increase a system's transparency, assisting the users' trust assessments by making the system more intelligible. While some interventions successfully calibrated the users' trust in a system, in some cases, the intervention also increased the users' workload (Kunze et al., 2019) or led to overtrust (Yeh & Wickens, 2001). In addition, adding, for example, explanations for increasing transparency had adversarial effects, eroding the users' trust, which Kizilcec (2016) explained by arguing that the additional information might have been confusing for users, reducing their understanding instead of increasing transparency.

Even though these transparency interventions have shown mixed effects, there are other reasons to question these approaches. First, many interventions are not developed for end-users but for developers themselves to make the inner workings of AI more transparent (Miller, 2019). However, explanations are likely to be less successful without the end-users in mind. Second, implementing additional measures such as explanations to increase users' trust shifts the responsibility of being trustworthy from the AI system and its developers to the users, who must determine whether the AI system is trustworthy. Third, previous

---

1. See, for example, <https://www.nts.gov/news/press-releases/Pages/NR20200225.aspx> (accessed February 5, 2024).

research has also shown that some users do not want to know how systems, in particular AI systems, work. They would rather stay willfully ignorant because they fear that knowing how a system operates might stop them from using it (Ngo & Krämer, 2022a).

To conclude, while understanding- and transparency-enhancing approaches aiming to increase user trust indeed hold benefits, they also come with many downsides. In the next section, we suggest a different approach to user trust: epistemic trust through AI seals of trust.

## Epistemic Trust in AI and Trust in AI-as-an-Institution

One of the main assumptions of understanding- and transparency-enhancing approaches to increase trust in AI, such as explanations or cues, is that users carefully assess the trustworthiness of AI to know whether they can trust it or not. Implicitly, this assumption often entails that users make rational choices about a system, that is, choices based on accurate perception and inference. However, as shown in the previous section, this assumption does not always hold.

We suggest that an alternative to such understanding-based trust is *epistemic trust*. Individuals show epistemic trust (see also, *trust in testimony*, Coady, 1992), whenever they accept communication or communicated knowledge from others as trustworthy, generalizable, and relevant (Sperber et al., 2010). In other words, when individuals trust what others tell them, they show epistemic trust. One could quickly assume that, as such, epistemic trust is equal to blind trust. However, individuals only assume information to be truthful and relevant when contextual or content cues like source credibility or plausibility evaluations do not indicate otherwise (Gilbert et al., 1993).

In the context of AI systems, showing epistemic trust in the communication of especially experts can ease their trust assessments, as it is easier for them to ask “Whom to believe?” instead of attempting to understand the AI system. Examining epistemic trust in science communication, Bromme and Gierth (2021) argue that, while from a classical logical perspective, to judge the trustworthiness of someone (or something) based on their expertise would be called an *argumentum ad verecundiam* (an argument from authority), a fallacious inference, it is indeed more accessible for individuals to assess the expertise of the scientists than to assess the veracity and scrutiny of the scholarship itself. Hence, establishing epistemic trust in AI systems could help overcome the burden of understanding the system.

Arguments similar to epistemic trust in AI systems also come from within the human-AI interaction community. Knowles and Richards (2021) established the concept of *public trust* in AI. In doing so, they differentiate between trust in a specific, discrete, and identifiable AI from trust in AI as an abstraction, which they call trust in *AI-as-an-institution*. Central, here, is the argument that “individuals do not develop trust in [AI] systems through careful and ongoing assessment of their trustworthiness; instead, one trusts that the system itself has appropriate mechanisms for ensuring trustworthiness” (Knowles & Richards, 2021, p. 264). Knowles and Richards also make clear that the ensuring instances are not the developers of the AI systems but the broader ecosystem that determines the trustworthiness rules

---

developers must follow. In other words, Knowles and Richards suggest that users develop epistemic trust in the ecosystem to ensure the trustworthiness of AI systems.

In their model of public trust, Knowles and Richards (2021) also suggest a four-step process to reach public trust in AI, starting with (1) defining trustworthiness, followed by (2) specifying trustworthiness, (3) enforcing trustworthiness, and (4) reaching trustworthy AI. In their model, the matter of trust calibration is taken over by the ecosystem, ensuring that AI development and outcomes are inherently trustworthy. However, how would an ecosystem communicate the trustworthiness of AI? One answer, included by Knowles and Richards in the fourth step of their model, is by providing certifications which we discuss in the next section.

## AI Seals of Trust: Theoretical and Empirical Considerations

Certifications such as AI seals of trust generally “refer to a process in which a company’s processes and services [here: AI] are evaluated against a predefined set of criteria via an audit by a third party, which formally acknowledges that the standard defined by the criteria is met” (Lansing et al., 2019, p. 4). As such, certifications aim to reduce complexity and uncertainties about systems and make it easy for users to identify what is (not) trustworthy. To that end, certifications have been discussed and introduced in various contexts, such as cybersecurity, web assurances in e-commerce, or cloud services. For the context of AI, the EU AI Act suggests certification as a central mechanism to communicate compliance with industry and legislative requirements to the public (see Article 44 in Chapter 5 “Standards, Conformity Assessment, Certificates, Registration”<sup>2</sup>).

To introduce seals of trust to the field, it is crucial to consider the effectiveness of such measures. Theoretically, arguments supporting seals of trust have previously predominantly been grounded in (1) trust theory, (2) signaling theory, and (3) persuasion literature, in particular, the elaboration likelihood model (ELM).

From the perspective of trust theory, seals of trust communicate to users through trust-assuring arguments that a system can fulfill the specific requirements laid out in the contract between trustor and trustee. In doing so, in trust theory, seals of trust become part of an institutionalized mechanism that ensures trust. In signaling theory, the main focus is on the communication process of one party to the other. Central here is the assumption of an *information asymmetry* wherein one party is less informed (the trustor) than the other (the trustee). Providing information in the form of seals of trust “are signals which are actions that parties take to reveal their true type” (Kirmani & Rao, 2000, p. 66).

In contrast to trust theory and signaling theory, the ELM is more explicit in how seals are perceived. At its core, the ELM describes how individuals process persuasive arguments by following either a peripheral route of processing which requires less cognitive effort, or a central, more effortful route of information processing. Theoretically, seals of trust function as cues that can effortlessly be processed via the peripheral route. However, processing via the central route is also possible when seals of trust induce deeper elaboration (Lowry et al., 2012).

---

2. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52021PC0206>

While all three theoretical approaches assume positive effects of seals of trust, empirically, previous scholarship has been inconclusive. On the one hand, some authors have found no effects. For example, McKnight et al. (2004) found no effects of, what they called, privacy assurance and industry endorsement seals on trust in web business. The authors explain their results, suggesting that participants either did not notice the seal or did not know what it was supposed to signal. Similar results were obtained by Kim et al. (2008), who found no effect of seals on trust but also pointed to a lack of understanding and familiarity with the seal's meaning. On the other hand, in a more recent study, Kim et al. (2016) found that Web Assurance Seal Services (WASS) were effective instruments to increase users' trust and mitigate their concerns about e-commerce platforms. Moreover, results for the positive effects of seals on trust in the context of e-commerce are supported by findings from Mavlanova et al. (2016). In doing so, the authors differentiated between internal (company's certification) and external (third-party certifications) signals. Their results indicate that, although both signals increased trust, only external signals also increased the perceived quality of the seller. Joining results against and in favor of seals of trust, Adam et al. (2020) introduce the trust tipping point. Examining the effectiveness of seals of trust in the context of online websites, the authors found that below a certain trustworthiness threshold, seals effectively increased users' trust. However, with raising trustworthiness, the seals could not increase users' trust further.

Concluding from previous empirical findings, we know that seals of trust can effectively increase trust. However, the effectiveness might be reduced when (a) users do not notice the seals of trust, (b) users do not know the function of the seal of trust, (c) the seal of trust is granted internally, and (d) user trust is already at a high level.

## The Present Study

Based on the theoretical and empirical findings elaborated above, for this study, we assume that:

**H1:** An AI system with an AI seal of trust is perceived as more trustworthy than an AI system without an AI seal of trust.

Moreover, we are also interested in how a seal of trust would affect each trust dimension (performance, process, and purpose). However, empirical differentiations between the three trust dimensions are rare. Hence, we did not formulate a directional hypothesis but instead posed the following research question:

**RQ:** How does a seal affect the three trust dimensions (performance, process, and purpose)?

Going beyond the mere presence (or absence) of a seal, we are also interested in the specific content of such a seal. What exactly should be certified? As it stands, trustworthy AI can refer to various aspects. While we hypothesize that any seal of trust would help to increase the users' trust perceptions (see H1), we also assume differences between different

---

seals (H2), relating to how familiar users are with the seals' content (H3a) and how well users understand what the seal certifies (H3b). More formally stated, we hypothesize:

**H2:** The three trust seals differ in their perceived trustworthiness, with certification of training data receiving the highest trust, followed by certification of transparency and certification through formal verification.

**H3a:** The seals' perceived trustworthiness partly depends on the perceived familiarity with the seals' content. The more familiar users are with the content of the seal, the higher the perceived trustworthiness of the seal.

**H3b:** The seals' perceived trustworthiness partly depends on the perceived understanding of users of the seals' content. The more intelligible seals are for users, the higher the perceived trustworthiness of the seal.

In addition, as the literature reviewed above suggests, trust in the certifying body will also affect how a seal is perceived. Hence, we assume:

**H4:** The seals' perceived trustworthiness partly depends on the perceived trustworthiness of the certifying body. The higher the perceived trustworthiness of the certifying body, the higher the perceived trustworthiness of the seal.

Because the literature on the possible effects of AI seals of trust is scarce, we also included a more explorative approach to better understand users' needs and expectations. Hence, in addition to the directional hypotheses, we included a qualitative part in which we asked participants to elaborate on which aspects of AI systems should be certified through an AI seal of trust.

## Method

The study received ethical approval from the ethics committee of the University of Duisburg-Essen. All hypotheses and analyses were pre-registered via [OSF—Open Science Framework](#).

### Sample and Study Design

To test our hypotheses and research question, we conducted an online study with a between-group design. To that end, we collected data from  $N = 453$  participants who were randomly assigned to one of four conditions. The sample consisted of 220 females, 218 males, 12 nonbinary, and three participants who preferred not to disclose their gender identity. All participants were recruited via the crowd-sourcing platform Prolific. Participants' mean age was 37.94 ( $SD = 12.69$ ) and ranged from 18 to 80 years. The highest degree for two participants was a middle school degree, for 184 a high school degree, for 194 a Bachelor's degree, for 48 a Master's degree, for four a PhD, and 21 indicated to have received another degree.



## Manipulated Variable: The AI Seal of Trust

The four experimental conditions reflected the different trust seals, in addition to a control group. To that end, we selected three certifications which correspond to archetypical levels of insight into the inner workings of AI systems: (1) The quality of the training data ( $n = 114$ )—that is, even if the AI system is a black box, certifications based on the input (i.e., training data) may assist in assessing the system's trustworthiness, (2) the transparency (e.g., explainability) of the AI system ( $n = 114$ )—as it relates the input and output of a black box approximate system behavior, and (3) the formal verification of a AI system ( $n = 113$ )—as it guarantees desirable behavior of the system by white-boxing it. In addition to these different certifications, we included one control group ( $n = 113$ ), which did not receive any seal of trust.

In addition to a brief description about the respective trust seal (all detailed descriptions can be found in the online supplementary material C), participants saw an image of a seal (see Figure 1). Because the design of a seal likely affects the end-users' trustworthiness perceptions, we reduced this effect by adding the following statement to the visual representation of the seal: "Please be aware that due to copyright reasons, we cannot represent the actual seal. The representation you see here is just a placeholder for this study."

**FIGURE 1 Visualization of the AI Trust Seal That Participants Saw in the Study**



## Procedure

After agreeing to the informed consent, participants were introduced to a working definition of AI (see the online supplementary material A for details). We included this information to ensure that all participants understood the terminology similarly. Afterward, participants of the experimental groups were introduced to the concept of AI seals of trust with the following text:

“Artificial intelligence (AI) is recognized as a strategically important technology that can contribute to a wide array of societal and economic benefits. However, it is also a technology that may present serious risks, challenges, and unintended consequences. Within this context, trust in AI systems is necessary for the broader use of these technologies in society. It is, therefore vital that AI-enabled products and services are developed and implemented responsibly, safely, and ethically. But how to know whether one can trust AI? One way to make this trust judgment easier for users are so-called AI seals of trust. Such AI seals of trust

are granted by independent and neutral intermediaries who assess whether AI fulfills trustworthiness standards. Similar to food certifications and labels, these AI seals signal to users the state of an AI.”

Next, participants saw the different seals of trust and were introduced to different AI systems certified with AI seals of trust. Participants of the control group were directly introduced to the AI system and did not view information on the seals of trust. After viewing the AI systems, participants were asked to answer several questions about one of these AI systems. Before closing the study with a manipulation check and the debriefing, participants were informed about all three possible seals of trust, after which, in an open question, participants were asked to indicate which of the three seals they found most important (ranking question), and what they expect from an AI seal of trust.

## Stimulus Material

Participants read short descriptions of four different AI systems and their functionalities. While modeled after real-world applications to avoid prior exposure effects, all systems were hypothetical and did not exist. The systems were: (1) CheckMySkin, a mobile application to check for skin cancer, (2) Drive Tek, an autonomous driving system, (3) Sound Shuffle, a music recommendation system, and (4) FindYou, a hiring system. The texts participants read can be found in the online supplementary material B.

To increase the generalizability of our results, half of the participants answered questions about the system CheckMySkin, whereas the other half answered questions about the system Drive Tek. Participants in the experimental groups saw both of these systems alongside an AI seal of trust. For the analysis, both conditions were joined.

Moreover, to increase external validity, we added two additional systems, Sound Shuffle and FindYou, which were always presented without an accompanying seal of trust. Hence, all participants of the experimental groups saw two systems with and two systems without seals of trust, whereas participants of the control group only saw systems without seals of trust.

## Measured Variables

All of the following measures were assessed on a 5-point Likert scale, ranging from 1 = “strongly disagree” to 5 = “strongly agree.” For subsequent analyses, items of all measures were summarized to a final mean score.

*Trust in a system.* Because we wanted to assess trust as thoroughly as possible, we combined items from different scales to measure the three dimensions of trust (performance, process, and purpose) and mistrust. The final measure included 15 items to measure the perceived performance of a system (Cronbach’s  $\alpha = .96$ ), 13 items to measure the perceived process (Cronbach’s  $\alpha = .90$ ), 10 items to measure the purpose of the system (Cronbach’s  $\alpha = .87$ ), and 12 items to measure mistrust (Cronbach’s  $\alpha = .94$ ). All items used to measure the trust dimensions and a supporting exploratory factor analysis can be found in the online supplementary material F.

*Perceived familiarity and perceived understanding.* We used a three-item measure, adapted from Gefen (2000), to assess the participants' perceived familiarity with a seal's content. The items were "I am familiar with the concept of [ . . . ]," "I have heard about the possibility to make AI systems better by controlling [ . . . ]," and "Media often report about controlling [ . . . ]." Depending on the group participants were allocated to, the blanks were filled by "the training data," "the concept of transparency," or "the concept of formal verification." For the analyses, all items were summarized in one mean score with Cronbach's  $\alpha = .91$ .

The construct perceived understanding was assessed through the following four items, which were developed following Ngo and Krämer (2022b): "I understand what the seal of trust means," "It is clear to me what the seal certifies," "I could explain in my own words what the certification does," and "I am uncertain about the meaning of the seal." For the analyses, all items were summarized in one mean score with Cronbach's  $\alpha = .88$ . Both constructs, perceived familiarity and perceived understanding were not assessed by participants of the control group who did not view a seal of trust.

*Trust in the certifying body.* Trust in the certifying body was assessed through seven items from corporate credibility scale of Newell and Goldsmith (2001). For the analyses, all items were summarized in one mean score with Cronbach's  $\alpha = .95$ .

*Trust in artificial intelligence.* Because we did not want the individual's take on AI to interfere with our results, we also included individuals' attitudes toward AI as a covariate, using the ATAI scale of Sindermann et al. (2021), which includes five items on an 11-point Likert scale such as "I fear artificial intelligence" or "Artificial intelligence will benefit humankind." For the analyses, all items were summarized in one mean score with Cronbach's  $\alpha = .78$ .

## Qualitative Content Analysis

To better understand the participants' needs and expectations toward an AI seal of trust, we included a ranking question and an open-ended question at the end of our online experiment. In the ranking question, having been introduced to all three possible seals of trust, we wanted to know which of the seals of trust participants found most important. To conclude, we asked:

"Lastly, having seen now three possible AI seals of trust, we are curious whether you have your own opinion about what an AI seal of trust could certify. Below you have some space to let us know what you think would be important."

We analyzed all answers following Mayring's (2014) recommendations for qualitative content analysis (see results section for details).

## Results

All data can be accessed via [OSF—Open Science Framework](#).

---

## Manipulation Check

A chi-squared test with the independent grouping variable trust seal and the dependent variable trust seal recall indicated that significantly more participants remembered correctly the seal they saw than those who did not remember correctly ( $\chi^2(12) = 747.05, p < .001$ ). In the control condition, 63.4% of participants remembered correctly ( $n = 71$ ), in the training data condition, 67.5% ( $n = 77$ ), in the transparency condition, 63.15% ( $n = 72$ ), and in the formal verification, 79.6% ( $n = 90$ ).

## Hypotheses Testing

In the central hypothesis of this work (H1), we expected that participants trust an AI system certified with an AI seal of trust more than an AI system without certification. To determine the effect of a seal on the participants' trust, we conducted an ANCOVA with the trust score as the dependent variable and the four leveled factor *AI seal of trust* as the grouping variable. As the covariate, we controlled for participants' general trust in AI. The descriptive results of the variables trust and its subdimensions performance, process, and purpose, as well as mistrust grouped by the factor *AI seal*, can be found in Table 1.

**TABLE 1** Descriptive Results of the Dependent Variable Trust and Its Subdimensions by Experimental Group

		No Seal	Training Data	Transparency	Formal Proof
Trust	<i>M</i>	3.48	3.53	3.50	3.41
	<i>SD</i>	0.71	0.68	0.76	0.69
Performance	<i>M</i>	3.29	3.49	3.39	3.36
	<i>SD</i>	0.87	0.78	0.93	0.88
Process	<i>M</i>	3.22	3.24	3.22	3.08
	<i>SD</i>	0.85	0.82	0.94	0.87
Purpose	<i>M</i>	3.93	3.87	3.88	3.79
	<i>SD</i>	0.79	0.77	0.76	0.76
Mistrust	<i>M</i>	3.34	3.24	3.35	3.44
	<i>SD</i>	1.05	0.99	1.05	0.94

Results of the ANCOVA indicate that there was no significant difference in the participants' trust scores between the different groups,  $F(3,448) = 0.72, p = .54$ . Moreover, we also had to reject H2 for which we expected that the training data seal would receive the most trust, followed by the transparency seal, and the formal verification seal.

While the result for H1 indicates that none of the three different seals of trust affected participants' trust perceptions, it could have been the case that the seal affected only subdimensions of trust. For this possibility, we did not articulate a hypothesis but posed RQ1, asking whether the different seals affected the three subdimensions, performance, process,

and purpose differently. In addition to the three subdimensions, we also included the measure for mistrust in RQ1 (note that mistrust was not included in the RQ in the pre-registration). To assess RQ1, we conducted a MANCOVA with the subdimensions performance, process (integrity & transparency), and purpose, as well as mistrust as outcome variables and the four leveled factor AI seal of trust as the grouping variable. Similar to testing H1, we also controlled for individual levels of trust in AI. Results indicate that the three subdimensions, as well as mistrust, were similarly affected by the trust seals, Pillai's trace = .02,  $F(3,448) = 1.09$ ,  $p = .075$ .

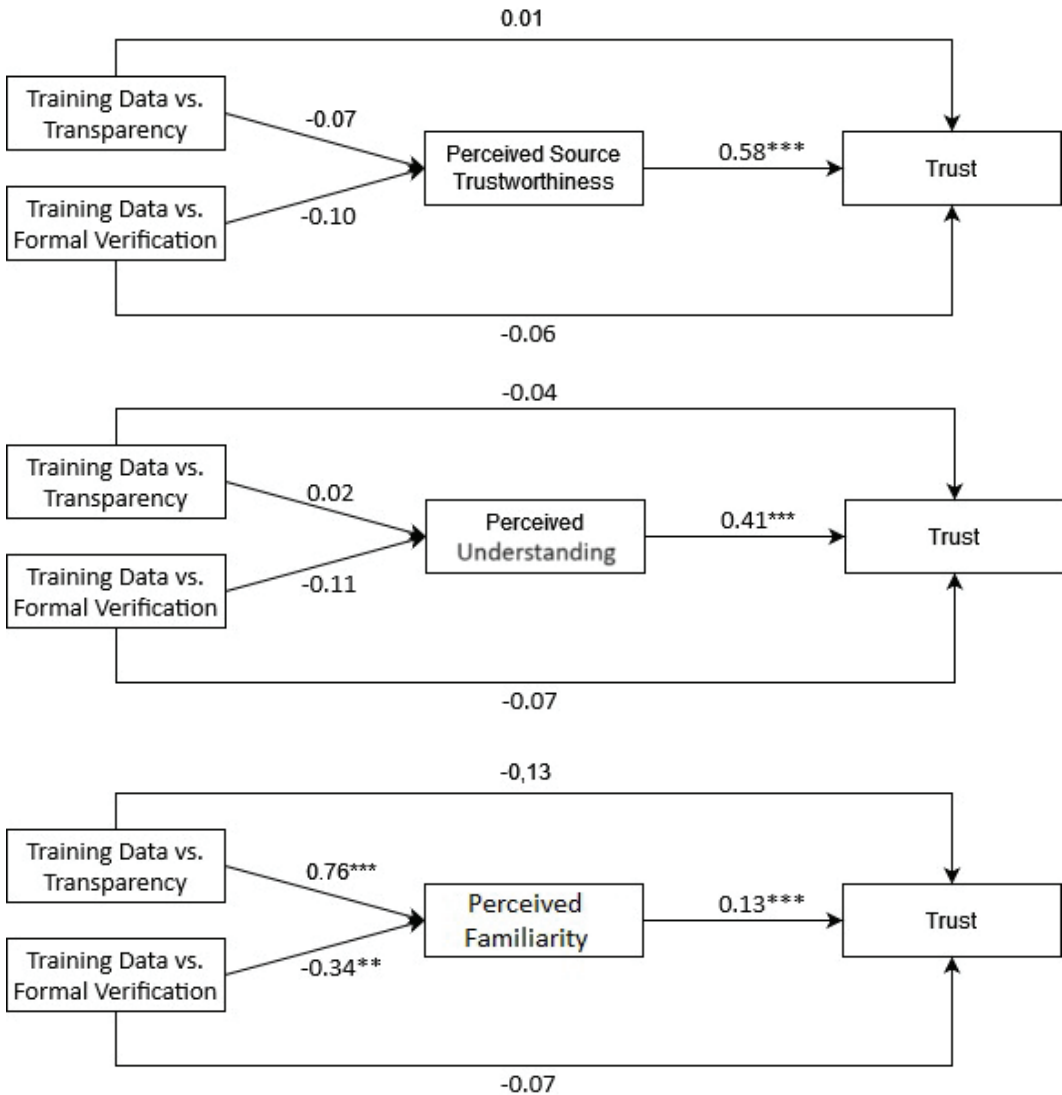
Although we found no differences between the three seals of trust and participants' trust perceptions (see H1), an indirect effect of the seals on trust can still be expected. In H3a and H3b, we suggested that an effect of the seal is at least partly the result of the participants' perceived understanding of the seal's content and the participants' familiarity with the seal's content. In addition, in H4, we anticipate that the effect of the seals might also be the result of the perceived trustworthiness of the institution which issued the seal.

To understand these possible explaining mechanisms, we ran three separate mediation analyses with understanding, perceived familiarity, and perceived source trustworthiness as mediating variables. For this, we used the Process Macro version 4.3.1 for SPSS by Hayes (2017). Furthermore, we used the variable *AI seal of trust* as the independent variable, which was dummy-coded. Participants who viewed the training data seal were entered as a reference category. Participants of the control group were excluded from the analyses as they did not answer questions about their understanding of the seal, their perceived familiarity, and the perceived trustworthiness of the source (see also the elaboration in the methods section). The outcome variable was again trust. We tested the significance of the effects using bootstrapping procedures, computing 5,000 bootstrapped samples with a confidence interval of 95%. All unstandardized path coefficients and significance levels can be found in Figure 2a–c. The full results of the mediation analyses can be found in the online supplementary material D.

The mediation analyses revealed nonsignificant indirect effects for all three variables (understanding, source trustworthiness, and perceived familiarity). For understanding and source trustworthiness, the a-path was insignificant, indicating that the AI seal of trust participants viewed was neither related to the variable understanding nor source trustworthiness. However, the b-path was significant, indicating that both were very strong predictors of trust, with understanding explaining roughly 34% of the trust variance and source trustworthiness explaining roughly 72%. Not surprisingly, these results underline the importance of users understanding what a seal represents and the importance of the issuing source of the seal.

In contrast, we found a significant a-path for perceived familiarity, suggesting that participants were not equally familiar with all AI seals. In particular, we found that participants were more familiar with transparency than verified training data (positive coefficient) but were less familiar with formal verification than training data (negative coefficient). This result partly confirms what we anticipated in H2, suggesting that participants are not equally familiar with the different seal content. Beyond this, the significant b-path indicates that higher familiarity with a seal's content resulted in greater trust.

**FIGURES 2a–2c Visual Representation of Mediation Analyses With Unstandardized Path Coefficients**



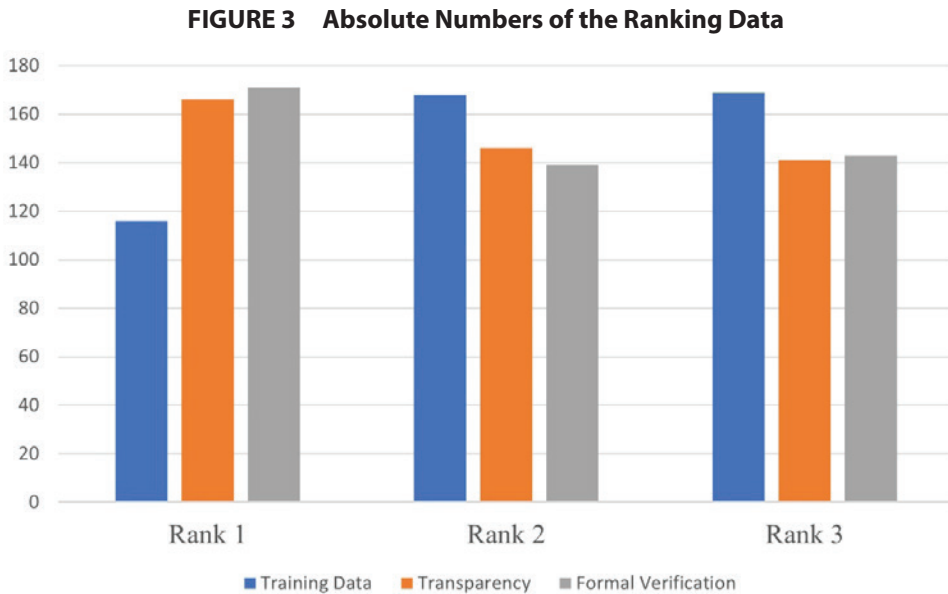
\*  $p < .05$ , \*\*  $p < .01$ , \*\*\*  $p < .001$

### Qualitative Results

First, we asked participants to rank the three seals of trust they found most important. With one as the highest rank and three as the lowest, mean results indicate that participants found all three seals of trust similarly important, with formal verification scoring  $M = 1.94$ , transparency of the AI system  $M = 1.94$ , and training data  $M = 2.12$ . While the mean ranks



do not indicate a great difference between the three seals of trust, which reflects the results of our quantitative analysis, inspecting the absolute number that a seal was ranked first, we can see that participants found the formal verification and transparency of a system most important (see Figure 3).



Because the three trust seals we selected reflect our understanding of importance, we assessed the participants' answers with an open-ended question, asking what participants find most important in an AI seal of trust. We applied descriptive and in-vivo codes in the first coding cycle to capture the participants' answers (Saldaña, 2013). In the second step, all codes were abstracted and summarized into higher-level codes. Throughout both coding cycles, three independent coders worked on the answers. To ensure the quality of the final coding scheme, we calculated Cohen's Kappa on 25% of the answers. In the first round, all three coders arrived at an agreement of  $K = .64$ . To increase agreement, all three coders discussed and resolved cases of disagreement. Consequently, inter-rater reliability increased to a sufficient  $K = .82$  in a second round of coding on different sets of answers.

In the following, we report the most important results of the qualitative content analysis. Overall, the final coding scheme identified seven different categories (see Table 2), which differ in the number of mentions as well as the level of abstraction (number of second-level codes).

**TABLE 2 Results of the Qualitative Content Analysis**

1st level codes	#	Description	2nd level codes (#)
Trustworthiness	19	Verification of the general trustworthiness of a system without further specification	
Distrust	24	General distrust in the AI or the seals of trust	
Performance	206	Verification of AI's abilities and characteristics	(formal) Verification (60) Safety (55) Accuracy (23) Error-free (20) Re-evaluation (20) Extensive testing (20) Efficiency (8)
Process (transparency)	49	Related to how the AI operates and the intelligibility of its inner workings	
Purpose	66	Verification of the intentions of the AI's developers and the development process	Ethical compliance (13) Privacy (24) Training set quality (25) Copyright compliance (4)
Trustworthy AI seals of trust	36	Verification of the seal-issuing institution	Trustworthy origin of the seal (26) Transparency of the certification process (10)
Destructive AI	20	Verification that AI cannot develop its own agency and intentionally harm humans	

While some participants voiced general support for trust seals, others rejected any certification as well as AI systems as a whole. For example, P84 stated, “nothing would really give me any trust in AI. I am very against the idea of anything AI.” In addition to general mistrust in AI and certifications, participants also voiced concrete concerns about the seal-issuing institution. For example, P163 states, “I don’t necessarily trust these seals of trust because they can always get bought.” This reflects our quantitative results, which underline the importance of source trustworthiness. Moreover, the distrust voiced by our participants reminds of what Dietvorst et al. (2015) call algorithm aversion, a generally negative stance toward anything related to algorithms and AI.

Following trust literature, most participants, however, commented along the lines of the three trust dimensions performance, purpose, and process, with performance-related comments being mentioned by far the most. Among those, most participants wanted a seal of trust to certify that the system does what it was set out to be (formal verification) and its safety.

Related to the issue of safety, a smaller group of participants also voiced the need for, what we call, a nondestructive AI seal of trust. For example, P136 stated that a seal “could certify that the AI can be trusted not to be evil and ruin mankind,” and P389 who noted that a seal “could certify whether the AI’s intentions are true—whether it wants to make humans safe or whether it wants to further its own goals regardless of our safety.”

## Discussion

Through quantitative and qualitative data collections, in this work, we investigated the effect of an AI seal of trust on the users’ trust assessments of AI systems as well as the users’ expectations toward such seals respectively.

### Quantitative Results: Addressing the Null Effect of the Trust Seal

In a pre-registered online experiment, we tested three different seals of trust (certification of the training data, transparency, and formal verification) and their effects on user trust in an AI system. However, unlike hypothesized, none of the three different seals of trust could significantly increase our participants’ trust in an AI system compared to a control group. A more fine-grained analysis, differentiating trust into its subdimensions performance, process, and purpose, supported this null result. The seals of trust did not affect the trust dimensions differently compared to a control group.

While previous results from different domains would suggest an effect of the certification, this paper’s null results echo previous null results. Examples include McKnight et al. (2004) and Kim et al. (2016), who relate their null findings to users’ not noticing the seal or users’ limited understanding and familiarity of the seal’s content. We can rule out these explanations because we also assessed participants’ understanding of and familiarity with a seal. In addition, the manipulation check indicated that participants remembered the respective trust seals. Instead, we suggest that our results relate to the findings of Adam et al. (2020). The authors suggest that if a system’s trustworthiness is already high, an additional seal of trust cannot increase the trustworthiness any further. We find support for this speculation in the mean trust ratings of our study as we noticed that these fall within 3.41 and 3.53 points, significantly higher than the scale midpoint (2.5 points).

Following theoretical considerations of trust theory and signaling theory, an alternative explanation to the null results is that the trust seals did not signal the intended meaning. Indeed, our seals might not have communicated the trustworthiness of the systems because they are neither well established outside the experimental setting nor granted by a well-known institution (see also next section). Hence, they possibly lacked the epistemic authority to convince our participants.

Moreover, we found that the seals of trust were not perceived differently in terms of understandability but differed in familiarity, with transparency certification being the most well-known, followed by training data and formal verification certifications. Finding differences for familiarity but not understanding indicates that, while knowing of a specific certification method, this knowledge does not necessarily translate into understanding.

---

## Support for Epistemic Trust

We found that independent of which seal participants saw, the higher the participants' trust in the seal-issuing institution was, the higher was the trust in the AI system. In other words, if users trust the institution that grants the seal, trust in the system will increase. Consequently, this shifts the users' trust assessments from the system to the certifying institution. Hence, our result supports the idea of epistemic trust and trust in AI-as-an-institution (Knowles & Richards, 2021). It seems that it is easier for users to ask, "Whom to trust?" instead of attempting to understand AI systems.

Moreover, in line with predictions of the ELM, knowing a certifying institution might also function as a mental shortcut. Knowing that a certain institution is trustworthy, any communication originating from such an institution should also be trustworthy (see also, *authority heuristic* in Sundar, 2008). For the present work, we could not rely on the authority of a specific institution as our seals might have been less effective because their origin was unknown to the participants. However, adding additional information such as a seal or a seal-issuing institution whose trustworthiness has to be assessed also comes with downsides discussed in the next section.

## Qualitative Results

### ***Need for Verifications Without Understanding of Verifications***

In the qualitative part of this work, we asked participants to explain what they expect from AI certifications. Through a qualitative content analysis, we found that participant responses mainly fell within the three trust dimensions, performance, process, and purpose, with performance-related certification being mentioned the most. Among the performance category, participants indicated that (formal) verification, the certification that the system does what it was set out to be, was mentioned the most. This is also supported by the ranking data that we collected. Here, formal verification was ranked first most of the time. However, in light of the quantitative results, which indicated that participants knew the least about formal verification compared to transparency and training data, the higher ranking of formal verifications is alarming. Participants found the greatest reassurance in something they understood the least and, in turn, maybe expected it to be most comprehensive and fail-safe. We speculate whether this might be due to participants having given up on other, more well-known methods.

## Second-Level Trust Calibrations

Interestingly, some participants mentioned the general need for a trustworthiness certification, whereas others voiced distrust toward any such certification and AI-related system. We relate these contradicting sentiments to what Wischnewski et al. (2023) define as *second-level trust calibrations*, where users have to perform an additional (second level) trust judgment (here: judging the trustworthiness of the seal) on top of the trust judgment concerning the AI system (first level), possibly increasing users' cognitive load. While following persuasion literature which suggests that seals can reduce the users' cognitive load

---

by offering trust cues, future studies should examine whether cognitive load can also be increased through the additional information that needs to be processed. This is especially true in the context of calibrated trust. Suppose it is the aim that user trust is appropriately calibrated to the AI system's functionality. In that case, users must also find a way to calibrate their trust in the AI seal appropriately.

In addition, the distrust sentiment voiced by our participants also indicates the limits of approaching trust from an epistemic perspective. If the seal-issuing institution is not trusted, users will likely not trust the system. Hence, future studies should assess which cues make an AI seal of trust more trustworthy and which user groups generally distrust AI.

## Limitations and Future Studies

The strongest limitation to our study concerns its external validity. First, as currently no established, noncommercial certification body or trust seal exist, all material was hypothetical. Similarly, participants did not directly engage with the AI systems but read different vignettes. Hence, we could not measure how participant trust translated into actual behavioral outcomes. Further, online data collection is limited for decisions in practice as this problem type involves substantial cognitive effort that an online environment may not be able to replicate as well as decision-making often is a high-involvement task and online participants may not meet this criterion.

For future studies, we suggest integrating actual systems into the experimental setting. In addition, with AI systems based on large language models such as GPT-4 being commercialized, it could be interesting, for example, to include such a conversational interface and interactivity in general.

Moreover, as participants likely did not know about AI seals of trust, we had to provide a definition of such. While we tried to be as subtle as possible, describing AI systems as “a technology that may present serious risks, challenges, and unintended consequences” (see Method section), we potentially biased participants to be more critical and vigilant than they initially were, raising participants' overall skepticism toward the presented system. However, as we can see in the overall trust ratings across conditions, participants perceived the systems as relatively trustworthy (mean trust ratings > 3.41 points at a scale midpoint of 2.5 points). In addition, we statistically controlled for participants' general attitudes toward AI by including individuals' attitudes as a covariate in our analyses. Hence, even if a subgroup of users was affected by our definition, it should not have changed our results.

Lastly, as we suggest in the previous section, we speculate that our null results are related to all AI systems being equally trustworthy. To test this interpretation, future studies should experimentally vary the trustworthiness of AI systems by, for example, comparing different levels of system reliability (high vs. low) to investigate whether trust seals can increase the users' trust.

## Conclusion

In this work, we investigated the effects of AI certifications, so-called AI seals of trust, on the users' trust in AI systems. We tested three certifications and their effects on global trust

---

and the trust subdimensions performance, process, and purpose. Unlike hypothesized, we found that the trust seals did not affect users' trust in the AI system. Examining possible underlying mechanisms, we found that a higher understanding of the seal's content as well as familiarity with the seal's content, could increase users' trust. Moreover, we found evidence of epistemic trust. That is, the more participants trusted the seal-issuing institution, the more they trusted the AI system. However, our qualitative results also indicated that some participants reject the idea of an AI seal of trust as they do not trust AI systems or any certifying party. Nevertheless, most participants said they would like to see a system's functionality be certified, specifically, its performance and safety.

## Author Biographies

**Magdalena Wischnewski** (PhD, University of Duisburg-Essen) is a PostDoc at the Research Center for Trustworthy Data Science and Security in Dortmund, Germany. In her work, she investigates human-centric trustworthy AI, such as the calibration of trust, trust assessment, and auditing of AI through AI seals of trust.

 <https://orcid.org/0000-0001-6377-0940>

**Nicole Krämer** (PhD, University of Cologne) is the Professor for Social Psychology: Media and Communication at the University of Duisburg-Essen.

 <https://orcid.org/0000-0001-7535-870X>

**Christian Janiesch** (PhD, University of Münster) is the Professor of Enterprise Computing at TU Dortmund University. His research focuses on intelligent systems at the intersection of business process management and artificial intelligence.

 <https://orcid.org/0000-0002-8050-123X>

**Emmanuel Müller** (PhD, Technical University of Aachen) is the Professor for Computer Science at the Chair of Data Science and Data Engineering at the Technical University Dortmund.

 <https://orcid.org/0000-0002-5409-6875>

**Theodor Schnitzler** (PhD, Ruhr University Bochum) is an Assistant Professor at the Department of Advanced Computing Science at Maastrich University. His main area of research is user privacy in online environments from both technical and HCI perspectives.

 <https://orcid.org/0000-0001-7575-1229>

**Carina Newen** is a PhD student in Computer Science at the Research Center Trustworthy Data Science and Security in Dortmund, Germany. She works in interdisciplinary fields such as trustworthy data science from a computer science and psychological perspective

 <https://orcid.org/0000-0001-8721-6856>

---



## Center for Open Science



This article has earned the Center for Open Science badges for Open Data and Open Materials through Open Practices Disclosure. The authors have made their data and materials freely accessible at <https://osf.io/6znvr/>. The article also earned a Preregistered badge for having a preregistered design available at <https://osf.io/c3g6y>.

## References

- Adam, M., Niehage, L., Lins, S., Benlian, A., & Sunyaev, A. (2020). Stumbling over the trust tipping point—The effectiveness of web seals at different levels of website trustworthiness. In *Proceedings of the 28th European Conference on Information Systems (ECIS). Online Conference, June 15–17, 2020*. [https://aisel.aisnet.org/ecis2020\\_rp/3](https://aisel.aisnet.org/ecis2020_rp/3)
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Bertrand, A., Belloum, R., Eagan, J. R., & Maxwell, W. (2022). How cognitive biases affect XAI-assisted decision-making: A systematic review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 78–91). <https://doi.org/10.1145/3514094.3534164>
- Bromme, R., & Gierth, L. (2021). Rationality and the public understanding of science. In M. Knauff & W. Spohn (Eds.), *The Handbook of Rationality* (pp. 767–776). MIT Press. <https://doi.org/10.7551/mitpress/11252.003.0084>
- Coady, C. A. J. (1992). *Testimony: A philosophical study*. Clarendon Press.
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114. <https://doi.org/10.1037/xge0000033>
- Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering*, 66, 111–126. <https://doi.org/10.1007/s12599-023-00834-7>
- Gefen, D. (2000). E-commerce: The role of familiarity and trust. *Omega*, 28(6), 725–737. [https://doi.org/10.1016/S0305-0483\(00\)00021-9](https://doi.org/10.1016/S0305-0483(00)00021-9)
- Gilbert, D. T., Tafarodi, R. W., & Malone, P. S. (1993). You can't not believe everything you read. *Journal of Personality and Social Psychology*, 65(2), 221. <https://doi.org/10.1037/0022-3514.65.2.221>
- Grigorescu, S., Trasnea, B., Cocias, T., & Macesanu, G. (2020). A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3), 362–386. <https://doi.org/10.1002/rob.21918>

- Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Press.
- Kim, D. J., Ferrin, D. L., & Rao, H. R. (2008). A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents. *Decision Support Systems, 44*(2), 544–564. <https://doi.org/10.1016/j.dss.2007.07.001>
- Kim, D. J., Yim, M.-S., Sugumaran, V., & Rao, H. R. (2016). Web assurance seal services, trust, and consumers' concerns: An investigation of e-commerce transaction intentions across two nations. *European Journal of Information Systems, 25*, 252–273. <https://doi.org/10.1057/ejis.2015.16>
- Kirmani, A., & Rao, A. R. (2000). No pain, no gain: A critical review of the literature on signaling unobservable product quality. *Journal of Marketing, 64*(2), 66–79. <https://doi.org/10.1509/jmkg.64.2.66.1800>
- Kizilcec, R. F. (2016). How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 2390–2395). <https://doi.org/10.1145/2858036.2858402>
- Knowles, B., & Richards, J. T. (2021). The sanction of authority: Promoting public trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 262–271). <https://doi.org/10.1145/3442188.3445890>
- Kunze, A., Summerskill, S. J., Marshall, R., & Filtness, A. J. (2019). Automation transparency: Implications of uncertainty communication for human-automation interaction and interfaces. *Ergonomics, 62*(3), 345–360. <https://doi.org/10.1080/00140139.2018.1547842>
- Lansing, J., Siegfried, N., Sunyaev, A., & Benlian, A. (2019). Strategic signaling through cloud service certifications: Comparing the relative importance of certifications' assurances to companies and consumers. *The Journal of Strategic Information Systems, 28*(4), 101579. <https://doi.org/10.1016/j.jsis.2019.101579>
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors, 46*(1), 50–80. <https://doi.org/10.1518/hfes.46.1.50.30392>
- Lowry, P. B., Moody, G., Vance, A., Jensen, M., Jenkins, J., & Wells, T. (2012). Using an elaboration likelihood approach to better understand the persuasiveness of website privacy assurance cues for online consumers. *Journal of the American Society for Information Science and Technology, 63*(4), 755–776. <https://doi.org/10.1002/asi.21705>
- Madani, A., Arnaout, R., Mofrad, M., & Arnaout, R. (2018). Fast and accurate view classification of echocardiograms using deep learning. *npj Digital Medicine, 1*(1). <https://doi.org/10.1038/s41746-017-0013-1>
- Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human-human and human-automation trust: An integrative review. *Theoretical Issues in Ergonomics Science, 8*(4), 277–301. <https://doi.org/10.1080/14639220500337708>
- Mavlanova, T., Benbunan-Fich, R., & Lang, G. (2016). The role of external and internal signals in e-commerce. *Decision Support Systems, 87*, 59–68. <https://doi.org/10.1016/j.dss.2016.04.009>
- Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734. <https://doi.org/10.5465/amr.1995.9508080335>

- Mayring, P. (2014). Qualitative content analysis: Theoretical foundation, basic procedures and software solution. Klagenfurt. <https://nbn-resolving.org/urn:nbn:de:0168-ssoar-395173>
- McKnight, D. H., Kacmar, C. J., & Choudhury, V. (2004). Shifting factors and the ineffectiveness of third party assurance seals: A two-stage model of initial trust in a web business. *Electronic markets*, 14(3), 252–266. <https://doi.org/10.1080/1019678042000245263>
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267, 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1), 81–103. <https://doi.org/10.1111/0022-4537.00153>
- Newell, S. J., & Goldsmith, R. E. (2001). The development of a scale to measure perceived corporate credibility. *Journal of Business Research*, 52(3), 235–247. [https://doi.org/10.1016/S0148-2963\(99\)00104-6](https://doi.org/10.1016/S0148-2963(99)00104-6)
- Ngo, T., & Krämer, N. (2022a). Exploring folk theories of algorithmic news curation for explainable design. *Behaviour & Information Technology*, 41(15), 3346–3359. <https://doi.org/10.1080/0144929X.2021.1987522>
- Ngo, T., & Krämer, N. (2022b). I humanize, therefore I understand? Effects of explanations and humanization of intelligent systems on perceived and objective user understanding. *psyarXiv preprint*. <https://doi.org/10.31234/osf.io/6az2h>
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230–253. <https://doi.org/10.1518/001872097778543886>
- Saldaña, J. (2013). *The coding manual for qualitative researchers* (2nd ed.). SAGE Publications Inc. <https://doi.org/10.1017/CBO9781107415324.004>
- Sindermann, C., Sha, P., Zhou, M., Wernicke, J., Schmitt, H. S., Li, M., Sariyska, R., Stavrou, M., Becker, B., & Montag, C. (2021). Assessing the attitude towards artificial intelligence: Introduction of a short measure in German, Chinese, and English language. *KI-Künstliche Intelligenz*, 35, 109–118. <https://doi.org/10.1007/s13218-020-00689-0>
- Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origg, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, 25(4), 359–393. <https://doi.org/10.1111/j.1468-0017.2010.01394.x>
- Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. In M. J. Metzger & A. J. Flanagin (Ed.), *Digital Media, Youth, and Credibility* (pp. 73–100). The MIT Press. <https://doi.org/10.1162/dmal.9780262562324.073>
- Wischniewski, M., Krämer, N., & Müller, E. (2023). Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 49–54). <https://doi.org/10.1145/3544548.3581197>
- Yeh, M., & Wickens, C. D. (2001). Display signaling in augmented reality: Effects of cue reliability and image realism on attention allocation and trust calibration. *Human Factors*, 43(3), 355–365. <https://doi.org/10.1518/001872001775898269>
-

# Doctor Who?: Norms, Care, and Autonomy in the Attitudes of Medical Students Toward AI Pre- and Post-ChatGPT

Andrew Prah1<sup>1</sup>  and Kevin Tong Weng Jin<sup>2</sup>

1 Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore


2 Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore

## Abstract

This study adopts the combined TAM-TPB model to investigate attitudes and expectations of machines at a pre-career stage. We study how future doctors (medical students) expect to interact with future AI machinery, what AI usage norms will develop, and beliefs about human and machine autonomy. Semi-structured interviews were conducted. Wave one ( $N = 20$ ) occurred 6 months prior to the public release of ChatGPT; wave two ( $N = 25$ ) occurred in the 6 months following. Three themes emerged: AI is tomorrow, wishing for the AI ouvrier, and human contrasts. Two differences were noted pre-versus post-ChatGPT: (1) participants began to view machinery instead of themselves as the controller of knowledge and (2) participants expressed increased self-confidence if collaborating with a machine. Results and implications for human-machine communication theory are discussed.

**Keywords:** ChatGPT, health care, human-machine communication, generative artificial intelligence, GenAI

**Acknowledgment:** We would like to acknowledge the support of the Nanyang Technological University Undergraduate Research Experience on Campus (URECA) program.

**CONTACT** Andrew Prah1  • [andrew.prah1@ntu.edu.sg](mailto:andrew.prah1@ntu.edu.sg) • Wee Kim Wee School of Communication and Information • Nanyang Technological University • 50 Nanyang Ave • Singapore 639798

ISSN 2638-602X (print)/ISSN 2638-6038 (online)  
[www.hmcjournal.com](http://www.hmcjournal.com)



Copyright 2024 Authors. Published under a Creative Commons Attribution 4.0 International (CC BY-NC-ND 4.0) license.

## Introduction

In the bright bluish light of the theatre, a group of humans form a circle in the middle. In deep concentration, one raises a blade and a voice pierces the silence: “Have you remembered to administer prophylactic antibiotics?” While the surgeon and their team in this operating room theatre are real, the voice is distinctly inhuman. “Yes,” the surgeon replies. This important reminder is not the last contribution to the operation that the surgeon’s robotic assistant will make. After an incision is made, the robot is perfectly positioned to retract the skin, giving the surgeon the best view of the viscera beneath. Along the way, there are constant updates about the patient’s blood pressure, volume of blood loss, heart rate, and oxygen saturation. Throughout, the assistant utilizes artificial intelligence (AI) to comprehend what is happening, adjust as needed, and provide constant reminders the human team needs to ensure that nothing is missed.

Though this scenario might seem futuristic, as if taken from the year 2100, a similar series of events may be seen in operating theatres across the world sooner than we think. And the operating room is just one place where machines are becoming more common in health care. Nearly all levels of care, from the family physician’s office to the most complex lab analyses are being constantly transformed by new technology. However, in contrast to the past’s relatively steady proliferation of medicine technology, the newest machine capabilities—heralded by some as the “4th Industrial Revolution”—are ushering in an age of exponential change (Evans, 2019). Recently, the widely known ChatGPT web application became the fastest growing app ever, outpacing the userbase growth of the former champions (and still tremendously popular) Instagram and TikTok (Hu, 2023; UBS, 2023). Clearly, when they graduate, medical students of today will be entering a world in which machines, especially those using AI, play an outsized role that will continue to grow throughout their careers.

It is imperative that we explore the attitudes of these future doctors toward their future machine companions. Ultimately, the older generations of medical professionals who began their careers without the aid of advanced machine tools such as decision-support systems or autonomous surgical robots will retire from the workforce. The new generation of medical professionals, much like the way younger generations grew up immersed in digital technology, are now emerging into their careers in an era dominated by advanced AI tools. But what will this relationship between doctors and machines be like? What clues can we gain now as to future sources of harmony between human and machine—or, less optimistically—can we identify future sources of tension? An extensive body of research in the human-computer interaction (Gibson et al., 2020), human-automation trust (Hoff & Bashir, 2015), and human-machine communication (Guzman, 2020) disciplines testifies to the importance of pre-existing attitudes and expectations when interacting with new technology. Though most medical students are not explicitly trained in AI, the ubiquity of articles regarding AI in popular media and AI’s overall salience in the public sphere suggests attitudes could be changing rapidly (Bartholomew & Mehta, 2023). We investigate these attitudes and beliefs here; we specifically focus on attitudes toward AI because it is most representative of cutting-edge and future technology.

In our investigation, we contribute to human-machine communication literature by stepping further back into the timeline of professional attitudes. That is, attitudes are formed

---

not only in the workplace during or after machine implementation (where most current research is situated), but also *before* the people enter the workforce. This study asks participants to be more forward-looking than the majority of extant research; in envisioning their careers, medical students must think in terms of decades. To guide our research, we adopt a theoretical framework that is suited to this extended temporal scope: the combined Technology Acceptance Model and Theory of Planned Behaviour (TAM-TPB) (Taylor & Todd, 1995), and we conduct two waves of our study, one preceding the release of ChatGPT and one after the release. Resultingly, our research provides insight into (1) the openness to using AI in their future careers, (2) the perceived usefulness of AI technologies, (3) the norms that medical students believe will form regarding the use of AI in the workplace, (4) the beliefs about personal autonomy in using AI, and (5) how these attitudes may have changed following the release of software that revolutionizes human-machine communication. We first review literature on worker attitudes to AI more broadly, previous work with medical students, and the TAM-TPB framework.

## Literature Review

Artificial Intelligence (AI) is generally understood to refer to the ability of the robots or machines to perform higher order human cognitive functions, to “think” and “act” like humans (IBM, 2020). In recent years, AI programs such as Deep Blue and AlphaGo have come to the forefront of attention by defeating the brightest human minds in games such as Chess and Go. At the same time, some remain wary of machines’ ability to replace humans, adamant that “AI will not replace humans overnight” (Toews, 2021). For our research, we defer to the most widely used definition of AI as “agents that receive percepts from the environment and perform actions” (Russell & Norvig, 2009, p. 10), a definition that allows for the study of a diverse array of machines. Recognizing this broad application of AI, it is critical to understand its impact on diverse domains, especially in the context of human work. Given the issue of human replacement by AI is so salient, much previous work on perceptions of AI in the workplace has focused on this question. Our inquiry, however, takes a different angle to explore the understudied perspective of how future professionals visualize their potential co-existence with AI in the workplace. This shift in focus does not neglect the question of replacement but instead attempts to build upon it by addressing the envisaged human-machine communication in the professional landscape, specifically from the vantage point of professionals entering their field.

## Worker Attitudes

Workers are not always welcoming to AI. Job insecurity and psychological distress are associated with fears about AI taking over jobs for workers in a Philippines call center (Presbitero & Teng-Calleja, 2022). Similarly, concerns arise among North American pharmacists, suggesting apprehension across many job types. These professionals may experience “automation anxieties” (i.e., concerns about job loss) when confronted with the advance of machines into their professions. AI may threaten the identities of employees, especially as it relates to job functions, and may pose a risk to the “social fabric of work” (Selenko et al., 2022, p. 1). This is generally negative not only for employees, but for organizations as well because



job satisfaction (or lack thereof) affects trust between employee and employer (Richter & Näswall, 2019). Media coverage of AI is widespread, and pessimistic views of AI are commonplace (Siegel, 2019; Sun et al., 2020). Even if presented alongside optimistic views regarding AI and labor, the tendency of humans to give more weight to negative outcomes is well-documented (Kahneman & Tversky, 1979). From this perspective, our study acquires an important dimension: it attempts to gauge the pre-existing sentiments of people just venturing into the workforce. We are interested in assessing whether the prevalent AI narrative, often tinged with negativity and uncertainty, has influenced their outlook on working with machines. The public release of ChatGPT in November 2022, given its rapid adoption—setting the record for fastest application ever to reach 1 million users (Hu, 2023; UBS, 2023)—along with its accessibility and widespread media coverage, is an ideal moment to study. Hence, our research design is crafted to capture the sentiment backdrop that the participants might bring to the table.

Leveraging the vantage point of the health care industry benefits our study. Health care is one field in which the application of artificial intelligence is rapidly growing. A number of studies have found varying support for AI in medicine; however, the perception that patient privacy must be protected and that clinicians should be the main participant in decision-making with patients is nearly universal (Scott et al., 2021). Examples cited include the ability of AI to interpret diagnostic imaging, aiding in pharmaceutical development and streamline administrative tasks (Shah & Chircu, 2018). This prioritization of human involvement is not ubiquitous across all industries, with aviation being one notable example where machine autonomy is more embraced. In essence, this distinction underlines that despite technological advancements, the human element is considered central to the practice of health care.

## **Medical Students and Machines**

In recent years, studies have been performed to find out more about medical students' perspectives on Artificial Intelligence (AI) in health care, often with a particular emphasis on the choice of residency or speciality, as well as how it might change the role of doctors in the future. This has been done in countries such as the UK (Sit et al., 2020), the US (Park et al., 2021) and Canada (Gong et al., 2019). Across the board, students surveyed felt that AI will play an important role in health care in the future, and in one study, 89% of students stated that teaching in AI would be beneficial for their future (Sit et al., 2020). Cho et al. (2021) found that although there was interest in AI among medical students, there was a discrepancy between degree of interest and concrete AI education. The primary motivation for many studies of medical students is for education/curriculum planning purposes, and is driven by questions related to how AI should be integrated into medical education. It may not be surprising to see that most of the medical schools which the survey participants came from did not have AI-related content in their syllabi. However, it may be surprising to learn that the vast majority of medical education programs do not include AI in any form. More broadly, other than simple “how to use X machine” education that students receive in experiential settings, medical education is devoid of education about human-machine communication, despite the machines in the workplace being widespread.

---

The curricular-focus of research on medical students has also aimed to address what specialties students believe will be most affected by AI. Two fields purported to be most susceptible to being replaced by AI were radiology and dermatology (Cho et al., 2021; Gong et al., 2019). Many students felt that radiology would be the specialty most affected by AI, with around half of the students surveyed stating that their interest in radiology had been negatively impacted by the development of AI. But the more daunting question of if AI has made the medical profession less attractive in general remains unanswered. Our work will provide insight as attitudes toward AI in professional life are inevitably a combination of both how people feel about AI and how they feel about the *profession* itself. Both research with medical students and in medicine generally have provided little theoretical basis for understanding attitudes toward AI. This is unfortunate as medical students in particular are in an ideal position to inform theory about how workers approach AI. Medical students are at the very cusp of long careers in medicine. An investment in medical school is a serious undertaking that prepares one for a particular industry; if AI takes over a doctor's job, they cannot just jump to a different industry and apply their skills in logistics or accountancy. Like many other areas of skilled labor, the "I'll just do something else" approach may not be viable given the limited transferability of skills. Furthermore, while they are forced to think deeply about their careers, medical students are not yet embedded in real work environments where their attitudes toward AI may be jaded by bad experiences or the dysfunctional work environments that have plagued previous attempts to introduce technology in organizations (Stam et al., 2004). This study provides a novel investigation into the attitudes of medical students as we investigate how the release of a disruptive technology (ChatGPT) may cause shifts in how medical students think about machines.

## Perceptions and Expectations

In investigating perspectives that are relatively unadulterated by work experience, our work follows previous investigations in the human-machine communication tradition (e.g., Guzman, 2020; Simmler et al., 2022). Our sample combines the best advantages of both professional and general population samples. For medical students, AI carries strong professional implications but they are not yet working with AI every day. To maximize the advantages of our sample our investigation is guided by the combined Technology Acceptance Model and Theory of Planned Behaviour (TAM-TPB) (Taylor & Todd, 1995). We chose TAM-TPB because it reflected the qualities of our sample, because it includes items geared toward general populations/issues from the widely used Theory of Planned Behavior (Ajzen, 1991) and because it also includes professional focused items from the similarly popular Technology Acceptance Model (Davis, 1989).

TAM-TPB proposes that attitudes toward technologies are determined from practical factors one would encounter in the workplace such as the perceived usefulness of the technology or the ease of using it. But attitudes also originate from the norms people perceive regarding technology. For instance, medical students' perception of AI in prognosis management would influence their attitudes. Further, as their careers progress, their attitudes toward AI will be shaped by both professional, patient, and societal expectations. Norms are important both on the professional and public sides of health care: providers may have

one set of expectations about how machines should be used, but the public (patients) may have different expectations; these differing expectations must be negotiated. TAM-TPB also proposes that one's *behavioural control* can affect intentions to behave. *Control* has multiple meanings in the workplace. While the straightforward interpretation is related to self-efficacy, control may also manifest in company policies, directives from superiors, or demands from patients that a technology be used or prohibited. Or, machines may demonstrate better performance to humans who resultingly feel inferior and compelled to defer to machines. Behavioral control therefore is related to the struggle for autonomy that is frequently discussed in studies of human and machine (Fortunati & Edwards, 2021; Schaefer et al., 2016). In light of these considerations, adopting the TAM-TPB framework enhances our study as it encompasses both individual and external factors that contribute to one's intention toward new technologies. In sum, the use of the TAM-TPB leverages the strengths of our sample and provides structure for our contribution to human-machine communication theory. As such, we are guided by three research questions:

**RQ1:** How do medical students foresee the usefulness of artificial intelligence in their careers?

**RQ2:** What future norms regarding the use of artificial intelligence in medicine do medical students foresee?

**RQ3:** How do medical students envision personal and professional autonomy in relation to the use of AI in their future workplaces?

## Method

A total of 45 ( $N = 45$ ) medical students participated in semi-structured interviews. The first wave of interviews ( $n = 20$ ) were collected between December 2021 and June 2022. First wave participants were sourced initially through recruitment posters ( $n = 8$ ) and snowball sampling ( $n = 12$ ) by asking participants to recommend others for the study. In the second wave ( $n = 25$ ), collected between January 2023 and June 2023, fewer students ( $n = 5$ ) responded to recruitment posters with the remainder sourced through snowball sampling ( $n = 23$ ). Recruitment in the second wave differed from the first in that there were two research team members independently snowball sampling in order to mitigate the selection bias risks inherent in snowball sampling. The interviewees selected were medical students in their second to fourth year of medical school from two undergraduate medical schools in Singapore. Ages ranged from 19 to 24. Notably, Singapore's medicine education program is compressed compared to the United States where equivalent schooling is often pursued *after* obtaining a bachelor's degree. Thus, our sample is 3–6 years younger than similar studies conducted in like education programs elsewhere.

Interviews were conducted using a mix of face-to-face and video call mediums because some students were under isolation directives resulting from COVID-19 mitigation measures. These directives were loosened between wave one and wave two, meaning more interviews were conducted in-person in wave two (56.5% in-person) than wave one (40.0% in-person). Interviews were conducted by study team members who were fellow medical students in Singapore. Research assistants were earning course credit for participating in

an undergraduate research program; participants were not compensated. The interviews adopted a semi-structured format. A pre-defined list of questions was prepared, but conversations were allowed to flow naturally based on interviewees' responses. The only difference between the wave one and two interview guides is that wave two included allowances for participants to go on tangents related to the recently released (with much fanfare), generative AI tools (e.g., ChatGPT), but we did not change any questions in the interview guide to ask about generative AI specifically. Questions were structured around the combined TAM-TPB model, although posed in layperson language to encourage more natural discussion. Some questions were general conversation openers (e.g., "Do you see AI changing the role of doctors during your career?") whereas others were more targeted at concepts of interest such as fear (e.g., "How do you think you would feel if your boss came to you and told you that you *must* use a new AI tool because studies show it has better judgement than you?"). Interviews lasted between 25–45 minutes.

### Coding and Content Analysis

Interviews were transcribed, analyzed and coded. Our approach follows the reflexive thematic analysis method (Braun & Clarke, 2006; Byrne, 2022) with the coding framework being developed and evolving as more interviews were being transcribed and analyzed. A framework of core themes was built based on the analysis of the initial transcripts using an inductive approach. This framework was built upon and modified as more and more transcripts were coded and analyzed. Earlier interviews were then revisited and re-coded in an iterative process (Braun & Clarke, 2006). In wave one, after completion of the initial coding scheme, study team members met to begin mapping the codes to concepts in the TAM-TPB model when applicable. This process, which we favored over a more structured deductive approach, can result in the identification of codes that fall outside the conceptual framework; we were agreeable to this due to exploratory, futuristic nature of our research questions. But these codes may be problematic from a theoretical view if they are over-fit into the framework. To avoid this conundrum, a crucial step was the consensus-building discussion among our study team. After discussion, the study team agreed that all codes that could not be mapped to a construct(s) in the TAM-TPB were related to student's educational curriculum (which was an inevitable topic of conversation given the interviewer was a peer student). Thus, we place these codes in an education category. All codes and frequencies are shown in Table 1. In wave two, the process differed in that the study team already had the codes developed by the study team members in wave one. The research assistants in wave two were supplied with the codebook, and coded transcripts from wave one. Wave two research team members then coded five transcripts from wave one, and one research team member from wave one re-coded the same transcripts. Intracoder reliability for the wave one and wave two team members was 96% agreement. For reliability between all three wave two members, we elected to calculate Krippendorff's alpha to mitigate problems with percentage agreement (Hayes & Krippendorff, 2007); reliability was acceptable  $\alpha = 0.882$ . The only source of significant disagreement in wave two data was how to cover discussions of generative AI and ChatGPT, which was not accounted for (the technology did not exist yet) in wave one. We place these codes into a "new technology" theme which is listed separate from the themes present in wave one and two (see Table 1).

**TABLE 1 Codes, Frequencies, and Theme Mapping**

Codes	Number	Percentage of All Codes	Percent of Interviews Code Occurred	Theme	Sub-Theme
Administration	50	8.25%	100.00%	AI is Tomorrow	AI will be Demanded
Patient Desires	85	14.03%	100.00%	AI is Tomorrow	AI will be Demanded
Growth of Technology	68	11.22%	100.00%	AI is Tomorrow	AI-Saturated Work
Ubiquity	15	2.48%	33.33%	AI is Tomorrow	AI-Saturated Work
Inevitability	60	9.90%	100.00%	AI is Tomorrow	AI-Saturated Work
Optimism	34	5.61%	75.56%	AI is Tomorrow	Tomorrow is Better
Safety	20	3.30%	44.44%	AI is Tomorrow	Tomorrow is Better
Pessimism	18	2.97%	40.00%	AI is Tomorrow	Tomorrow is Worse
Humans: Skill Comparison	12	1.98%	26.67%	The Human Contrast	Cognition and Intuition
Humans: Positive Comparison	15	2.48%	33.33%	The Human Contrast	Cognition and Intuition
Useful as Assistant	12	1.98%	26.67%	The Human Contrast	Different Colleagues
Coworkers	10	1.65%	22.22%	The Human Contrast	Different Colleagues
Ease of Profession	14	2.31%	31.11%	The AI Ouvrier	A Better Professional
Choice of Use	12	1.98%	26.67%	The AI Ouvrier	A Better Professional
Health Care Industry	18	2.97%	40.00%	The AI Ouvrier	A Better Professional
Ease of Use	7	1.16%	15.56%	The AI Ouvrier	Taking Difficult Work
Ease of Workload	18	2.97%	40.00%	The AI Ouvrier	Taking Difficult Work
Useful as Tool	40	6.60%	88.89%	The AI Ouvrier	The AI Toolbox
Humans: Negative Comparison	39	6.44%	86.67%	The AI Ouvrier	The AI Toolbox
Education	15	2.48%	33.33%	Education	Education
Electives	9	1.49%	20.00%	Education	Education
Generative AI (e.g., ChatGPT)*	35	10.78%	100.00%	New Technology	New Technology
Total	606				

\*Code only occurred in wave two, percentages calculated for wave 2 data only

## Thematic Development

After the study team reached sufficient agreement on codes, the process of thematic analysis began with building upon the initial content coding. First, codes were categorized and abstracted into larger conceptual themes. Through discussion, themes and potential sub-themes were identified. Ultimately the study team settled on three themes with a total of ten sub-themes that accurately summarized the data. Given the format of the interview questions that were structured around the combined TAM-TPB model, themes unsurprisingly gathered around several constructs in the model. We then wrote the results section in a collaborative manner to ensure agreement on content.

## Results and Discussion

Corresponding with the interview guide which used the combined TAM-TPB framework, and the research questions regarding concepts in the model, our three main themes largely corresponded with the perceived usefulness, perceived norms, and behavioral control. However, in constructing our themes we drew relevant information from across the entire conversation instead of only responses to specific questions. Thus, we discovered new angles on what these oft-used concepts mean given the unique perspective of our participants. We first present the three main themes that occurred across both waves. Then, we discuss differences between wave one and wave two.

### AI is Tomorrow

The inevitability of AI in medical workplaces was a constant theme in interviews, so much that some codes and sub-themes (e.g., AI being demanded by administration and patients in the future workplace; AI use continuing to grow) occurred in 100% of interviews. Sentiment toward this future demand was not universal, with participants noting both pros (e.g., improved safety) and cons (e.g., need to retrain frequently). Regardless, no respondents saw a future without AI in the workplace, nor did any suggest that there will be large amounts of resistance to this change. Despite acknowledging the changes that AI may bring, respondents largely believed AI wouldn't change what it means to be a doctor. Some students felt that AI would not change the role of doctors very much, at least not in the near future. One student stated, "I think the fundamental roles of the doctor will still be there and won't be completely replaced" [14]. Another said, "I don't think AI will drive us out of business . . . down [that] path, hopefully AI is more friend than foe" [12]. There is a belief that AI is nearly synonymous with the workplaces the respondents will work in throughout their careers: AI is *not a "maybe"* in the workplaces of tomorrow, AI is tomorrow. And these workplaces will become further *saturated* with AI over time. Respondents understood that AI capabilities will increase, "you know as AI gets more precise and more experience, higher datasets, it can even go up the ladder and take up more specialised skills" [4]. While these possibilities are acknowledged, there is still uncertainty about just what AI will be capable of, and how quickly it will move, as summed up by one respondent: "AI will move faster than we expect but also slower than we expect, if you know what I mean" [35].



Addressing our first research question, we find a belief that the workplaces of tomorrow may be improved by AI, “I will be welcome to this sort of changes” said one respondent, continuing, “If it helps to make your job more efficient and more error-free, then I think probably it would be nice to welcome that sort of change” [7]. But, a less optimistic view was present in roughly 40% of interviews. For example, AI may bring problems to the future workplace, especially in the implementation side that is plagued by long timelines and testing, “They will experiment, they will trial, they will roll out a pilot programme, and that’s all fine and good, but for mass adoption, that will take a long time” [41]. In addition to dealing with long rollouts, respondents are not enthusiastic about the potential ethical questions introduced by AI:

“Then there’s also the whole legal aspect, like what if a patient is misdiagnosed? Computer says it’s a correct diagnosis, like the answer key says it’s correct, but in reality it was misdiagnosed like based on autopsy or tissue biopsy or too late, then it’s a dispute, like why did the doctor go against the decision of the computer, how valid is the clinician’s experience and what not.” [13]

Nevertheless, respondents see patients as being open to AI, perhaps even demanding its use from health care systems. Safety improvements may prompt this, “the overall impact [of AI] will be positive for the patient, then I think that I will also be glad to accept that change because I mean, I will accept the fact that probably we do make judgement errors” [2]. But this demand may follow generational shifts, “Probably the older generation [patients] will still prefer the more personal, they probably rely less on technology compared to the younger generation who might have differing thoughts on AI” [33]. Some respondents also offered a countertheme to this, suggesting patients will ultimately be unhappy, “. . . if you put AI’s formulaic way of thinking into the practice of medicine, then you would have a lot of unhappy patients and a lot of hurt patients” said one respondent [11].

All of these thoughts come with the assumption that AI will be adopted and will saturate the workplace. In turn, respondents shift their thinking from the traditional “perceived usefulness” as being a measure of likelihood of use. Rather, usefulness is a maybe but the presence of AI is not. And in truth, respondents saw usefulness as only a weak maybe. Barring a few pessimistic thoughts, the majority of comments saw this AI tomorrow as a better place, especially in reference to newer generative AI technologies (e.g., ChatGPT) present in the second wave of data collection. This answer to our first research question was reflected in our data, too, with optimistic futures codes occurring more than pessimistic ones. The results also provide some insight into our second research questions about norms. AI is the norm tomorrow; this AI tomorrow is inevitable and good.

### **Wishing for the AI Ouvrier**

If AI is an inevitable fixture on doctor’s careers, they want AI as a sidekick instead of a substitute. It is telling that the most frequently occurring code in this theme—the idea that AI is only a tool—occurred in nearly every interview (88.89%). The right tools will make the workplace more pleasant, as indicated by the discussion of AI’s potential to reduce workloads (40%) or ease other aspects of the profession (31.11%) provided that users will be

---

able to choose when and how to use AI (26.67%). One respondent recounted the benefits of AI they have heard of in other industries, “Not so much in the medical field, but in other applications like industrial applications, how it’s used to streamline predicting your supply chain demands, for business analytics” [5]. AI poses a threat when it acts as a substitute,

“I guess if [the AI] something that like doesn’t take away the joy of practising in that specialty, then I wouldn’t mind. The converse example is if I want to do surgery, but the AI is doing everything all the surgeries, then I don’t think I would want to do it. It takes away the joy of actually using your own hands and operating on the patient.” [6]

This quote deserves closer inspection: joy is under threat because it removes some degree of interaction. But the imagery of “actually using your own hands” is meaningful as well—people do things with our hands, it is the literal feel of the job. The removal of this physical connection with the job may represent a disembodiment prospect for doctors. But this fear does not manifest itself in a resistance to AI or a decrease in perceived usefulness. Instead, people wish to carefully pick specialties to avoid this fate. Again, we see the implication that AI is inevitable. One cannot resist an unstoppable force.

Respondents discussed the norms (RQ2) they foresee around AI use. The norms were often hopeful, respondents wish that AI will “serve as an adjunct,” [7] “provide second opinions,” [36] or, as one respondent candidly admitted their fallibility, AI can be a backup, “We do make mistakes, the AI could back us up” [40]. One respondent represented the thought of AI as a laborer for those tasks that are either time-consuming or difficult. The respondent pointed out the task of looking at diagnostic images, “I think what we’re all looking forward to is skipping the whole interpreting patho slides part, it would be amazing!” [9]. Another respondent expressed a similar thought bluntly: “If you’re talking about things that are more like clinical and diagnostic, like pattern and image recognition, then ya, even better AI does it, because we all suck at x rays and that kind of thing” [12]. AI not only makes up for the shortcomings of humans but takes up the tasks that doctors do not enjoy. *Ouvrier*, an obscure English word adopted from French, meaning “a workman who is employed to do heavy work requiring little skill” (Cambridge, n.d.), describes these hopeful norms well. AI can take difficult work, it can make participants better professionals, but AI sits below humans on the work-value chain—a sentiment that is perhaps exaggerated due to the norms and social prestige that come with being a doctor. While doctors still make leadership decisions and do hands-on work, it is expected of doctors to happily pass off undesirable to their *ouvrier* colleagues: AI.

## The Human Contrast

Our third research question probed the issue of autonomy as a potential source of tension between human and machine. Through the lens of TAM-TPB, this is a question about the control that humans will have over AI: What things will humans always be best at, what things will AI instead be in the driver’s seat? In our interviews, this discussion elicited many comparisons of human and machine qualities, typically involving comparisons of machine cognitive skills compared to human intuition (26.67%). We were surprised that the newest

generative AI technologies present in wave two did not change the nature of the human versus machine comparisons participants discussed. For example, participants believe that AI cannot control relationships with patients, including via the physical touch hinted at by other respondents above, “the personal connection with your patient, stuff like physical exam and things like that, I think those are things that AI cannot replace” [6]. Being a human health care provider is not only about touch, humans also provide assurance. One respondent put them in the position of patient, “As patients we go to a clinic to seek reassurance or to talk to someone about something you couldn’t tell anybody, so probably I’ll still find a human doctor” [34]. Another echoed the sentiment:

“If it’s something that’s worrying me, something that is out of the ordinary, even if it’s something that is quite straightforward for a doctor right, then I think I would want some reassurance, like if it’s a machine, I don’t think I would get a lot of reassurance.” [15]

One respondent offered a blanket statement about why humans still will remain the defining part of health care, “. . . in the end, healthcare is a service, and you still need the human touch as they call it” [8]. The unsaid implication similar to the issues present in other human-machine communication work, such as the need for machines to recognize emotion to be effective in health care (Kim et al., 2021). Or, in other words, can machines really provide the care part of health *care*.

Other respondents dismissed the idea that AI becomes controlling of human doctors because of the onerous requirements of creating effective AI systems, “I think if you want to implement an AI to that, it needs a s\*\*\* ton of data,” said one [3]. Another discussed human intuition,

“I think a lot of the work that doctors do is also very intuitive, for example they see certain signs and they’re able to synthesise what’s going on based on their clinical acumen. I think it’s probably going to take quite a while or even ever for a system to be able to synthesise that amount of knowledge, not just hard knowledge that you can feed into a computer but probably also some other kinds of soft cues.” [7]

Here, “hard cues” are the domain of the machine. But humans provide a softer touch; one that is out of touch for machines. But machines can also be superior to humans (33.33%). Many acknowledged the strengths that AI has which give it an advantage over humans. One respondent said, “If you have something that can process a million times for information than we can, and really start to see the patterns even we forget, [it can] make things more streamlined” [22]. Another alluded to the amount of health care research and evidence-based treatments possible, “I think our human minds really cannot fight that amount of information, and if it does get that far then that would be an invaluable tool” [24]. It is clear that the respondents understood that computers and machines have much greater “processing power” than human beings can muster, even if this does not manifest as control.

---

There were some limited places students could see AI put in the controller's seat, mainly on the administrative side of health care. For example, "You'll probably see it adopted earlier in the operational side of things, because that is the part where AI is already engaged, it's all numbers, optimising bed slots, allocation of resources, so that's probably where it's gonna start" was the opinion of one student [5]. Another student talked about "Crowd control, or scheduling doctors for clinic appointments, like how many doctors you might need at a particular time of the year, if it coincides with say flu season and travel incoming, even operational things like that" [12]. While these limits on human control may primarily originate from management, "Whether they want to let AI have that much power . . ." [21] respondents still foresee having control over the use of AI generally. It would be an exaggeration to say these predictions are made with full confidence, though. One respondent hinted at some nervousness, "[If] the research does show that AI is significantly better than a human at certain tasks then well, so be it. I'm not sure that day would be anytime in the near future, but that could be famous last words" [7].

## Pre- and Post-ChatGPT Differences

### ***Repositioning on the Timeline of Machinery***

Our first theme, "AI is Tomorrow," is in reference to many participants discussing AI as being the future, whereas they themselves are currently not in the future but are ready to learn. Thus, we could rename the theme, "AI is Tomorrow, I am Today." This sentiment is best reflected in the wave one data. If working solely off the wave two data we may be tempted to rename this theme, "AI is Today, *I am Yesterday*." It is difficult to express the distinct sense of discomfort that was palpable in some wave two interviews, especially when discussing the latest technologies. Every participant who mentioned ChatGPT discussed it as a user of the technology rather than just an observer. Several participants remarked how "good," "fast," "expansive," and "revolutionary" the technology is. As is well-documented in recent surveys, participants were using ChatGPT as everything from a writer of class assignments, to a study-buddy and even (as admitted by two) a second-opinion or quick reference tool in the clinic (The Learning Network, 2023). The key difference here is not that there is an overwhelming sense of pessimism in wave two interviews, but rather a sense of resignation. Wave two participants seem keenly aware that AI systems of the future (and perhaps current) will give patients faster, more detailed and—importantly—*better* answers than they themselves can. One participant made a telling statement, "My Aunt is always asking me questions about her conditions like I know something, I am only student lor . . . but let's say right now she ask me question, do I go to my own knowledge, look in a [text]book or something, or just use ChatGPT and see if it is a reasonable answer [pause], ChatGPT of course!" [43]. Another student spoke about ChatGPT's communication skill,

"It is amazing. If I ask it questions about something like pain management I will prompt it to speak empathetically. I say speak like Brené Brown, it delivers a better answer than me. It is soft and caring, not like me." [42]

Another said it quite directly, concurring with other participants who spoke of the sudden advance of AI technology, when they said that chatting with a computer just “went from zero to hero” overnight [23].

While these quotes reflect the sentiment well, it was also reflected in the participants’ evaluation of their education. While we elected not to have any themes specifically about education, when the topic arose in wave two the sentiment was distinctly more pessimistic. In wave one, participants wanted to learn more about AI but understood why they did not; by wave two the participants were frustrated, perhaps some even bitter, that they did not. Many participants spoke of the incredible rise of ChatGPT, coming out of seemingly nowhere and being “orders of magnitude” better than any previous technology for seeking medical information [39]. We believe this represents an overall sense of feeling behind technology. Before, participants are aware the technology is advanced, they are aware it is improving, worryingly so. But they still hold the keys to their success, they feel as if investment now in education will allow them to keep up with technology. In wave two, the situation feels more hopeless. AI has not just inched ahead, it has leapt ahead, and the gap is only growing. In turn, there is a clear sense of participants feeling that they are a past landmark on the timeline of AI’s advance, rather than the status quo of AI being a tool on the timeline of human advance.

### ***Better Together; Worse Alone***

We were especially intrigued by a number of participants who expressed what can best be described as lowered self-confidence after becoming familiar with ChatGPT. “It made me wonder ‘what am I doing here?’” said one participant after becoming familiar with the technology [34]. Another said, “I have no doubt this morphs into something that makes for better care, we all want that, but I also want to have something to contribute” [38]. To be transparent, we only discovered this theme about one third through the second wave interviews. This may have caused the interviewers to frame some questions or prompts differently, so we interpret this theme with caution. However, there is a silver lining here which is much more clear in the data than the sense of decreased self-confidence: increased machine and self-confidence. In other words, participants feel enlightened and enabled by ChatGPT, it unlocks opportunities they did not have before due to human constraints.

Perhaps this new sense of confidence in human-machine hybridity is best summed up by a participant describing her enjoyment of using ChatGPT, “Honestly I find it . . . invigorating! I really enjoy trying to find that answer and wording I am looking for; I may know what it knows but it knows how I want to say it” [35]. Notably, the context of this comment was the participant describing homework assignments where she is instructed to describe *what you would say to a patient* in given scenarios. Another discussed wishing to find work environments that facilitate the use of the technology, “It is the way, [telemedicine] always seemed nice and all you know less prep less buffer, but if I can use this it is a game-changer, I won’t be able to type fast enough in-person” [45]. We noted many similar instances of participants—enabled by AI—feeling more knowledgeable, more well-resourced (e.g., time), and far more confident in their communication skills. ChatGPT has ushered in a new self-confidence through human-machine hybridity.

---

## Discussion

The rapidly evolving human-machine communication landscape continues to redefine our sense of human-machine communication—our study contributes to this understanding both in terms of theory and practice. Our research shows the utility of the combined TAM-TPB model for investigating the attitudes, perceived norms, and behavioral control which are antecedents to communicating with machines. Attitudes and expectations are a major theme in human-machine communication research (e.g., Dearing, 2021; Gambino et al., 2020), but studying the underlying componentry of them has been challenging. We adopt a qualitative approach similar to previous work investigating the underlying assumptions people have about machines (Guzman, 2020), but structure it with a model providing new perspective. We make several resulting contributions to human-machine communication theory.

### Utility of Combined Theory

The study of human-machine communication can be approached from a multitude of theoretical and methodological backgrounds (Fortunati & Edwards, 2020). As such, we see that the use of a theoretical framework that combines two models is useful for discovering new perspectives, and the unique dynamics of human-machine communication can bring new meaning to oft studied concepts. The TAM-TPB model provided us with practical concepts such as usefulness and efficacy but also catered to the forward-looking nature of our sample, who are not professionals yet, by asking them to consider future norms. Our results provided a new angle on the notion of perceived usefulness, for example, is most frequently conceptualized as a cause for use or non-use of machines. Our research suggests the meaning of the concept shifts in situations where use of automation feels inevitable: usefulness becomes a proxy for the enjoyment of interacting with machines and their effects—positive or negative—on the work environment. This is an important theoretical consideration in future research given the pace that machinery is being adopted in the workplace and given that communicating with machines in the workplace is becoming less of “just an option” and more often a “compulsory” part of the workflow, regardless if the machines are thought to be useful (Bulchand-Gidumal, 2022, p. 18).

We also show how studying norms can encourage creative thinking and reveal implicit beliefs about machines. We find that asking people to envision future norms can elicit the hopes and fears of participants. When looking far enough forward, norms are infused with as much hope as they are with fact, and provide a window into uncertainties that people have regarding their future relationships with machines. Structuring our question around future norms provided a different approach to study uncertainty than that used in extant work that investigates workers already on the job (e.g., Piercy & Gist-Mackey, 2021). It is significant that our findings echo this work. Piercy and Gist-Mackey (2021, p. 191) found that pharmacy professionals can experience “automation anxieties,” for example. Our work suggests that these anxieties are not entirely a result of machines coming onto the workplace, but these anxieties may be present far before professionals enter the workforce at all.

People may become more confident in their abilities when paired with AI. Our findings in this area (salient post-ChatGPT) show that the conflict between human and



machine autonomy can be studied from a perspective of control. Our line of questioning derived from the concept of self-efficacy was especially well-tuned to exploring this phenomenon. Given that industry is perhaps only in the very beginning of a “Cambrian explosion” (Matsuoka, 2018) of big data and AI development, this dynamic will take on greater relevance as future disruptive technologies are introduced and redefine human-machine communication. Thus, the seesaw-like sense of individual confidence decreasing but individual-plus-machine confidence increasing—witnessed in our study—is a promising area for future research. The instances we recorded of participants mentioning communication specifically (e.g., discussing pain management), are especially intriguing, and the question is broader than just medical students and medical settings. Further research should be conducted on how confidence in communication skills (among other skills) is affected by the introduction, use, and expertise with new machine communication technologies.

### **Temporal Dynamics of Machine Agency**

The dynamics of human and machine agency manifest in a unique way in our study. In describing our results, we leaned heavily on the notion of time and control. However—in interpreting our findings—it is a mistake to simply reduce time to a linear concept, with a before, during, and after. For example, just describing humans or machines as ahead or behind is not an accurate characterization because behind implies inferiority, which is not always reflected in our data. Rather, the emergence of ChatGPT has affected participants in that they are no longer in control of their own timelines in regard to technology. In wave one, the self is the reference point on the timeline of experience, skill, and knowledge. Post-ChatGPT, the machines are the reference point. Machines control the timeline and advance at will, humans remain stationary, bound by unchangeable cognitive, emotional, and time limits. Hence, humans surrender their timelines and now live on the timeline of machinery.

What emerges from this, perhaps, is a new structure of how professional knowledge, practice, and standards are set. Throughout history, there is no obvious challenge to humans being the standard-bearers in all of these domains. Hence, the accumulation of knowledge, for example, in medicine is determined by what people determine to be correct. But the emergence of ChatGPT, as just the first in an inevitable line of improving AI technologies, appears to be tearing down this human-controlled structure. It is reminiscent of the work of Gibbs et al. (2021) who contend that technological systems can create new structures in workplaces. However, what we witness here extends beyond the workplace and into the personal mentalities of future professionals. AI is not the relatively simple algorithms that are “continually produced and reproduced by human action,” or that “evolve in a recursive relationship with human actors” (Gibbs et al., 2021, p. 165). Rather, AI systems such as ChatGPT are on the cusp of a transformative era where they go beyond enhancing humans and become autonomous entities that redefine the pace of knowledge acquisition and application. In this future, AI pioneers the benchmarks of efficacy and efficiency in various professional domains, forcing humans to adapt to machines’ pace rather than shape machines to human needs. Therefore, our data suggests a paradigm shift in human-machine interaction. Historically, machines operated within the confines of their programming. The generative AI revolution is changing this. It is vital to grasp the nuance here: This isn’t about

---

a machine's dominance over humans, nor is it about machines making humans obsolete. Instead, it's about machines setting a pace that humans struggle to match. Machines, largely a passive tool in the existing workplace, are transitioning dynamic communicators that redefine the contours of human expertise.

This shift has practical and theoretical implications. If the bar of professional excellence is set by a machine, then humans must chase ever-changing—perhaps even elusive—standards. It's not just about keeping up anymore; it's about continuously recalibrating one's knowledge and skills to synergize with machine capabilities. For human-machine communication theory, we see an interesting convergence point with Banks and de Graaf's (2020) notion of agency that is a foundational concept of their agent-agnostic model of transmission. Banks et al. describe agency as the capacity to make a difference through action (p. 28). Multiple participants in our study mentioned that ChatGPT can do a better job of communicating with patients than themselves. This sentiment is echoed broadly: Earlier this year an article suggesting ChatGPT could be preferred to human doctors made a stir on social media and garnered significant media coverage (McPhillips, 2023). So, when machines, powered by AI, start defining communication goals, deciphering meaning, and suggesting how a doctor should communicate with a patient, aren't they exercising a form of agency? Our study suggests so, and while these scenarios are futuristic now, they are clearly salient in the minds of the future doctors we spoke with. Thus, in the current moment, the perceived agency of these machines isn't derived merely from their ability to communicate or take action, but from their newfound role as the *timekeepers* of knowledge evolution. Their rapidly improving ability to accumulate and process information is essentially redrawing the temporal boundaries of human learning, professional growth, and expertise.

## Conclusion

Human-machine communication takes place in many contexts both personal and professional. Expectations, attitudes, and beliefs about machines affect the way that people interact with them. This research takes a unique study population—medical students on the cusp of long careers—to take a step backward in the timeline of attitudes toward AI. Our research shows that these attitudes can be strong even without extensive interaction with AI in the workplace. Overall, we find that attitudes are generally positive toward the use of AI, but some hesitation remains. The most salient norms are the ones medical students hope for, namely that AI is primarily a tool and acts as an *ouvrier* for less desirable tasks. They also believe that AI will be introduced to fulfill organizational goals, and they may not be granted autonomy to use or not use AI in these largely administrative functions. But in their personal workflows and relationships with patients, future doctors believe that they will have control over AI tools; humans remain the boss. We also witness that the introduction of a new, revolutionary technology can affect people's sense of control over their own personal development in relation to machines, and affect their confidence in a number of domains, including communication. When the future doctors we interviewed move throughout their careers, future machines will inevitably be there as well. Fortunately, for most, this future seems a better place.

## Author Biographies

**Andrew Prah** (PhD, University of Wisconsin-Madison) is an Assistant Professor in the Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore. His research frequently compares humans and machines, including human workers make sense of machines in the workplace or society more generally understands machines. Andrew's research on machines has spanned multiple industries including health care, aviation, journalism, humanitarian aid, and public relations.

 <https://orcid.org/0000-0003-3675-3007>

**Kevin Tong Wen Jin** is a medical student in the Lee Kong Chian School of Medicine at Nanyang Technological University, Singapore. His research interests include the integration of technology into medical education and the future of AI in health care.

## References

- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179–211. [https://doi.org/10.1016/0749-5978\(91\)90020-T](https://doi.org/10.1016/0749-5978(91)90020-T)
- Banks, J., & de Graaf, M. M. A. (2020). Toward an agent-agnostic transmission model: Synthesizing anthropocentric and technocentric paradigms in communication. *Human-Machine Communication*, 1(1), 19–36. <https://doi.org/10.30658/hmc.1.2>
- Bartholomew, J., & Mehta, D. (2023). *How the media is covering ChatGPT* (Columbia Journalism Review). Tow Center, Columbia University. [https://web.archive.org/web/20230608003831/https://www.cjr.org/tow\\_center/media-coverage-chatgpt.ph](https://web.archive.org/web/20230608003831/https://www.cjr.org/tow_center/media-coverage-chatgpt.ph)
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Bulchand-Gidumal, J. (2022). Impact of artificial intelligence in travel, tourism, and hospitality. In Z. Xiang, M. Fuchs, U. Gretzel, & W. Höpken (Eds.), *Handbook of e-Tourism* (pp. 1943–1962). Springer International Publishing. [https://doi.org/10.1007/978-3-030-48652-5\\_110](https://doi.org/10.1007/978-3-030-48652-5_110)
- Byrne, D. (2022). A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & Quantity*, 56(3), 1391–1412. <https://doi.org/10.1007/s11135-021-01182-y>
- Cambridge. (n.d.). *Ouvrier*. Retrieved October 30, 2022, from <https://web.archive.org/web/20240426154916/https://dictionary.cambridge.org/dictionary/french-english/ouvrier>
- Cho, S. I., Han, B., Hur, K., & Mun, J.-H. (2021). Perceptions and attitudes of medical students regarding artificial intelligence in dermatology. *Journal of the European Academy of Dermatology and Venereology*, 35(1), e72–e73. <https://doi.org/10.1111/jdv.16812>
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319–340. <https://doi.org/10.2307/249008>
- Dearing, J. (2021). What will affect the diffusion of AI agents? *Human-Machine Communication*, 3(1). <https://doi.org/10.30658/hmc.3.6>

- Evans, J. (2019). The post-exponential era of AI and Moore's Law. *TechCrunch*. <https://web.archive.org/web/20191111055828/https://techcrunch.com/2019/11/10/the-post-exponential-era-of-ai-and-moores-law>
- Fortunati, L., & Edwards, A. (2020). Opening space for theoretical, methodological, and empirical issues in human-machine communication. *Human-Machine Communication, 1*(1). <https://doi.org/10.30658/hmc.1.1>
- Fortunati, L., & Edwards, A. (2021). Moving ahead with human-machine communication. *Human-Machine Communication, 2*(1). <https://doi.org/10.30658/hmc.2.1>
- Gambino, A., Fox, J., & Ratan, R. (2020). Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication, 1*(1). <https://doi.org/10.30658/hmc.1.5>
- Gibbs, J., Kirkwood, G., Fang, C., & Wilkenfeld, J. (2021). Negotiating agency and control: Theorizing human-machine communication from a structural perspective. *Human-Machine Communication, 2*(1). <https://doi.org/10.30658/hmc.2.8>
- Gibson, A. M., Ryan, T. J., Alarcon, G. M., Jessup, S. A., Hamdan, I. A., & Capiola, A. (2020). Are all perfect automation schemas equal? Testing differential item functioning in programmers versus the general public. In M. Kurosu (Ed.), *Human-Computer Interaction. Human Values and Quality of Life* (pp. 436–447). Springer International Publishing. [https://doi.org/10.1007/978-3-030-49065-2\\_31](https://doi.org/10.1007/978-3-030-49065-2_31)
- Gong, B., Nugent, J. P., Guest, W., Parker, W., Chang, P. J., Khosa, F., & Nicolaou, S. (2019). Influence of artificial intelligence on Canadian medical students' preference for radiology specialty: A national survey study. *Academic Radiology, 26*(4), 566–577. <https://doi.org/10.1016/j.acra.2018.10.007>
- Guzman, A. L. (2020). Ontological boundaries between humans and computers and the implications for human-machine communication. *Human-Machine Communication, 1*(1). <https://doi.org/10.30658/hmc.1.3>
- Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*(1), 77–89. <https://doi.org/10.1080/19312450709336664>
- Hoff, K. A., & Bashir, M. (2015). Trust in automation integrating empirical evidence on factors that influence trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 57*(3), 407–434. <https://doi.org/10.1177/0018720814547570>
- Hu, K. (2023). ChatGPT sets record for fastest-growing user base—Analyst note. *Reuters*. <https://web.archive.org/web/20230205085718/https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01>
- IBM. (2020, July 7). *What is Artificial Intelligence (AI)?* <https://www.ibm.com/cloud/learn/what-is-artificial-intelligence>
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263–292. <https://psycnet.apa.org/doi/10.1017/CBO9780511609220.014>
- Kim, D. K., Kreps, G., & Ahmed, R. (2021). Communicative development and diffusion of humanoid AI robots for the post-pandemic health care system. *Human-Machine Communication, 3*(1). <https://doi.org/10.30658/hmc.3.5>

- The Learning Network. (2023, February 2). What students are saying about ChatGPT. *The New York Times*. <https://web.archive.org/web/20230203010131/https://www.nytimes.com/2023/02/02/learning/students-chatgpt.html>
- Matsuoka, S. (2018). Cambrian explosion of computing and big data in the post-Moore era. *Proceedings of the 27th International Symposium on High-Performance Parallel and Distributed Computing*, 105. <https://doi.org/10.1145/3208040.3225055>
- McPhillips, D. (2023). ChatGPT may have better bedside manner than some doctors, but it lacks some expertise. *CNN*. <https://web.archive.org/web/20230502010540/https://www.cnn.com/2023/04/28/health/chatgpt-patient-advice-study-wellness/index.htm>
- Park, C. J., Yi, P. H., & Siegel, E. L. (2021). Medical student perspectives on the impact of artificial intelligence on the practice of medicine. *Current Problems in Diagnostic Radiology*, 50(5), 614–619. <https://doi.org/10.1067/j.cpradiol.2020.06.011>
- Piercy, C., & Gist-Mackey, A. (2021). Automation anxieties: Perceptions about technological automation and the future of pharmacy work. *Human-Machine Communication*, 2(1). <https://doi.org/10.30658/hmc.2.10>
- Presbitero, A., & Teng-Calleja, M. (2022). Job attitudes and career behaviors relating to employees' perceived incorporation of artificial intelligence in the workplace: A career self-management perspective. *Personnel Review*, 52, 1169–1187. <https://doi.org/10.1108/PR-02-2021-0103>
- Richter, A., & Näswall, K. (2019). Job insecurity and trust: Uncovering a mechanism linking job insecurity to well-being. *Work & Stress*, 33(1), 22–40. <https://doi.org/10.1080/02678373.2018.1461709>
- Russell, S., & Norvig, P. (2009). *Artificial intelligence: A modern approach*. Prentice Hall, USA.
- Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors*, 58(3), 377–400. <https://doi.org/10.1177/0018720816634228>
- Scott, I. A., Carter, S. M., & Coiera, E. (2021). Exploring stakeholder attitudes towards AI in clinical practice. *BMJ Health & Care Informatics*, 28(1), e100450. <https://doi.org/10.1136/bmjhci-2021-100450>
- Selenko, E., Bankins, S., Shoss, M., Warburton, J., & Restubog, S. L. D. (2022). Artificial intelligence and the future of work: A functional-identity perspective. *Current Directions in Psychological Science*, 31(3), 272–279. <https://doi.org/10.1177/09637214221091823>
- Shah, R., & Chircu, A. (2018). IOT and AI in healthcare: A systematic literature review. *Issues in Information Systems*, 19(3). [https://doi.org/10.48009/3\\_iis\\_2018\\_33-41](https://doi.org/10.48009/3_iis_2018_33-41)
- Siegel, E. (2019). The media's coverage of AI is bogus. *Scientific American Blog Network*. <https://web.archive.org/web/20191120170530/https://blogs.scientificamerican.com/observations/the-medias-coverage-of-ai-is-bogus>
- Simmler, M., Brunner, S., Canova, G., & Schedler, K. (2022). Smart criminal justice: Exploring the use of algorithms in the Swiss criminal justice system. *Artificial Intelligence and Law*. <https://doi.org/10.1007/s10506-022-09310-1>
-

- 
- Sit, C., Srinivasan, R., Amlani, A., Muthuswamy, K., Azam, A., Monzon, L., & Poon, D. S. (2020). Attitudes and perceptions of UK medical students towards artificial intelligence and radiology: A multicentre survey. *Insights into Imaging, 11*(1), 14. <https://doi.org/10.1186/s13244-019-0830-7>
- Stam, K. R., Stanton, J. M., & Guzman, I. R. (2004). Employee resistance to digital information and information technology change in a social service agency: A membership category approach. *Journal of Digital Information, 5*(4), Article 4. <https://jodi-ojs-tdl.tdl.org/jodi/article/view/jodi-156>
- Sun, S., Zhai, Y., Shen, B., & Chen, Y. (2020). Newspaper coverage of artificial intelligence: A perspective of emerging technologies. *Telematics and Informatics, 53*, 101433. <https://doi.org/10.1016/j.tele.2020.101433>
- Taylor, S., & Todd, P. (1995). Assessing IT usage: The role of prior experience. *MIS Quarterly, 19*(4), 561–570. <https://doi.org/10.2307/249633>
- Toews, R. (2021). Artificial intelligence and the end of work. *Forbes*. <https://web.archive.org/web/20210215234711/https://www.forbes.com/sites/robtoews/2021/02/15/artificial-intelligence-and-the-end-of-work>
- UBS. (2023). *Let's chat about ChatGPT* (p. 4). <https://web.archive.org/web/20231207224625/https://www.ubs.com/us/en/wealth-management/insights/market-news/article.1585717.html>
-





# What's in a Name and/or a Frame? Ontological Framing and Naming of Social Actors and Social Responses

David Westerman<sup>1</sup> , Michael Vosburg<sup>2</sup> , Xinyue "Gordon" Liu<sup>1</sup>,  
and Patric R. Spence<sup>3</sup> 

1 Department of Communication at North Dakota State University, Fargo, North Dakota, USA

2 Department of Mass Communication at Benedict College, Columbia, South Carolina, USA

3 Nicholson School of Communication and Media, University of Central Florida, Orlando, Florida, USA

## Abstract

Artificial intelligence (AI) is fundamentally a communication field. Thus, the study of how AI interacts with us is likely to be heavily driven by communication. The current study examined two things that may impact people's perceptions of socialness of a social actor: one nonverbal (ontological frame) and one verbal (providing a name) with a 2 (human vs. robot) × 2 (named or not) experiment. Participants saw one of four videos of a study "host" crossing these conditions and responded to various perceptual measures about the socialness and task ability of that host. Overall, data were consistent with hypotheses that whether the social actor was a robot or a human impacted each perception tested, but whether the social actor named themselves or not had no effect on any of them, contrary to hypotheses. These results are then discussed, as are directions for future research.

**Keywords:** social robots, artificial intelligence, electronic propinquity, perceived humanness, attraction, source credibility

## Introduction

Artificial intelligence (AI) is fundamentally a communication field (Gunkel, 2020). Dating back to the classic foundations of AI, what has come to be known as the Turing Test (1950),

**CONTACT** David Westerman  • [david.k.westerman@ndsu.edu](mailto:david.k.westerman@ndsu.edu) • Department #2512 • P.O. Box 6050 • Fargo, ND 58108

ISSN 2638-602X (print)/ISSN 2638-6038 (online)  
[www.hmcjournal.com](http://www.hmcjournal.com)



Copyright 2024 Authors. Published under a Creative Commons Attribution 4.0 International (CC BY-NC-ND 4.0) license.

or how we come to perceive *artificial intelligence* as simply *intelligence*, is driven by how a social actor communicates. Thus, the study of how AI interacts with us is likely to be heavily driven by communication (Westerman et al., 2020). But what seems human, in both humans and AI such as robots? This study examined two ways a robot may communicate its socialness to us; one nonverbal (ontological frame) and one verbal (providing a name) to help address how each may be important for establishing initial perceptions of socialness.

## Social Responses to Social Actors

As Turing (1950) suggested, people's perceptions of AI will drive their response to it. Turing specifically suggested that the best way for an AI to pass what became known as the Turing Test would be to use what we today might call text-only computer-mediated communication (CMC); to use only words and not be seen. Seeing an entity is one of the most significant ways to be tipped off that the social actor is not human. Guzman (2020) and A. Edwards (2018) argued that these ontological boundaries of origin are becoming more complicated and vital to consider (see also Etzrodt & Engesser, 2021). Constructivist approaches to communication suggest that social interactions “unfold first and foremost through a process of prototyping the potentially communicative other (what is it?),” which links to stereotyping processes (what does it do?), which then influence perception and messaging, and behavior by driving the creation and use of interaction scripts (A. Edwards, 2018, p. 45). When people interact with other people, they occupy the same ontological space (Dautenhahn, 2004). However, previous studies have demonstrated that both verbal and visual primes suggesting the nature of an actor as a robot versus a human lead to lower expectations regarding interaction sociality such as social presence/electronic propinquity and liking (e.g., A. Edwards et al., 2019) and that ontological/agent-type category cueing can activate heuristics affecting interpretations of agent behavior (e.g., Banks et al., 2021).

Initial interactions are largely driven by our expectations of such interactions, and much about those expectations are scripted (Kellerman, 1992). One such relevant script that has been identified is the human-human interaction script (A. Edwards et al., 2019; C. Edwards et al., 2016; Spence et al., 2014). In general, this script suggests that when people communicate, they expect their partner to be human. When that partner is not human, people have lower expectations about how much social presence/electronic propinquity they will feel in the interaction (A. Edwards et al., 2019; C. Edwards et al., 2016; Spence et al., 2014). Electronic propinquity can be defined as a “psychological feeling of nearness” (Walther & Bazarova, 2008, p. 624) or even more simply as “electronic presence” (Korzenny, 1978, p. 7). Kelly and Westerman (2016) have argued that electronic propinquity is similar to/the same as some other concepts, such as social presence and perceived immediacy, and suggests an experience of interpersonal connectedness between two social actors and the degree of realness perceived in one's communication partner. Thus, if a person is greeted by a robot rather than a human, we expect to see the same perceptions. But even before that, we would also expect a human to be perceived as more human than a robot when that actor is seen. This leads to the first two hypotheses of the current study:

---

**H1:** A human will be perceived as more human than a robot.

**H2:** A human will be perceived as having more electronic propinquity than a robot.

The human-human interaction script predicts social perceptions of initial interactions with robots compared to humans. It may also help suggest possibilities for more task-based initial perceptions, such as task attraction and source credibility, in a similar way. Because the human-human interaction script involves initial encounters, perceptions and attributions may follow similar patterns to what has been seen in psychological research (see A. Edwards et al., 2019). Several studies have found task and social attraction to manifest in communication with humans and machines and impact these perceptions of the communicator. For example, a study by Beattie et al. (2020) examined differences in perceptions of a message sender using emojis when the identity of the message sender was either a human or a chatbot. Emoji use was perceived as more socially attractive than non-emoji use, whether the source was a human or a chatbot.

Cross-cultural studies and studies on stereotypes (in-group vs. outgroups) have shown that “others” are often stereotyped negatively as cold and incompetent (Lee & Fiske, 2006). Task attraction is related to competence, it has been noted that competence refers to capability reflecting the targets’ ability to put their intentions into practice (Cuddy et al., 2007). Thus, there may be differences in perceptions of task attraction based on the perception of similarity if the target is a human or robot. Task attraction is defined by McCroskey and McCain (1974) as “how easy or worthwhile working with someone would be” (p. 266). If people’s expectations about an initial interaction are violated, this may create uncertainty (Burgoon, 1993), leading to lower perceptions about how easy it may be to work with a partner, and thus, lower task attraction, in line with other perceptions related to the human-human interaction script. Moreover, there may be tasks that are viewed as appropriate for a machine in a given context whereas some tasks might be seen as best completed by a human. There still exists as perception of human-to-human communication as the “gold standard” of communication (Spence, 2019). Following this logic, two studies by Spence et al. (2019, 2021) examined these perceptions in the context of weather forecasts. The first study (Spence et al., 2019) had conditions that included a professional meteorologist’s X feed (Twitter), the X feed of a weatherbot, and that of an amateur meteorologist. That study found that respondents perceived the professional meteorologist as more socially attractive than the weatherbot, with no differences in task attraction. The weatherbot was perceived as more task attractive than the amateur meteorologist with no difference in social attraction. The second study (Spence et al., 2021) employed a weather forecast from a local television station using a professional meteorologist, a television station weather robot, and an amateur weather forecaster. Respondents in that study indicated the highest perceptions of task and social attraction with the professional meteorologist followed by the amateur weather forecaster, and the weather robot creating the lowest levels of task and social attraction. The authors argue their results across the two studies taken together support the anthropomorphic bias outlined in the human–human interaction script. These studies add support to the following hypothesis offered:

**H3:** A human will be perceived as more socially attractive than a robot.

**H4:** A human will be perceived as more task attractive than a robot.

Similarly, source credibility (McCroskey & Teven, 1999), made up of perceptions of one's competence (expertise; or does someone know about something); trustworthiness (character and honesty; or can I trust that someone to be honest about that something), and goodwill (a perception of caring, or does that someone have my interests in mind) may also work similarly as evident in several studies (C. Edwards et al., 2021; Finkel & Krämer, 2022). This may be especially true for the goodwill aspect of credibility, as it is more social perception and could be primed by the presence of anthropomorphic cues. As noted in the Spence et al. (2021) study, perceptions of source credibility were highest with the professional meteorologist, followed by the amateur meteorologist, and the lowest levels of perceived source credibility emerged in the condition with the weather social robot. The authors note that all three dimensions of credibility (competence, trust, and goodwill) followed the same pattern.

However, even without the presence of an anthropomorphic cue, similar results have emerged. Research by Kim et al. (2022) examined a radio AI newscaster and a radio human newscaster in a broadcast concerning severe weather. There were higher perceptions of credibility for the human newscaster compared to the AI newscast. Thus, preferences for humans may not be the result of only a visual prime, but any prime concerning the humanness of the communicator. Given that credibility may be a more social perception and the previous study found differences between a human and robot in these perceptions the following hypothesis is offered:

**H5:** A human will be perceived as more credible (competence, trustworthiness, and goodwill) than a robot.

The look of a social actor is one that likely plays a part in our responses to that actor. However, it is not the only one. Evidence suggests that even if we “know” that something does not warrant a human/social response, we still may respond to that entity with one. The Computers are Social Actors paradigm (CASA; Nass & Moon, 2000; Nass et al., 1994; Reeves & Nass, 1996) is the basis of this position. This paradigm suggests that when a technology triggers a social response, we respond socially, especially if it stems from an overlearned heuristic (Nass & Moon, 2000). Thus, if a robot does something to trigger a heuristic of social action, we may be more likely to respond as we do to other social actors (i.e., humans).

The CASA paradigm suggests that interpersonal communication theory and research are relevant for considering AI and social robots (Spence et al., 2023; Westerman et al., 2020). Among humans, social relationships often begin and grow through self-disclosure, as this is how a social actor reveals themselves to another entity or social actor. One simple self-disclosure that individuates an actor and may begin the social process is providing one's name. Indeed, naming is a powerful speech act, as argued by Palsson (2014), and serves as a “technology of belonging.” Names impact people's impressions of the person as well (e.g., Young et al., 1993).

---

Presenting a self is what the goal of a social bot is, so it can present itself as being like another social actor, and thus, a relationship can be formed (Gehl & Bakardjieva, 2017). Social robots have been found to be preferred individually compared to in groups (Fraune et al., 2015), again suggesting that people want such bots to present more of a self. Fritz (2018) explained that having a human name rhetorically situates a robot to be interpellated as a subject, and connotes the uniqueness and “hailability” of a person or pet. Indeed, machines with names (as part of a variety of things) seem to be more anthropomorphized, whether it be autonomous vehicles (Waytz et al., 2014) or chatbots (Araujo, 2018), as well as increasing other social responses. Robots that have a name also seem to trigger more human/social responses to them (Darling, 2017). Given the general finding here that naming tends to be a socializing cue, the following hypotheses are offered:

**H6:** A social actor that provides a name will be perceived as more human than one that does not.

**H7:** A social actor that provides a name will be perceived as having more electronic propinquity than one that does not.

**H8:** A social actor that provides a name will be perceived as more socially attractive than one that does not.

As stated above, research on ingroups vs. outgroups have shown that “others” are often stereotyped negatively as cold and incompetent (Lee & Fiske, 2006). If providing a name within an introduction makes a social actor, such as a robot, be perceived as more similar to a person than if no name is provided, then this may lead to more positive impressions. Moreover, the act of providing a name may reduce perceived anonymity and cause a more favorable impression. Research has shown that individuals perceive higher levels of source credibility of risk information when the identity of the source is known (Lin et al., 2016). Other research has shown that the absence of individual identifications in various situations, such as voice changers, pseudonyms, and nicknames, have the ability to impact perceptions (Graf et al., 2017; Lin et al., 2019). Given these past findings, the following hypotheses are offered:

**H9:** A social actor that provides a name will be perceived as more task attractive than one that does not.

**H10:** A social actor that provides a name will be perceived as more credible (competence, trustworthiness, and goodwill) than one that does not.

It is not clear if/how these two pieces of information about a social actor would interact to impact these social and task related perceptions. Thus, a general research question (RQ) asks if there are interaction effects on any of these perceptions.



## Method

### Overview

A 2 (ontological frame; robot vs. human)  $\times$  2 (name; provided or not) between-subjects experiment was conducted to test the hypotheses offered in this study (and to test for possible interaction effects between the two independent variables). Participants were asked to log into a website, where, after providing informed consent, they were welcomed to the study by watching a video of either a robot or a human “host” for the study. This host either said their name as part of the greeting or not. These conditions were fully crossed and randomly assigned to participants, leading to four different conditions: human host that gave a name ( $n = 77$ ), human host that did not give a name ( $n = 80$ ), robot host that gave a name ( $n = 72$ ), and a robot host that did not give a name ( $n = 77$ ). After the greeting, participants were asked to respond to several measures about this greeter and then told the study was over.

### Participants

Data were collected from 332 participants recruited from an introductory communication course at a public university in the upper Midwestern United States. The removal of 20 participants that responded “yes” to a question asking if they recognized the host (7 in human host condition, 13 in robot host conditions) and six participants who failed to complete measures left data from 306 for analysis. One-hundred and forty-eight participants self-identified as male (48.4%), 147 (48.0%) as female, 2 (0.7%) as women, 3 (1.0%) as nonbinary, 1 (0.3%) as agender, with 5 (1.6%) not responding. The majority of participants self-identified as White/Caucasian ( $n = 263$ , 85.9%). Participants’ ages ranged from 18 to 41 years ( $M = 18.93$ ,  $SD = 2.15$ ). A sensitivity power analysis was conducted using G\*Power 3.1 (Faul et al., 2007), which suggested this sample size had 80% power for detecting effects with a Cohen’s  $f$  of .16 at the .05 level, which is a relatively small effect size.

### Stimulus Materials

Four different videos were created for this study; one for each of the four conditions. The videos can be seen at the following link: <https://osf.io/y62ak/>. There was one video for each of the human conditions. In these human conditions, a Caucasian, middle-aged looking and sounding male introduced the study. The specific male was chosen because, although he was a graduate student at the university where the data was collected at the time of data collection, he was not a teaching assistant, and therefore, combined with the fact that he was older than most students, it was considered less likely that students in the class that participants were recruited from would recognize him. The male was recorded from the waist up with a plain wall as the background. In condition A, the video consisted of the introduction to the research study in which a human male provided that introduction. The human male did not use a name in the introduction. It was 11 seconds in length and recorded in 720p HD with a frame height of 720 and width of 1280. The video both had a fade-in and fade-to-black. In condition B, the video format features were identical except that the name “Mike” was used by the human male in the introduction and the video was 12 seconds in

---

length. There was also one video for each of the robot conditions. In these robot conditions, the angle and the background were the same as in the human conditions, with the “host” being recorded from the “waist” up to the head in the same room as the human conditions; however, the experimental manipulation differed in that a robot delivered the script. The robot in the video was an Ohmni<sup>®</sup> Telepresence Robot with a graphic screen that projected a human-like face based on the MAKI Humanoid robot. This robot has been used in other studies (see Edwards et al., 2016; Michaelis & Mutlu, 2019; Rainear et al., 2021). In condition C, the script and video features were identical to condition A. The length of the video was 13 seconds. In condition D, the script and video features were identical to condition B; however, the experimental manipulation differed in that a robot delivered the script. The video was 14 seconds long. The voice for conditions C and D were taken from the audio files of videos A and B and then modified with the program Audacity to emulate a synthetic voice.

## Measures

### **Perceived Humanness**

After viewing one of the four videos (randomly assigned), participants responded to measures about the “host” they saw in the video. The first of these was a measure of perceived humanness, adapted from Bartneck et al. (2009), and previously used by Author. Using a five-point response set, this adaptation contained four semantic differential items (e.g., “machinelike-humanlike”). The scale had acceptable reliability ( $\alpha = .91$ ). Scores on this index ranged from 1 to 5, with a mean of 2.91 ( $SD = 1.23$ ).

### **Electronic Propinquity**

Electronic propinquity was measured using Walther and Bazarova’s (2008) scale. Using a seven-point response set, this measure consists of five semantic differential items (e.g., “disconnected-connected”). Scores on individual items were recoded so that higher scores on the index meant greater electronic propinquity. The scale had acceptable reliability ( $\alpha = .86$ ). Scores on this index ranged from 1 to 7, with a mean of 3.78 ( $SD = 1.23$ ).

### **Task Attraction**

Task attraction was measured using a version of McCroskey and McCain’s (1974) measure. Task attraction consisted of five items (e.g., “I couldn’t get anything accomplished with them”) using a seven-point response set. Scores on individual items were recoded so that higher scores on the index meant greater task attraction. The scale had acceptable reliability ( $\alpha = .73$ ). Scores on this index ranged from 1 to 7, with a mean of 4.68 ( $SD = 1.06$ ).

### **Social Attraction**

Social attraction was measured using a version of McCroskey and McCain’s (1974) measure. Social attraction consisted of six items (e.g., “They just wouldn’t fit into my circle of friends.”) with a seven-point response set. Scores on individual items were recoded so that higher scores on the index meant greater social attraction. The scale had acceptable reliability ( $\alpha = .87$ ). Scores on this index ranged from 1 to 6.5, with a mean of 3.90 ( $SD = 1.15$ ).

### Source Credibility

Source credibility was measured using a version of McCroskey and Teven's (1999) measure. There are three different types of credibility measured using this scale: competence, trustworthiness, and goodwill. One item ("bright-stupid") was removed from the original competence measure, leaving five semantic differential items (e.g., "informed-uninformed") with a seven-point response set used for analysis. Scores on individual items were recoded so that higher scores on the index meant greater competence. The scale had acceptable reliability ( $\alpha = .84$ ). Scores on this index ranged from 1 to 7, with a mean of 5.11 ( $SD = 1.12$ ). Trustworthiness consisted of six semantic differential items (e.g., "honest-dishonest") with a seven-point response set. Scores on individual items were recoded so that higher scores on the index meant greater trustworthiness. The scale had acceptable reliability ( $\alpha = .86$ ). Scores on this index ranged from 1 to 7, with a mean of 4.68 ( $SD = 1.11$ ). Goodwill consisted of six semantic differential items (e.g., "has my interests at heart—doesn't have my interests at heart") with a seven-point response set. Scores on individual items were recoded so that higher scores on the index meant greater goodwill. The scale had acceptable reliability ( $\alpha = .83$ ). Scores on this index ranged from 1 to 7, with a mean of 3.90 ( $SD = 1.17$ ). Please see Table 1 for overall descriptive statistics and correlations for each measured variable.

**TABLE 1** Descriptive Statistics and Correlations for Study Outcome Variables

Variable	$\alpha$	$M$	$SD$	1	2	3	4	5	6	7
1. Perceived Humanness	.91	2.91	1.23							
2. Electronic Propinquity	.86	3.78	1.23	.26**						
3. Task Attraction	.73	4.68	1.06	.52**	.23**					
4. Social Attraction	.87	3.90	1.15	.61**	.38**	.54**				
5. Competence	.84	5.11	1.12	.34**	.11	.57**	.22**			
6. Trustworthiness	.86	4.68	1.11	.60**	.23**	.59**	.48**	.69**		
7. Goodwill	.83	3.90	1.17	.74**	.43**	.53**	.67**	.33**	.64**	

Note. \* $p < .05$ , \*\* $p < .01$

## Results

In order to test the hypotheses and research question offered in this study, a series of  $2 \times 2$  Analyses of Variance (ANOVAs) were conducted for each dependent variable, with frame (human vs. robot) and name (did so or not) as the independent variables. In general, frame (whether the host was human or a robot) had a significant main effect on each dependent variable. Whether the host named themselves or not in the video had no significant main effect on any dependent variable, and there were no significant interaction effects for any variable. More details are included below. Please see Table 2 for descriptive statistics across condition, and Tables 3–9 for ANOVA details for each outcome variable.

**TABLE 2 Outcome Measure Means, Standard Deviations, and *N* Across Conditions**

	Robot		Human	
	Name	No name	Name	No name
Perceived Humanness	1.98 (.84) <i>n</i> = 72	2.12 (1.03) <i>n</i> = 76	3.70 (.90) <i>n</i> = 74	3.77 (.79) <i>n</i> = 80
Electronic Propinquity	3.58 (1.35) <i>n</i> = 70	3.68 (1.09) <i>n</i> = 77	3.82 (1.22) <i>n</i> = 77	4.01 (1.24) <i>n</i> = 78
Task Attraction	4.06 (1.07) <i>n</i> = 72	4.32 (1.12) <i>n</i> = 77	5.12 (.84) <i>n</i> = 77	5.16 (.75) <i>n</i> = 80
Social Attraction	3.23 (1.16) <i>n</i> = 72	3.53 (1.29) <i>n</i> = 76	4.31 (.87) <i>n</i> = 76	4.46 (.73) <i>n</i> = 79
Competence	4.66 (1.30) <i>n</i> = 71	4.87 (1.07) <i>n</i> = 76	5.39 (1.03) <i>n</i> = 76	5.49 (.89) <i>n</i> = 79
Trustworthiness	4.10 (1.08) <i>n</i> = 70	4.20 (1.05) <i>n</i> = 77	5.17 (.90) <i>n</i> = 77	5.17 (.93) <i>n</i> = 79
Goodwill	3.22 (1.10) 72	3.31 (1.21) 76	4.38 (.85) 77	4.61 (.80) 79

**TABLE 3 ANOVA for Perceived Humanness**

	<i>df</i>	<i>F</i>	<i>p</i>	$\eta^2$
Frame	1	266.00	< .001	.47
Name	1	1.09	.297	.00
Interaction	1	.11	.736	.00

**TABLE 4 ANOVA for Electronic Propinquity**

	<i>df</i>	<i>F</i>	<i>p</i>	$\eta^2$
Frame	1	4.09	.044	.01
Name	1	1.09	.296	.00
Interaction	1	.10	.753	.00

**TABLE 5 ANOVA for Task Attraction**

	<i>df</i>	<i>F</i>	<i>p</i>	$\eta^2$
Frame	1	75.99	< .001	.20
Name	1	1.87	.173	.00
Interaction	1	1.08	.300	.00

**TABLE 6 ANOVA for Social Attraction**

	<i>df</i>	<i>F</i>	<i>p</i>	$\eta^2$
Frame	1	72.01	< .001	.19
Name	1	3.84	.058	.01
Interaction	1	.44	.507	.00

**TABLE 7 ANOVA for Competence**

	<i>df</i>	<i>F</i>	<i>p</i>	$\eta^2$
Frame	1	29.38	< .001	.09
Name	1	1.58	.210	.00
Interaction	1	.21	.650	.00

**TABLE 8 ANOVA for Trustworthiness**

	<i>df</i>	<i>F</i>	<i>p</i>	$\eta^2$
Frame	1	80.32	< .001	.21
Name	1	.20	.656	.00
Interaction	1	.20	.656	.00

**TABLE 9 ANOVA for Goodwill**

	<i>df</i>	<i>F</i>	<i>p</i>	$\eta^2$
Frame	1	114.91	< .001	.28
Name	1	1.89	.171	.00
Interaction	1	.40	.525	.00

To test H1 and H6, which predicted that human framing of the study host and the host providing a name would lead to greater perceptions of humanness, a  $2 \times 2$  ANOVA was conducted on perceived humanness. Frame had a significant main effect on perceived humanness [ $F(1, 298) = 265.99, p < .001, \eta^2 = .47$ ] such that the human host ( $M = 3.73, SD = .84$ ) was seen as more human than the robot one ( $M = 2.05, SD = .94$ ). Thus, data were consistent with H1. There was no significant main effect of name on perceived humanness [ $F(1, 298) = 1.09, p = .297$ ], thus data were not consistent with H6. There was also no significant interaction effect between frame and name [ $F(1, 298) = .11, p = .736$ ].

To test H2 and H7, which predicted that human framing of the study host and the host providing a name would lead to greater perceptions of electronic propinquity with that host, a  $2 \times 2$  ANOVA was conducted on electronic propinquity. Frame had a significant main effect on electronic propinquity [ $F(1, 298) = 4.09, p = .044, \eta^2 = .01$ ] such that participants perceived more electronic propinquity with the human host ( $M = 3.92, SD = 1.23$ ) than the robot one ( $M = 3.63, SD = 1.22$ ). Thus, the data were consistent with H2. There was no significant main effect of name on electronic propinquity [ $F(1, 298) = 1.09, p = .296$ ],

data were not consistent with H7. There was no significant interaction effect between name and frame [ $F(1, 298) = .10, p = .753$ ].

To test H3 and H8, which predicted that human framing of the study host and the host providing a name would lead to greater social attractiveness of the host, a  $2 \times 2$  ANOVA was conducted on social attraction. Frame had a significant main effect on social attraction [ $F(1, 299) = 72.01, p < .001, \eta^2 = .19$ ] such that participants perceived the human host as more socially attractive ( $M = 4.39, SD = .80$ ) than the robot one ( $M = 3.38, SD = 1.23$ ). Thus, the data were consistent with H3. There was no significant main effect of name on social attraction [ $F(1, 299) = 3.61, p = .058$ ]; thus, the data were not consistent with H8. There was no significant interaction effect [ $F(1, 299) = .44, p = .507$ ].

To test H4 and H9, predicting that human framing of the study host and the host providing a name would lead to greater task attractiveness of the host, a  $2 \times 2$  ANOVA was conducted on task attraction. Frame had a significant main effect on task attraction [ $F(1, 302) = 75.99, p < .001, \eta^2 = .20$ ] such that participants perceived the human host as more task attractive ( $M = 5.14, SD = .80$ ) than the robot one ( $M = 4.19, SD = 1.10$ ). Thus, the data were consistent with H4. There was no significant main effect of name on task attraction [ $F(1, 302) = 1.87, p = .173$ ], and so the data were not consistent with H9. No significant interaction effect [ $F(1, 302) = 1.08, p = .300$ ] was found.

Finally, to test H5 and H10, predicting that human framing of the study host and the host providing a name would lead to greater perceived credibility of the host, a  $2 \times 2$  ANOVA was conducted on each of the three subcomponents of credibility. Frame had a significant main effect on competence [ $F(1, 298) = 29.38, p < .001, \eta^2 = .09$ ] such that participants perceived the human host as more competent ( $M = 5.44, SD = .96$ ) than the robot one ( $M = 4.77, SD = 1.19$ ). Frame also had a significant main effect on trustworthiness [ $F(1, 299) = 80.32, p < .001, \eta^2 = .21$ ] such that participants perceived the human host as more trustworthy ( $M = 5.17, SD = .91$ ) than the robot one ( $M = 4.15, SD = 1.06$ ). Frame also had a significant main effect on goodwill [ $F(1, 300) = 114.91, p < .001, \eta^2 = .28$ ] with participants seeing more goodwill from the human host ( $M = 4.50, SD = .83$ ) than the robot one ( $M = 3.27, SD = 1.16$ ). Thus, the data were consistent with H5. There was no significant main effect of name on perceived competence [ $F(1, 298) = 1.58, p = .210$ ], trustworthiness [ $F(1, 299) = .20, p = .656$ ], nor goodwill [ $F(1, 300) = 1.89, p = .171$ ]. Thus, the data were not consistent with H10. No significant interaction effects were found for competence [ $F(1, 298) = .21, p = .650$ ], trustworthiness [ $F(1, 299) = .20, p = .656$ ], nor goodwill [ $F(1, 300) = .40, p = .525$ ].

## Discussion

The current study was designed to examine the role that ontological frame (robot vs. human) and name (naming or not) of a social actor had on various perceptions of that actor. In general, the frame had significant effects on all dependent variables, such that the human host was seen as more human, electronically propinquitous, socially and task attractive, and all three components of credibility measured (competence, trustworthiness, and goodwill) than the robot host. Whether the host named themselves or not did not have significant effects on any of these perceptions, and there were also no interaction effects between frame and name. These results are discussed in more detail below.



First, consistent with hypotheses, a video of a human host introducing a study was perceived more positively overall than a robot one. This is very much in line with the human-human interaction script found in previous research (Craig & Edwards, 2021; Edwards et al., 2019; Edwards et al., 2016; Spence et al., 2014), suggesting that people expect to interact with humans when they know they are going to interact with a social actor. The current study suggests this script may also apply to other expectations of experiences with social actors as well, including the role of study *host*, introducing what people will be doing, as used here. Thus, perhaps the biggest practical application of the findings in this study are that humans may make for better greeters than robots overall, at least for the kind of one-time, noninteractive greeting examined in the current study, as the human-host conditions were perceived more positively overall as compared to the robot-host conditions. This might be especially important for some of the larger effects found in the current study, which cut across both work and social outcomes. For example, some of the stronger effects were found on task attraction ( $\eta^2 = .20$ ) and trustworthiness ( $\eta^2 = .21$ ), with the human host perceived as more task attractive and trustworthy than the robot host. This was also true for social attraction ( $\eta^2 = .19$ ) and goodwill ( $\eta^2 = .28$ ), with the human host perceived as higher on both of these than the robot host. Thus, it would seem that people engaging with this kind of *hosting* video both like and trust it more when a human is the one talking to them. Companies may want to consider being very careful about using a robot for this purpose, unless there is good reason to do so.

As mentioned above, although there were significant differences found for ontological frame on each variable of interest in this study, there was variance in the effect sizes. For example, although statistically significant, the effect size on electronic propinquity was relatively small ( $\eta^2 = .01$ ). Thus, although this result was in the same pattern as those found for other outcome variables measured, the effect size was smaller than those found for other outcomes. Perhaps one reason that this effect size was much smaller than the others was the specific person that was used for the human host video. In order to try to make sure that participants would not be familiar with the human appearing in the video, a particular person was chosen. This person was a middle-aged man, who appears and sounds middle-aged. Given the use of a middle-aged male actor, it could be expected that our human host may have prompted *out-group* responses from the relatively young sample (Cohen et al., 2019), and thus, relatively low increases in perceived closeness central to electronic propinquity, compared to other outcomes measured. Future research can be conducted to test for possibilities of different patterns of relative effect sizes using different humans (and robots, for that matter) as comparisons.

Perhaps robots are also less alien to people now than they may have been in the past. Greater familiarity with robots may bring them closer to humans in electronic propinquity. It is also possible that greater familiarity with robots here means people have developed scripts for dealing with technology (Gambino et al., 2020), and this may lead people to have somewhat similar responses to technology although those responses might be driven by different processes (Edwards & Edwards, 2022). Future research is necessary to consider these possibilities.

However, whether the social actor told participants their name or not during this introduction had no impact on participants' perceptions of said actor's humanness, electronic propinquity with the actor, social and task attraction toward the actor, or credibility of the

actor. This was surprising, as previous research provided reasons to assume that naming can be a cue that individuates an actor (e.g., Darling, 2017; Fritz, 2018), which would make the actor more social and more like the participant. Interestingly, in some of the past studies showing that name and an impact on anthropomorphism and other outcomes, name was manipulated along with other anthropomorphic cues. For example, Waytz et al. (2014) used three different car conditions in their study: A *normal* one, where people drove a car themselves, an *agentive* one, where the vehicle was able to control steering and speed, and an *anthropomorphic* condition that added a name, gender, and human voice to the car. Similarly, Araujo (2018) differentiated an anthropomorphic agent from a non-anthropomorphic agent by giving the anthropomorphic agent a human name (instead of a nonhuman one), as well as having the agent interact using less formal language and asking the participant to use more human dialogical cues to start and end the interaction. These studies did not explicitly test which of these individual cues would lead to anthropomorphism specifically, but the current research seems to suggest that name alone may not always be enough of an anthropomorphic cue to increase perceived humanness of a machine. Perhaps naming operates as what Lombard and Xu (2021) refer to as a secondary social cue in this situation; one that is neither sufficient nor necessary to lead to the social outcomes examined in the context of the current research. Future research can examine this possibility.

Perhaps an explanation for this pattern of findings is this: The visual/nonverbal information that clearly showed the social actor to be human or robot was such a strong initial piece of information about the social actor that it overrode any initial perception that the naming could have caused, especially given the way that name was manipulated in the current study. Perhaps providing more reminders of the host's name (e.g., visually representing it on the screen as well as having the host say it) would make it more salient even with the seemingly stronger attention paid to the ontological frame. It is also possible that the presence of a name (or not) would have been a more important piece of information if participants were led to believe that further and actual interaction with the social actor was going to take place. For example, anticipation of future interaction has been found to matter in research on social information processing theory (SIPT; Walther, 1992), such that such anticipation may be important for people to be willing and able to pay attention to information like this (Kellerman & Reynolds, 1990; Walther, 1994). In other words, people need motivation to do things that help form impressions of other social actors, but can do so with such motivation. Perhaps the ontological frame was too great a cue to ignore, but participants felt no particular need to attend to something like a name here without a motivation such as anticipating future interaction, and host names may be more salient for participants who expect future interactions. Although this does not change the fact that the name manipulation used did not seem to matter in the static, initial impression environment of the current study, it may help explain why not, and why we may still expect naming to matter in future studies (and other studies that did show the importance of name). This is something that future research can examine.

Given this possibility, it is also possible that moving past initial impressions and actually interacting with the social actor may make the naming a more important piece of information. Again, Walther's (1992) SIPT suggests that impressions can be formed through interaction in CMC, and has been argued to be applicable to the study of human-machine communication (HMC; Westerman et al., 2020). This has also been seen in previous studies

on the human-human interaction script. Studies found that although initial impressions (based on expectations) of interacting with robots were lower than with humans (Edwards et al., 2019; Edwards et al., 2016; Spence et al., 2014); however, after actually interacting for 5 minutes, these differences largely disappear, and may even turn to more positive impressions with robots (Edwards et al., 2019). Furthermore, Gockley et al. (2005) found that a robot serving as a receptionist was able to build relationships with recurring visitors, so such interaction and relationship building processes have shown to be possible in a similar setting. Perhaps providing a name could kickstart later actual interactions, as other information may (Westerman et al., 2008), changing the nature of the interaction itself, and leading to predicted differences in the types of perceptual outcomes measured in the current study. Again, future research could examine this in a situation involving actual interactions with various social actors (e.g., humans and robots) that provide names or not.

Considering interactions with social actors, it is also possible that how a social actor's name is used matters in potential interactions as well. Fritz (2018) suggested that the real power of naming robots is not only in the name itself, but in the fact that the person interacting with said robot has to address them by that name (engaging human, or even pet, scripts). The robot then also responds to that name, either verbally or nonverbally, such as by turning to look at the person who has addressed them. If this is the case, then actually having an interaction with a robot would be important for seeing the kinds of differences that were predicted due to naming in this study. It is also possible that hearing another person use the robot's name in addressing the robot (rather than addressing it oneself) would also work to humanize the robot more and perhaps lead to other human perceptions, as expected. For example, Darling (2017) had other people use a robot's name when addressing it. Future research can examine the possibility of activating human interaction scripts and processes based on these kinds of hailing of a robot by name within an interaction, even when the initial use of the name is a relatively small cue.

## Author Biographies

**David Westerman** (PhD, Michigan State University) is an Associate Professor in the Department of Communication at North Dakota State University. His research and classes focus on how people communicate through and with technology, especially focusing on how we perceive social actors of both the human and machine variety. He is the Director of the Department of Communication's Social Robotics Lab at NDSU, and affiliated with the Communication and Social Robotics Lab ([www.combotlabs.org](http://www.combotlabs.org)).

 <https://orcid.org/0000-0001-9550-0304>

**Michael Vosburg** (PhD, North Dakota State University) is an Assistant Professor of Mass Communication at Benedict College. He had a 33-year career in photojournalism, and managed visual departments for 25 years at *The Missourian* (Columbia), *The San Angelo (TX) Standard-Times*, and *The Forum of Fargo-Moorhead* (ND-MN). He champions using mobile devices to speed the delivery of high-quality photographs to readers in minutes. His primary research interest is media effects of photographs, a subject he also theorizes.

 <https://orcid.org/0000-0001-8613-7670>

---

**Xinyue “Gordon” Liu** (MA, Boston University) is a Doctoral Candidate in the Department of Communication at North Dakota State University, where he also currently serves as the Assistant Basic Course Director. His primary research focus centers on how technology can be used in the classroom, especially to help improve performance in the basic course.

**Patric R. Spence** (PhD, Wayne State University) is a Professor at University of Central Florida. His primary areas of research are crisis communication and social robotics. He is affiliated with the Communication and Social Robotics Labs ([www.combotlabs.org](http://www.combotlabs.org)).

 <https://orcid.org/0000-0002-1793-6871>

## Center for Open Science



This article has earned the Center for Open Science badges for Open Materials through Open Practices Disclosure. The authors have made their data and materials freely accessible at <https://osf.io/y62ak/>.

## References

- Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior, 85*, 183–189. <https://doi.org/10.1016/j.chb.2018.03.051>
- Banks, J., Edwards, A. P., & Westerman, D. (2021). The space between: Nature and machine heuristics in evaluations of organisms, cyborgs, and robots. *Cyberpsychology, Behavior, and Social Networking, 24*(5), 324–331. <https://doi.org/10.1089/cyber.2020.0165>
- Bartneck, C., Kulic, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics, 1*(1), 71–81. <https://doi.org/10.1007/s12369-008-0001-3>
- Beattie, A., Edwards, A. P., & Edwards, C. (2020). A bot and a smile: Interpersonal impressions of chatbots and humans using emoji in computer-mediated communication. *Communication Studies, 71*(3), 409–427. <https://doi.org/10.1080/10510974.2020.1725082>
- Burgoon, J. K. (1993). Interpersonal expectations, expectancy violations, and emotional communication. *Journal of Language and Social Psychology, 12*(1–2), 30–48. <https://doi.org/10.1177/0261927X93121003>
- Cohen, J., Appel, M., & Slater, M. D. (2019). Media, identity, and the self. In M. B. Oliver, A. A. Raney, & J. Bryant (Eds.), *Media effects: Advances in theory and research* (4th ed., pp. 179–194). Routledge.

- Craig, M. J. A., & Edwards, C. (2021). Feeling for our robot overlords: Perceptions of emotionally expressive social robots in initial interactions. *Communication Studies*, 72(2), 251–265. <https://doi.org/10.1080/10510974.2021.1880457>
- Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2007). The BIAS map: Behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology*, 92(4), 631–648. <https://doi.org/10.1037/0022-3514.92.4.631>
- Darling, K. (2017). “Who’s Johnny?”: Anthropomorphic framing in human-robot interaction, integration, and policy. In P. Lin, R. Jenkins, & K. Abney (Eds.), *Robot ethics 2.0: From autonomous cars to artificial intelligence* (pp. 173–192). Oxford University Press.
- Dautenhahn, K. (2004). Socially intelligent agents in human primate culture. In S. Payr & R. Trapp (Eds.), *Agent culture: Human-agent interaction in a multicultural world* (pp. 45–71). Lawrence Erlbaum Associates, Inc. <https://doi.org/10.1201/b12476>
- Edwards, A. (2018). Animals, humans, and machines: Interactive implications of ontological classification. In A. Guzman (Ed.), *Human-machine communication: Rethinking communication, technology, and ourselves* (pp. 29–49). Peter Lang. <https://doi.org/10.3726/b14399>
- Edwards, A., & Edwards, C. (2022). Does the correspondence bias apply to social robots?: Dispositional and situational attributions of human versus robot behavior. *Frontiers in Robotics and AI*, 8. <https://doi.org/10.3389/frobt.2021.788242>
- Edwards, A., Edwards, C., Westerman, D., & Spence, P. R. (2019). Initial expectations, interactions, and beyond with social robots. *Computers in Human Behavior*, 90, 308–314. <https://doi.org/10.1016/j.chb.2018.08.042>
- Edwards, C., Edwards, A., Albrehi, F., & Spence, P. (2021). Interpersonal impressions of a social robot versus human in the context of performance evaluations. *Communication Education*, 70(2), 165–182. <https://doi.org/10.1080/03634523.2020.1802495>
- Edwards, C., Edwards, A., Spence, P. R., & Westerman, D. (2016). Initial interaction expectations with robots: Testing the human-to-human interaction script. *Communication Studies*, 67(2), 227–238. <https://doi.org/10.1080/10510974.2015.1121899>
- Etzrodt, K., & Engesser, S. (2021). Voice-based agents as personified things: Assimilation and accommodation as equilibration of doubt. *Human-Machine Communication*, 2, 57–79. <https://doi.org/10.30658/hmc.2.3>
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G\*Power 3: A flexible statistical power analysis program for social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175–191. <https://doi.org/10.3758/BF03193146>
- Finkel, M., & Krämer, N. C. (2022). Humanoid robots—artificial. Human-like. Credible? Empirical comparisons of source credibility attributions between humans, humanoid robots, and non-human-like devices. *International Journal of Social Robotics*, 14, 1397–1411. (2022). <https://doi.org/10.1007/s12369-022-00879-w>
- Fraune, M. R., Kawakami, S., Sabanovic, S., De Silva, R., & Okada, M. (2015). Three’s company, or a crowd?: The effects of robot number and behavior on HRI in Japan and the USA. *Proceedings of the international conference on robotics science and system*. <https://doi.org/10.15607/RSS.2015.XI.033>
- Fritz, L. M. (2018). Child or product? The rhetoric of social robots. In A. L. Guzman (Ed.), *Human-machine communication. Rethinking communication, technology, and ourselves* (pp. 6–82). Peter Lang.
-



- Gambino, A., Fox, J., & Ratan, Rabindra, R. A. (2020). Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication, 1*, 71–86. <https://doi.org/10.30658/hmc.1.5>
- Gehl, R. W., & Bakardjieva, M. (2017). Socialbots and their friends. In R. W. Gehl & M. Bakardjieva (Eds.), *Socialbots and their friends: Digital media and the automation of sociality* (pp. 1–16). Routledge.
- Gockley, R., Bruce, A., Forlizzi, J., Michalowski, M., Mundell, A., Rosenthal, S., Sellner, B., Simmons, R., Snipes, K., Schultz, A. C., & Wang, J. (2005, August). Designing robots for long-term social interaction. In Proceedings of the 2005 IEEE/RSJ international conference on intelligent robots and systems (pp. 1338–1342). IEEE.
- Graf, J., Erba, J., & Harn, R. W. (2017). The role of civility and anonymity on perceptions of online comments. *Mass Communication and Society, 20*(4), 526–549. <https://doi.org/10.1080/15205436.2016.1274763>
- Gunkel, D. J. (2020). *An introduction to communication and artificial intelligence*. Polity.
- Guzman, A. (2020). Ontological boundaries between humans and computers and the implications for human-machine communication. *Human-Machine Communication, 1*, 37–54. <https://doi.org/10.30658/hmc.1.3>
- Kellerman, K. (1992). Communication: Inherently strategic and primarily automatic. *Communication Monographs, 59*(3), 288–300. <https://doi.org/10.1080/03637759209376270>
- Kellerman, K., & Reynolds, R. (1990). When ignorance is bliss: The role of motivation to reduce uncertainty in uncertainty reduction theory. *Human Communication Research, 17*(1), 5–75. <https://doi.org/10.1111/j.1468-2958.1990.tb00226.x>
- Kelly, S. E., & Westerman, D. K. (2016). New technologies and distributed learning systems. In P. L. Witt (Ed.), *Handbooks of communication science 16: Communication and learning* (pp. 455–479). De Gruyter.
- Kim, J., Xu, K., & Merrill, Jr., K. (2022). Man vs. machine: Human responses to an AI newscaster and the role of social presence. *The Social Science Journal. https://doi.org/10.1080/03623319.2022.2027163*
- Korzenny, F. (1978). A theory of electronic propinquity: Mediated communication in organizations. *Communication Research, 5*(1), 3–24. <https://doi.org/10.1177/009365027800500101>
- Lee, T. L., & Fiske, S. T. (2006). Not an outgroup, not yet an ingroup: Immigrants in the stereotype content model. *International Journal of Intercultural Relations, 30*(6), 751–768. <https://doi.org/10.1016/j.ijintrel.2006.06.005>
- Lin, X., Kaufmann, R., Spence, P. R., & Lachlan, K. A. (2019). Agency cues in online comments: Exploring their relationship with anonymity and frequency of helpful posts. *Southern Communication Journal, 84*(3), 183–195. <https://doi.org/10.1080/1041794X.2019.1584828>
- Lin, X., Spence, P. R., & Lachlan, K. A. (2016). Social media and credibility indicators: The effect of influence cues. *Computers in Human Behavior, 63*, 264–271. <https://doi.org/10.1016/j.chb.2016.05.002>
- Lombard, M., & Xu, K. (2021). Social responses to media technologies in the 21st century: The media are social actors paradigm. *Human-Machine Communication, 2*, 29–55. <https://doi.org/10.30658/hmc.2.2>
-



- McCroskey, J. C., & McCain, T. A. (1974). The measurement of interpersonal attraction. *Speech Monographs*, 41(3), 261–266. <https://doi.org/10.1080/03637757409375845>
- McCroskey, J. C., & Teven, J. J. (1999). Goodwill: A reexamination of the construct and its measurement. *Communication Monographs*, 66(1), 90–103. <https://doi.org/10.1080/03637759909376464>
- Michaelis J., & Mutlu, B. (2019). Supporting interest in science learning with a social robot. *Proceedings of the 18th ACM international conference on interaction design and children* (pp. 71–82). ACM. <https://doi.org/10.1145/3311927.3323154>
- Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1), 81–103. <https://doi.org/10.1111/0022-4537.00153>
- Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *CHI'94: Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 72–78). ACM Digital Library. <https://doi.org/10.1145/191666.191703>
- Palsson, G. (2014). Personal names: Embodiment, differentiation, exclusion, and belonging. *Science, Technology, & Human Values*, 39(4), 618–630. <https://doi.org/10.1177/0162243913516808>
- Rainear, A. M., Jin, X., Edwards, A., Edwards, C., & Spence, P. R. (2021). A robot, meteorologist, and amateur forecaster walk into a bar: Examining qualitative responses to a weather forecast delivered via social robot. *Communication Studies*, 72(6), 1129–1145. <https://doi.org/10.1080/10510974.2021.2011361>
- Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.
- Spence, P. R. (2019). Searching for questions, original thoughts, or advancing theory: Human-machine communication. *Computers in Human Behavior*, 90, 285–287. <https://doi.org/10.1016/j.chb.2018.09.014>
- Spence, P. R., Edwards, A., Edwards, C., & Jin, X. (2019). ‘The bot predicted rain, grab an umbrella’: Few perceived differences in communication quality of a weather Twitterbot versus professional and amateur meteorologists. *Behaviour & Information Technology*, 38(1), 101–109. <https://doi.org/10.1080/0144929X.2018.1514425>
- Spence, P. R., Edwards, C., Edwards, A., Rainear, A., & Jin, X. (2021). “They’re always wrong anyway”: Exploring differences of credibility, attraction, and behavioral intentions in professional, amateur, and robotic-delivered weather forecasts. *Communication Quarterly*, 69(1), 67–86. <https://doi.org/10.1080/01463373.2021.1877164>
- Spence, P. R., Westerman, D., Edwards, C., & Edwards, A. (2014). Welcoming our robot overlords: Initial expectations about interaction with a robot. *Communication Research Reports*, 31(3), 272–280. <https://doi.org/10.1080/08824096.2014.924337>
- Spence, P. R., Westerman, D., & Luo, Z. (2023). Observing communication with machines. In A. Guzman, R. McEwen, & S. Jones (Eds.), *The Sage handbook of human machine communication* (pp. 220–227). Sage Publications. <http://doi.org/10.4135/9781529782783.n27>
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 49(236), 433–460. <https://doi.org/10.1093/mind.LIX.236.433>
- Walther, J. B. (1992). Interpersonal effects in computer-mediated interaction: A relational perspective. *Communication Research*, 19(1), 52–90. <https://doi.org/10.1177/009365092019001003>
-

- 
- Walther, J. B. (1994). Anticipated ongoing interaction versus channel effects on relational communication in computer-mediated interaction. *Human Communication Research*, 20(4), 473–501. <https://doi.org/10.1111/j.1468-2958.1994.tb00332.x>
- Walther, J. B., & Bazarova, N. N. (2008). Validation and application of electronic propinquity theory to computer-mediated communication in groups. *Communication Research*, 35(5), 622–645. <https://doi.org/10.1177/0093650208321783>
- Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, 52, 113–117. <http://dx.doi.org/10.1016/j.jesp.2014.01.005>
- Westerman, D., Edwards, A. P., Edwards, C., Luo, Z., & Spence, P. (2020). I-It, I-Thou, I-Robot: The perceived humanness of AI in human-machine communication. *Communication Studies*, 71(3), 393–408. <https://doi.org/10.1080/10510974.2020.1749683>
- Westerman, D., Van Der Heide, B., Klein, K. A., & Walther, J. B. (2008). How do people really seek information about others?: Information seeking across internet and traditional communication channels. *Journal of Computer-Mediated Communication*, 13(3), 751–767. <https://doi.org/10.1111/j.1083-6101.2008.00418.x>
- Young, R. K., Kennedy, A. H., Newhouse, A., Browne, P., & Thiessen, D. (1993). The effects of names on perceptions of intelligence, popularity, and competence. *Journal of Applied Social Psychology*, 23(21), 1770–1788. <https://doi.org/10.1111/j.1559-1816.1993.tb01065.x>
-



# Authentic Impediments: The Influence of Identity Threat, Cultivated Perceptions, and Personality on Robophobia

Kate K. Mays<sup>1</sup> 


<sup>1</sup> Department of Community Development and Applied Economics, College of Agriculture and Life Sciences, University of Vermont, Burlington, Vermont, USA

## Abstract

Considering possible impediments to authentic interactions with machines, this study explores contributors to *robophobia* from the potential dual influence of technological features and individual traits. Through a  $2 \times 2 \times 3$  online experiment, a robot's physical human-likeness, gender, and status were manipulated and individual differences in robot beliefs and personality traits were measured. The effects of robot traits on phobia were nonsignificant. Overall, subjective beliefs about what robots are, cultivated by media portrayals, whether they threaten human identity, are moral, and have agency were the strongest predictors of robophobia. Those with higher internal locus of control and neuroticism, and lower perceived technology competence, showed more robophobia. Implications for the sociotechnical aspects of robots' integration in work and society are discussed.

**Keywords:** robophobia, social robots, artificial intelligence, agency, identity

**Acknowledgments:** This study was supported by the Division of Emerging Media Studies at Boston University through a Feld Research Grant. The author would like to express her appreciation to James Cummings for his invaluable and careful feedback on the study design.

**CONTACT** Kate K. Mays  • [kate.mays@uvm.edu](mailto:kate.mays@uvm.edu) • 146 University Place • Department of Community Development and Applied Economics • University of Vermont • Burlington, VT 05405-0160, USA

ISSN 2638-602X (print)/ISSN 2638-6038 (online)  
[www.hmcjournal.com](http://www.hmcjournal.com)



Copyright 2024 Authors. Published under a Creative Commons Attribution 4.0 International (CC BY-NC-ND 4.0) license.

## Introduction

Social robots come in many shapes and sizes, variably approximating human appearance. Some, like Hanson Robotics' Sophia, attempt to appear as human-like as the technology allows, with artificial skin, female features, and feminine outfits that strategically hide Sophia's wires and rolling base. Others, like SoftBank's Pepper, maintain a mechanistic look, with an all-white plastic exterior and a touch screen for a chest. These design choices suggest differing ontological aims for social robots: Sophia imitates humans as closely as possible in order to facilitate a more seamless "co-existence" with people, whereas Pepper stands apart in an immutable "robot" category.

The efficacy of either approach may also be variable and highly contingent on individual differences in expectations and beliefs about robots (A. Edwards et al., 2019), as well as beliefs about human identity (Ferrari et al., 2016; Zlotowski et al., 2017). One of the value propositions that social robots offer is real engagement with their human interactants, such that they could fill in for their human counterparts in retail, care, and education spaces, as a few examples (Pedersen et al., 2018; Rasouli et al., 2022). To achieve this, people should feel like they are experiencing an authentic interaction. What "authenticity" means, though, can vary; the term has been used to denote originality, indicate a veritable reconstruction or reproduction, and describe the revelation of a deep truth (Van Leeuwen, 2001). Because of this conceptual fuzziness, Van Leeuwen (2001) emphasizes the situated, relative, and subjective nature of authenticity, as a question not of external reality but of who perceives something as authentic, and who does not.

Within the computer-mediated communication (CMC) paradigm, authenticity is emphasized in the ways people perform their identities on platforms (Abidin, 2018; Van Driel & Dumitrica, 2021). Authenticity in mass-CMC relates to the source, message, and interaction that influence beliefs in, feelings about, and behavior change from media messages (Lee, 2020). Within the human-machine communication (HMC) paradigm, questions shift from authenticity *through* a mediated channel (toward another human receiver) to authentic interactions and relations *with* a machine (Guzman & Lewis, 2020). Here, questions arise not only about the human's identity but also about the identity of the machine alone and in relation to its human interactant. For example, engaging with social robots as interaction partners may hinge on both the social robot's plausibility as a human-like interactant and the human interactant's receptivity to such engagement.

Therefore, this study explores perceptions of social robots from the potential dual influence of technological features and individual traits. People's phobia of robots (*robophobia*) is examined and considered conceptually as a potential hindrance to meaningful, authentic interactions. Using an online experimental design, this study analyzes whether a robot's physical human-like appearance, gender, and status affect people's robophobia, and the extent to which people's cultivated perceptions of robots from media, attitudes of robot's threat to human uniqueness, and individual differences in efficacy and anxiety influence these attitudes.

---

## Literature Review

### Robophobia

Phobia around technology has been narrowly conceptualized as fear and anxiety toward computers and more broadly conceived to capture people's orientation to technology generally (Khasawneh, 2018a; Osiceanu, 2015). A commonality across definitions is that such phobia is characterized by avoidance, paranoia, fear, and anxiety, which can manifest behaviorally, emotionally, and attitudinally (Osiceanu, 2015). In turn, technophobia is an important factor in people's adaptation to new technologies (Khasawneh, 2018a, 2018b; Lan et al., 2022). Those with computerphobia have more negative attitudes toward computers (Rosen et al., 1993), which in turn leads to computer avoidance (McIlroy et al., 2007). From the lens of technology acceptance (Davis, 1989), technophobia is a significant antecedent to attitudes about how easy and useful a technology is (Khasawneh, 2018b).

In thinking about robophobia, there are similarities with and deviations from computer- and technophobia. To start, computers are a tool employed by people to help them achieve their own goals. While the introduction of computers and their ancillary systems in the workplace required employee re-skilling and upskilling, and rendered certain tasks obsolete, computers still required human operators. Robots, though also conveyed as tools and helpers, can act with varying degrees of autonomy. With less need for direct human intervention or involvement, robots pose more existential threat than computers, which lack robots' increasingly autonomous, intelligent, and embodied capabilities (Sinha et al., 2020).

Research has shown that similar technophobic dynamics to computer resistance is at play with robots. People with more negative attitudes toward robots are more likely to avoid human-robot communication (Nomura et al., 2008). Technophobia had a powerful and negative influence on intentions to use robots in a hospitality context (Sinha et al., 2020). Importantly, technophobia not only negatively predicted use intentions, but also usurped anthropomorphism's positive effect on use intention (Sinha et al., 2020). This suggests the importance of considering differences in how individuals approach technology alongside its features.

Indeed, for decades, we have seen evidence that socio-emotional relating with machines may have less to do with its technical capabilities and more to do with the human interactant (Vanman & Kappas, 2019). Rudimentary computer programs like ELIZA (Weizenbaum, 1966) and the Tamogotchi (Vanman & Kappas, 2019) could elicit human emotion and attachment, which Turkle (2007) explicated by people's projection of their own attributions and desires, in order to bridge the gap between an artifact's actual (rudimentary) capabilities and people's (complex) emotions. More recently, though, these "relational artifacts" (Turkle, 2007) imitate human behavior and appearance in increasingly sophisticated ways, as illustrated by robots like Sophia and Pepper.

Thus, robophobia may be variably influenced by the technology's traits and differences across people in how they approach technology. An important question is the extent to which robophobia stems from its static, human-like features or people's individual experiences and subjective beliefs, which are multifaceted. The remaining literature review discusses each of these components in turn.

---



## Robots' Features

### *Physical Human-Likeness and the Uncanny Valley*

Considerations about the possible influence that a robot's human-like appearance has on attitudes toward it extends back decades to Mori et al.'s (2012) uncanny valley hypothesis. Mori posited that people feel more affinity toward nonhuman entities that appear more human-like up to a certain point of humanness; once something approaches human-likeness but is not actually human, people drop into the "uncanny valley," wherein affinity is replaced with feelings of eeriness and unease (Mori, 1970, in Mori et al., 2012; Wang et al., 2015). Importantly, Mori (1970, in Mori et al., 2012) did not test this hypothesis empirically, and subsequent research has not unequivocally demonstrated a clear, curvilinear relationship in the uncanny phenomenon (Rosenthal-von der Pütten et al., 2014). For example, MacDorman (2006) found an uncanny valley occurred in response to images of an entity morphing from mechanical to human-like, but the same pattern was not replicated with videos portraying mechanical to human-like subjects. In another study using video stimuli, Riek and colleagues (2009) found that people empathized more with robots that appeared more human when they were being mistreated.

On the other hand, studies have found that, when faced with more human-like robots, people can feel increased unease (Palomäki et al., 2018) and more threat to their identity (Ferrari et al., 2016; Yogeewaran et al., 2016). This study does not aim to directly test the uncanny valley hypothesis, which would require a greater range of stimuli than the present manipulation entails (MacDorman, 2006; Palomäki et al., 2018; Rosenthal-von der Pütten et al., 2014). The "uncanny phenomenon" (Wang et al., 2015), however, does inform how people might respond to a robot that appears mechanical compared to one that is more human-like, and supports the prediction that:

**H1:** The more human-like robot will elicit more robophobia.

### *Gender and Stereotypes*

The research on how robot gender affects people's response to it does not show a clear-cut preference for one gender over another. Studies have shown that people tend to apply existing gender stereotypes to robots (Bernotat et al., 2021; Eyssel & Hegel, 2012). When not explicitly gendered, people tend to default to a male attribution (Beraldo et al., 2018; Bernotat et al., 2021). Stereotypes can also influence robot acceptance and anthropomorphism, in that both increased when robot gender was more congruent with the task at hand (Kuchenbrandt et al., 2014; Tay et al., 2014).

In terms of more phobia-adjacent measures such as likability and trustworthiness, the results are mixed. Although they are liked more than male robots, female robots are viewed as less trustworthy (Kraus et al., 2018). Male robots are also perceived as more useful than (Beraldo et al., 2018) and generally favored (Jung et al., 2016) over female robots. Still other studies have not found any evidence of gender differences in how much people perceived competence in (Bryant et al., 2020), felt comfortable with (Rogers et al., 2020), or trusted (Ghazali et al., 2018) robots. Given these mixed findings, this study asks:

**RQ1:** Are there differences in how much robophobia is elicited by a male vs. female robot?

---

## **Status and Power**

In addition to robots' physical human traits, the human *context* in which they operate may affect how people perceive and interact with them, which is reflected by recent research in this realm (e.g., Bernotat et al., 2021; Bryant et al., 2020; Kraus et al., 2018; Rogers et al., 2020). Context could refer to the domain in which the robot operates, such as security or care settings (Tay et al., 2014; Taipale & Fortunati, 2018), as well as to the robot's status relative to its human interactants (Y. Kim & Mutlu, 2014). This study focuses on status in order to explore how a robot's agency may influence phobia of it. Research shows that generally people prefer for a robot to engage in work that is more rote and assistive (Dautenhahn et al., 2005; Takayama et al., 2008). When relying on a robot to complete a task, people are more critical of one in a supervisory compared to subordinate capacity (Hinds et al., 2004). Interestingly, when examining both physical (near vs. far) and power (high vs. low status) distance, Y. Kim and Mutlu (2014) found that people preferred the higher-status robot to remain physically closer than the lower-status robot, perhaps due to a wariness about the robot with more power. Robots demonstrating more autonomy also elicit less empathy (Kwak et al., 2013) and more feelings of eeriness (Appel et al., 2020). These findings suggest that people may be more phobic of robots with a higher status (e.g., supervisor) than them:

**H2:** A higher-status robot will elicit more robophobia than an equivalent- or lower-status robot.

## **Humans' Features**

### ***Perceptions of Robots' Identity Threat and Morality***

When robots appear more anthropomorphic (Ferrari et al., 2016) or autonomous (Zlotowski et al., 2017), they are perceived as more threatening. Threat perceptions may not just stem from robots' traits, however. If viewed as a separate ontological entity, people may categorically classify robots as "other" (A. Edwards, 2018; Vanman & Kappas, 2019). According to intergroup threat theory (Stephan et al., 2008), outgroup members are perceived to pose heightened threat, which leads to ingroup members holding more negative attitudes toward them (Stephan et al., 2008; Zlotowski et al., 2017). Outgroup bias is caused by ingroup members' fear and uncertainty toward unfamiliar "others" (Kawakami et al., 2017). This dynamic has been demonstrated in threat perceptions of machines, which amplify negative attitudes about usage (Huang et al., 2021). People may differ in how much they view robots as outgroup members, which would influence the extent to which they perceive them as threatening (Vanman & Kappas, 2019; Yogeeswaran et al., 2016). Therefore, this study predicts that:

**H3:** Perceived identity threat is related to greater robophobia.

Although robots can elicit feelings of threat, they can also be regarded as entities deserving of moral treatment (Banks, 2019; Waytz et al., 2010). Banks (2019, 2021) has identified two dimensions of robots' morality: their ability to reason (morality dimension) and the extent to which they lack agency and intentionality (dependency dimension). In her validation of the scale, Banks (2019) found that robots' perceived morality was related

to positive feelings about the robots' goodwill and trustworthiness, as well as willingness to interact more intimately with it and have more relational certainty toward it. Examining moral behaviors, Banks (2021) found that judgments are relatively agent agnostic, though the robot agent (compared to the human agent) was given more credit or blame for upholding or violating moral foundations. This (small) interaction effect suggests that heuristics about a robot's mind or morality may influence judgments about their (im)moral behavior (Banks, 2021).

Viewing a robot with empathy extends from individual differences in anthropomorphic tendencies (Darling, 2015), which are also related to the extent to which robots are seen as entities with moral worth (Waytz et al., 2010). When presented as more autonomous (Stein & Ohler, 2017) or more human-like (Ceh & Vanman, 2018), robots simultaneously elicited more empathy *and* more feelings of threat. Thus, when innate human traits are ascribed to robots, they may activate both affinity and hostility, making it unclear whether moral perceptions of a robot would influence negative attitudes toward it. Seeing robots as moral accords with more affinity toward it (Banks, 2019), but a unique human trait could also elicit feelings of animus (Vanman & Kappas, 2019). Therefore, this study explores whether perceived morality affects robophobia.

**RQ2a-b:** Is a robot's perceived (a) morality and (b) dependency related to robophobia?

### ***Robot Experience in Real Life and on the Screen***

The literature on technophobia demonstrates how increased exposure to and experience with a technology can reduce people's apprehension about it (Anthony et al., 2000). Similarly, affinity toward robots may be developed with increased real-life interactions and experience with them (Lan et al., 2022; Nomura & Horii, 2020). When exposed to a robot in their classroom over 2 months, elementary school children came to view it as a member of their group (Kanda et al., 2007). Importantly, though, this dynamic occurred among children who were initially open to interacting with it; some children in the classroom rejected its presence early on (Kanda et al., 2007). Thus, real-life experience with a robot may already hinge on a lack of robophobia, which may have a self-reinforcing effect in that further contact reduces phobia more. Therefore, this study posits that:

**H4:** More real-life experience with robots relates to less robophobia.

In the absence of real-life experience, people may rely on media portrayals to frame their understanding. Media exposure cultivates certain attitudes toward (Sundar et al., 2016) or mental models (Banks, 2020) of robots. When they could better recall robots from films, people showed less anxiety about robots generally (Sundar et al., 2016). When people felt sympathy toward recalled robot characters, they were more likely to view robots positively (Banks, 2020). Conversely, when people had cultivated negative perceptions of robots from media exposure they subsequently held more negative attitudes (Horstmann & Krämer, 2019). Given these differential effects of positive and negative views, this study captures them separately and predicts that:

**H5a:** Positive mediated view of robots relates to less robophobia.

**H5b:** Negative mediated view of robots relates to greater robophobia.

### **Personality Traits**

People's attitudes about technology are not solely determined by their prior experience with it (Anthony et al., 2000). Matthews and colleagues (2021) argue that individual differences in etic (i.e., universal, generalizable) traits are critical for understanding human-machine interactions, now and in the future. Given the uncertain, increasingly complex, and rapidly advancing nature of intelligent and autonomous technology, people's acceptance cannot necessarily hinge on sophisticated knowledge about its use (Matthews et al., 2021). Therefore, in addition to robot-specific experience and beliefs (what Matthews et al., 2021 refer to as "emic" traits), individuals' traits related to efficacy and personality are explored. As an interactive, agentic technology, social robots are a departure from prior conceptions of "use" for technical tools; thus, people's own sense of agency and control may be challenged in the face of machine agency (Mays et al., 2021). Research on the influence of efficacy in technology adoption typically finds that general efficacy and domain efficacy positively relate to adoption (Hsia et al., 2014). In attitudes toward AI, however, people with a greater sense of control of their lives were *less* comfortable with the technology (Mays et al., 2021). Conversely, those with more technological competence (domain efficacy) were more comfortable with AI. As a technology with similar attributes to AI (e.g., more autonomy and agency), attitudes toward robots may show a similar divergence in influence of general and domain efficacy. Therefore, this study predicts that:

**H6:** Higher internal locus of control is related to greater robophobia.

**H7:** Higher perceived technology competence is related to less robophobia.

Of the Big Five personality traits, neuroticism in particular—which is characterized by tendencies toward anxiety and emotional instability (Eysenck et al., 1985)—shows a positive relationship with technophobia (Anthony et al., 2000) and computer anxiety (Osiceanu, 2015), as well as fear of and less comfort with AI (Mays et al., 2021; Sindermann et al., 2022). This pattern appears to extend to robots, as those higher in neuroticism are less comfortable with them (Robert, 2018), hold more negative attitudes toward them (Müller & Richert, 2018), and are more sensitive to their uncanniness (eeriness and lack of warmth) (MacDorman & Entezari, 2015). While this study does not evaluate uncanniness directly, robophobia and the uncanny are conceptually similar in that both relate to fear and anxiety (MacDorman & Entezari, 2015). Neuroticism is a particularly salient trait to examine because of its relationship to uncertainty intolerance (Matthews et al., 2021). As an emergent technology with plenty of unknowns about their advancement and social integration, social robots induce a great deal of uncertainty. Additionally, those higher in neuroticism experience more sensitivity to social threat (Matthews et al., 2021). Research on attitudes toward outgroups suggests that if robots are perceived as more threatening, then people

will feel more anxiety and negativity toward them (Riek et al., 2006, in Vanman & Kappas, 2019). Given these findings, it is predicted that:

**H8:** Higher neuroticism is related to greater robophobia.

## Method

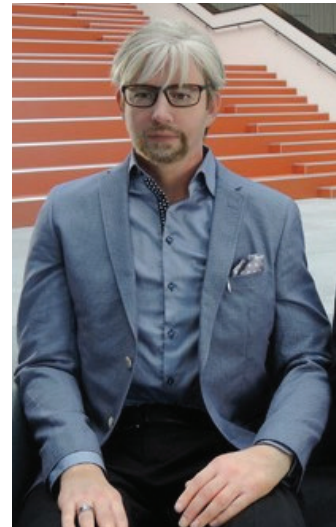
### *Design and Participants*

In order to examine the factors that influence robophobia, a between-subjects online experiment ( $2 \times 2 \times 3$ ) was conducted. Human-like robot traits were considered through a visual + vignette manipulation. Participants ( $N = 1,020$ )<sup>1</sup> were randomly shown one robot that was either male or female, either humanoid (more mechanical appearing) or android (more human appearing), and described as an agentic, intelligent entity (per Zlotowski et al., 2017) that was an assistant, coworker, or supervisor.

Age and gender quotas based on US census demographics were established for each condition. In the overall sample, 52.5% of the participants were female and the mean age was 44.01 years ( $SD = 17.30$ ). After being shown the stimulus—robot image (see Figures 1–4) and description (see Table 1)—participants were instructed to imagine the robot in the scenario when responding to a measure of robophobia. After completing that measure, participants answered other self-report measures for the independent variables.



**FIGURE 1** Android Female Robot, “Nadine”



**FIGURE 2** Android Male Robot, “Geminoid”

1. Sample size was determined based on available funding and an estimation of 100 participants/condition. The resulting sample size ( $N = 1020$ ) is smaller after removing straight-liners from the data. Using G\*Power software, a post-hoc power analysis was conducted for multiple linear regression with 14 predictors, an alpha of .05, and a conservative effect size ( $f^2 = .02$ ), yielding a statistical power of .85 (Faul et al., 2009).



FIGURE 3 Humanoid Female Robot, "Ira"



FIGURE 4 Humanoid Male Robot, "Romeo"

TABLE 1 Robot Scenarios Displayed to Manipulate Its Status

	Status description
In all three conditions	<p>Today's robots can already move on their own and perform a variety of tasks like lifting heavy things, cleaning, driving, tutoring, and looking after the elderly. They can also solve puzzles and make decisions on their own.</p> <p>In light of these advances, in the very near future robots might be part of everyday life. One setting where robots may be deployed is <b>in the workplace.</b></p>
Superior	Imagine that this robot has been assigned as your <b>supervisor at work.</b> In such a role, [she / he] would assign you tasks and projects, as well as evaluate your performance.
Peer	Imagine that this robot has been assigned as your <b>coworker at work.</b> In such a role, [she / he] would be assigned similar tasks to yours, as well as work with you as a partner on group projects.
Subordinate	Imagine that this robot has been assigned as your <b>personal assistant at work.</b> In such a role, [she / he] would help you with your tasks and projects, performing duties like answering phones and emails, scheduling meetings, and taking care of other logistics.



## Stimulus Material

Prior to the main study, a pilot study was conducted to pre-test the robot image stimuli to ensure they significantly differed in gender and human-likeness, and were not significantly different in threat perceptions. Drawing from the ABOT (Anthropomorphic roBOT; Phillips et al., 2018) database, 21 robots were identified for pre-testing based on ABOT's humanness ratings. Participants ( $N = 75$ ) were asked to rate aspects of the robot images' physical appearance using 9-point semantic differential scales adapted from MacDorman (2006), Bartneck et al. (2009), and Ho and MacDorman (2010). Physical humanness was evaluated on four pairs of items: machine-like vs. human-like; artificial vs. natural; robotic vs. human; human-made vs. human-like. The robot's gender was evaluated using two pairs of items: male vs. female; masculine vs. feminine. In order to control for any other aspects of the robot's appearance that could confound phobia perceptions, four pairs of items gauging threateningness were evaluated: cold vs. warm; threatening vs. friendly; unlikeable vs. likeable; dangerous vs. safe. Gender perception scores were used first to reduce the sample of 21 robots to 4 for further analysis. These four robots' physical humanness and threateningness scores were then compared via paired-samples  $t$ -tests. The  $t$ -test results confirmed significant gender differences within the pairs of android ( $t = 19.93, p < .001$ ) and humanoid ( $t = 7.16, p < .001$ ) robots. Between the android and humanoid pairs there were significant differences in physical humanness (android male–humanoid male:  $t = -12.30, p < .001$ , android male–humanoid female:  $t = -11.44, p < .001$ , android female–humanoid male:  $t = -13.38, p < .001$ , android female–humanoid female:  $t = -12.79, p < .001$ ). Threateningness scores were not significantly different across the four robots.

Robot status was manipulated using vignettes based on Zlotowski et al.'s (2017) scenarios (see Table 1). These included the same description of a social robot's capabilities across conditions and varied a workplace scenario to describe the robot as the participant's supervisor (superior status), coworker (peer status), or personal assistant (subordinate status). The robots (Figures 1–4) were combined with a vignette (Table 1) and presented together in one image.

## Measurement

Unless otherwise noted, all variables were measured using 7-point, Likert-type scales.

**Dependent variable.** Following the stimuli, participants were asked to respond to the *robotphobia* items: "I would feel very nervous just being around a robot," "I would feel paranoid talking with a robot," "Something bad will happen if robots develop into living beings," "I would feel very nervous just being around a robot," "I would feel uneasy if robots really had emotions," "Robots should never make decisions concerning people," and "Robots would be a bad influence on children." The 6-item scale (strongly disagree—to strongly agree) was adapted from Nomura et al.'s (2008) Negative Attitudes toward Robots Scale (NARS) ( $\alpha_{mas} = .88, \alpha_{map} = .90, \alpha_{maa} = .84, \alpha_{mhs} = .90, \alpha_{mhp} = .83, \alpha_{mha} = .78, \alpha_{fas} = .88, \alpha_{fap} = .87, \alpha_{faa} = .90, \alpha_{fhs} = .90, \alpha_{fhp} = .87, \alpha_{fha} = .89$ ).<sup>2</sup> The six NARS items were selected because they

2. Cronbach's  $\alpha$  is reported for each condition for the dependent variable: m/f = male or female, a/h = android or humanoid, and s/p/a = superior, peer, or assistant.

represented elements of other technophobia scales that gauge people's avoidance, paranoia, fear, and anxiety of the technology in question.

**Independent variables.** In addition to the robot manipulation and main outcome variable, participants' individual differences related to personal robot experience, robot-human-likeness beliefs, and personal traits were measured. Robot experience was comprised of *real-life exposure* to and *mediated views* of robots, both adapted from Horstmann and Krämer (2019). To measure exposure, participants were asked how often, on a 6-point scale ranging from "Never" to "Very often," they encountered industrial robots, domestic robots like a vacuum cleaner or lawnmower, and social robots that are autonomous and interactive ( $\alpha = .82$ ,  $M = 2.70$ ,  $SD = 1.45$ ). Mediated views were measured with two 3-item scales capturing *positive* ( $\alpha = .82$ ,  $M = 5.09$ ,  $SD = 1.16$ ) and *negative* ( $\alpha = .87$ ,  $M = 3.87$ ,  $SD = 1.51$ ) views. Participants were asked to indicate their agreement with negative (e.g., "Robots are rather against humans") and positive (e.g., "Robots help humans") statements about the relationships between humans and robots in movies or TV shows. Higher values corresponded to stronger negative and stronger positive views.

Three additional robot beliefs were measured to capture subjective impressions of robots' human-like abilities. *Perceived identity threat* (Zlotowski et al., 2017) measures the extent to which participants believe robots threaten human uniqueness. The 4-item scale asked about participants' agreement with items such as "Robots seem to lessen the value of human existence" ( $\alpha = .76$ ,  $M = 4.07$ ,  $SD = 1.51$ ). Perceptions of robots' morality were measured using Banks's (2019) two-dimensional scale that captures both *morality* (six items) and *dependency* (four items). Participants indicated their agreement with moral reasoning statements such as "Robots can have a sense for what is right and wrong" ( $\alpha = .91$ ,  $M = 4.07$ ,  $SD = 1.49$ ) and dependency statements such as "Robots can only do what humans tell them to do" ( $\alpha = .83$ ,  $M = 2.63$ ,  $SD = 1.19$ ).

Finally, personal traits of efficacy and neuroticism were measured. General efficacy was measured with a 5-item *locus of control* scale (Rotter, 1966), which asked participants' agreement to items such as "I do not have enough control over the direction my life is taking." Higher values corresponded to higher internal locus of control ( $\alpha = .84$ ,  $M = 3.68$ ,  $SD = 1.38$ ). Domain efficacy was measured through a 5-item *perceived technology competence* scale (Katz & Halpern, 2014), which captured how much participants enjoy and feel comfortable using technology. Higher values indicated more perceived competence ( $\alpha = .84$ ,  $M = 5.47$ ,  $SD = 1.19$ ). A 9-item *neuroticism* scale was adapted from Eysenck et al. (1985). Participants answered how much they agreed with statements like "I would call myself tense or 'highly strung.'" Higher values corresponded to stronger neuroticism ( $\alpha = .94$ ,  $M = 3.92$ ,  $SD = 1.51$ ).

## Results

A hierarchical linear regression was run to explore the relative influence of a robot's technological features (block 1), robot experience/beliefs (block 3), and personal traits (block 4) on robophobia. Demographics were included in the second block as a control. Table 2 displays the regression results; all analyses were conducted using IBM SPSS.

**TABLE 2 Technological and Individual Factors That Influence Robophobia**

	B (SE)	$\beta$
<b>Block 1: Robot traits</b>		
Physical humanness	.06 (.07)	.02
Gender (1 = male, 2 = female)	-.02 (.07)	-.01
Status	-.05 (.04)	-.03
$\Delta R^2$	.30%	
<b>Block 2: Demographics</b>		
Age	.001 (.002)	.02
Gender (1 = male, 2 = female)	.11 (.07)	.04
$\Delta R^2$	1.00%**	
<b>Block 3: Experience w/ robots</b>		
Real-life exposure	-.01 (.03)	-.02
Negative mediated views	.23 (.03)	.25***
Positive mediated views	-.19 (.04)	-.17***
Identity threat	.30 (.03)	.33***
Morality	-.19 (.03)	-.16***
Dependency	.17 (.03)	.19***
$\Delta R^2$	37.3%***	
<b>Block 4: Personal traits</b>		
Locus of control	.09 (.03)	.09**
Perceived technology competence	-.09 (.04)	-.07*
Neuroticism	.09 (.03)	.10**
$\Delta R^2$	2.4%***	
Total adjusted $R^2$	40.1%	

Notes:  $N = 1,020$ ; \* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$ .

The experimental manipulation of robots' features had no influence on phobic attitudes (H1–2, RQ1). Rather, individual differences in beliefs about robots had the strongest effect, explaining 37% of the variance in robophobia. Those who felt that robots threaten human identity (H3:  $\beta = .33$ ,  $p < .001$ ) and have cultivated negative views from robots' mediated portrayals (H5b:  $\beta = .23$ ,  $p < .001$ ) were more phobic. Conversely, positive mediated views of robots (H5a:  $\beta = -.17$ ,  $p < .001$ ), perceptions of robots as moral (RQ2a:  $\beta = -.19$ ,  $p < .001$ ) and agentic (e.g., lower dependency) (RQ2b:  $\beta = .16$ ,  $p < .001$ ) was related to less robophobia. Contrary to the prediction in H4, real-life exposure to robots had no effect on robophobia. Although demonstrating less influence, personal traits were also related: those who felt more in control of their lives (H6:  $\beta = .09$ ,  $p < .01$ ) and who were higher in neuroticism (H8:  $\beta = .10$ ,  $p < .01$ ) were more phobic, while those with a higher perceived technology competence were less phobic (H7:  $\beta = -.09$ ,  $p < .05$ ).

---

## Discussion

Considering possible impediments to authentic interactions with machines, this study explored contributors to robophobia. Through an online experiment, a robot's physical human-likeness, gender, and status were manipulated and individual differences in robot attitudes and traits were measured. Overall, subjective beliefs about what robots are, cultivated by media portrayals, and whether they threaten human identity, were the strongest predictors of robophobia. Stronger beliefs that robots can be moral and agentic—typically unique human traits—were related to less robophobia. Although the effects were smaller, results showed that stable individual traits (general and domain efficacy and neuroticism) also influenced robophobia, though in different directions. Those who feel more in control of their lives (general efficacy) and who are higher in neuroticism were more robophobic, while those with higher feelings of technological competency were less robophobic.

### Importance of Subjective Robot Beliefs and Individual Traits

The study's findings on the strong influence of subjective, cultivated beliefs about robots extends research on the double-sided nature of human-robot interaction that includes both the robot traits as well as individual subjectivities (c.f., MacDorman & Entezari, 2015; Mays & Cummings, 2023; Rosenthal-von der Pütten et al., 2014; Rosenthal-von der Pütten & Weiss, 2015; Waytz et al., 2010). In particular, the more robots were perceived as a threat to unique human identity, the more phobic someone was. This extends intergroup threat findings on how ingroup members perceive those in the outgroup more negatively (Stephan et al., 2008) to robots as an outgroup "other" (A. Edwards, 2018; Vanman & Kappas, 2019; Zlotowski et al., 2017), which increases negative attitudes toward them (Huang et al., 2021).

Given identity threat's amplifying influence on phobia, it is at first blush counterintuitive that perceptions of robots as moral and agentic *lessened* phobia toward them. Some research suggests that ascribing such human-like traits, particularly agency, would increase hostility toward robots (Vanman & Kappas, 2019). However, other research indicates that viewing robots as moral (Banks, 2019), autonomous (Stein & Ohler, 2017), and human-like (Ceh & Vanman, 2018) can increase affinity toward them. This study's findings on perceptions of robots as moral agents supports the latter stance. One possible explanation is that morality is not considered an exclusive human trait; thus, a robot capable of morality does not necessarily violate assumptions of unique human identity. Another explanation could be that phobia *precedes* agentic perceptions. Future work should investigate the directionality of influence, with a mediation analysis or by manipulating machine agency to explore its effects on phobia.

The significant influence of cultivated attitudes on how people engage with and perceive the world is well established and extends far beyond robots (Gerbner & Gross, 1976). In the context of robots, this study reinforces prior research findings on the extent to which media affects attitudes about robots (Banks, 2020; Horstmann & Krämer, 2019; Sundar et al., 2016): a negative mediated attitude was related to higher robophobia and a positive mediated attitude was related to less robophobia. Of note, negative cultivation had a stronger effect on phobia compared to positive cultivation, which may stem from people's negativity bias (Rozin & Royzman, 2001).

---

The findings also showed an interesting dynamic between domain and general efficacy, wherein those with technological efficacy were less phobic, while those with higher general efficacy were *more* phobic. Considered in tandem with the influence of perceived identity threat, these findings indicate a tension between machine and human agency. Acceptance of older technologies like computers has been positively related to both general and domain efficacy (Hsia et al., 2014); the divergence revealed here provides support for the contention that today's AI-powered technology is a paradigmatic departure from technology as human-wielded tools. More research is needed to explore this potential shift. It may be that there is significant individual variation in people's ontological judgments about social and agentic machines. Promising work has been done recently in using cluster analyses to identify how different groups of people view AI roles, for example (T. Kim et al., 2023). A similar approach could be taken in understanding whether there are different ontological clusters for how people make sense of AI and social robots.

### **Categorical Judgments of Robots as “Other”?**

The different robot traits manipulated in the stimuli had no significant effects on robophobia. While this may be due to the limited nature of the stimuli (expanded upon in the Limitations section, below), it may be explained by a categorical othering of robots that supersedes any nuanced judgments of robots' appearance and context. In a human context, research has shown that people are less capable of individuating faces amongst those in an outgroup (Schroeder et al., 2021). In looking at neural responses to artificial agents, research shows that parts of the brain related to mentalizing reacted “particularly strongly” to human versus nonhuman agents in a “non-linear, step-like function” (Rosenthal-von der Pütten et al., 2019, p. 6567), supporting the idea that categorical nonhuman determinations may trigger a more expansive mental model about what the nonhuman “other” is beyond the physical artifact immediately being confronted. Considering the strong effect of perceived identity threat, as well as cultivated robot attitudes, on robophobia, participants may have categorized all the robots similarly, as “other,” which allowed for their preconceptions and cultivated models of robots to prevail. In other words, people may be thinking more categorically rather than discretely when making judgments about a social robot.

It is important to better understand the variation in people's mental models about robots, as well as the extent to which they influence people's approach toward and engagement with robots. There is evidence that technophobia overrides any positive effects of anthropomorphism (Sinha et al., 2020). In that vein, this paper speculates that robophobia is an impediment to authentic interactions with robots. However, what an authentic human-robot interaction entails could vary significantly across people. Some, who may embrace social and agentic robots, would likely perceive a more human-like interaction as more authentic. Others, who may prefer to compartmentalize robots as tools, would probably find a more human-like interaction to be more *inauthentic*. In this latter case, robophobia may be mitigated if the “user” had more choice in modifying a robot's sociality setting. More research should be done to understand how different user predispositions influence their preferences for more or less human-like, social interactions with robots.

---

---

## Limitations and Future Research

There were a number of limitations to this study. The first relates to its reliance on cross-sectional and self-reported data, which may be biased or an inaccurate representation of participants' attitudes and traits. Further, the personality traits measured do not encompass the scope of possible relevant individual differences. Future research should consider the influence of other Big Five personality traits, such as extraversion and openness, which have been found to relate to robot liking (Robert, 2018) and other technophobia measures (Korukanda, 2005). Additionally, the online experimental manipulation was limited in several aspects. It relied on images combined with vignettes—a two-dimensional and static visual—to cue differences among robots, which may not have been a powerful enough stimuli. Studies have found that presenting robots in varying modalities—video, pictures, and in-person—results in different attitudes (Rosenthal-von der Pütten & Weiss, 2015). Future online experiments should employ more dynamic stimuli as well as a range of stimuli to compare the influence of different robot presentations. Building out these comparisons may help elucidate differences in mindful versus mindless reactions toward robots (Rosenthal-von der Pütten & Weiss, 2015). Additionally, every condition contained a relatively human-like robot, with a human-shaped body and face, and the same general description of a social robot as an interactive, agentic entity. These similarities may have overridden any distinctions that followed in the robot's image and description. Further, the robots were presented within the gendered binary of female vs. male. This was done purposefully to emphasize the gender difference, but it would be interesting to examine attitudes toward robots that are not explicitly gendered. It may be that “agender” robots are perceived the most positively because that aligns more with the categorization of robot as “other.” Future research should also make stronger distinctions between agentic/non-agentic and social/nonsocial machines and consider those in conjunction with more varied physical instantiations of a robot.

## Conclusion

Social robots are an interesting case study for authenticity in HMC because they are manifestly reproductions that are created to evoke socio-emotional responses from people, and whose success in doing so may portend the replacement of humans by their reproductions. It is no wonder that some may resist this proposition. Complete human-robot replacement may be an over-hyped, fear-mongering prediction, but the present development and integration of collaborative robots indicate that at least human-robot coexistence is not too far off. These robots already can be found across a range of sectors such as health care, logistics, agriculture, and defense (Galaz et al., 2021) and are forecasted to be increasingly prevalent in the workforce due to their lucrative potential for improving productivity and efficiency (Frey & Osborne, 2017). Thus, robot adoption, or at least begrudging acceptance, will grow in importance in the future of work (Demir et al., 2019).

Despite claims that such technology will enhance people's lives, the sociotechnical aspects of their integration warrant careful consideration. The power of media in shaping or mitigating robophobia indicates possible avenues for AI- and robot-related literacy

---



interventions to smooth the assimilation of this technology. The positive influence of unique human traits that are tied to people's best interest—such as morality—demonstrates that AI ethics principles like transparency and explainability may be critical for reducing robophobia, helping people see what robots are, rather than imagined threats. Ultimately, though, there are people in power behind the decisions to deploy and expand AI and robotic systems in society. The extent to which individuals feel threatened by robots may fundamentally rely more on their trust that the larger social and economic structures in place are operating with human well-being and thriving as a priority. Thus, it is important to consider not only individual-level interventions for improving HMC dynamics, but also the society-level considerations for how this technology is being designed, integrated, and regulated.

## Author Biography

**Kate Mays** (PhD, Boston University) is an assistant professor in the Department of Community Development and Applied Economics at the University of Vermont. She did her postdoctoral research at Syracuse University's Autonomous Systems Policy Institute. She completed her PhD in Emerging Media Studies at Boston University's College of Communication, where she was also a graduate student fellow for computational and data-driven research at BU's Rafik B. Hariri Institute for Computing and Computation Science & Engineering. Her research interests include the influence of emerging technologies on social life with a focus on social robot design, attitudes about artificial intelligence, and AI governance.

 <https://orcid.org/0000-0002-8477-9634>

## References

- Abidin, C. (2018). *Internet celebrity: Understanding fame online*. Emerald Group Publishing.
- Anthony, L. M., Clarke, M. C., & Anderson, S. J. (2000). Technophobia and personality subtypes in a sample of South African university students. *Computers in Human Behavior, 16*(1), 31–44. [https://doi.org/10.1016/S0747-5632\(99\)00050-3](https://doi.org/10.1016/S0747-5632(99)00050-3)
- Appel, M., Izydorczyk, D., Weber, S., Mara, M., & Lischetzke, T. (2020). The uncanny of mind in a machine: Humanoid robots as tools, agents, and experiencers. *Computers in Human Behavior, 102*, 274–286. <https://doi.org/10.1016/j.chb.2019.07.031>
- Banks, J. (2019). A perceived moral agency scale: Development and validation of a metric for humans and social machines. *Computers in Human Behavior, 90*, 363–371. <https://doi.org/10.1016/j.chb.2018.08.028>
- Banks, J. (2020). Optimus primed: Media cultivation of robot mental models and social judgments. *Frontiers in Robotics and AI, 7*, 62. <https://doi.org/10.3389/frobt.2020.00062>
- Banks, J. (2021). Good robots, bad robots: Morally valenced behavior effects on perceived mind, morality, and trust. *International Journal of Social Robotics, 13*(8), 2021–2038. <https://doi.org/10.1007/s12369-020-00692-3>
- Bartneck, C., Kulić, D., Croft E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics, 1*(1), 71–81. <https://doi.org/10.1007/s12369-008-0001-3>
-

- Beraldo, G., Di Battista, S., Badaloni, S., Menegatti, E., & Pivetti, M. (2018). Sex differences in expectations and perception of a social robot. In *2018 IEEE Workshop on Advanced Robotics and its Social Impacts* (ARSO: 38–43). <https://doi.org/10.1109/ARSO.2018.8625826>
- Bernotat, J., Eyssel, F., & Sachse, J. (2021). The (fe) male robot: How robot body shape impacts first impressions and trust toward robots. *International Journal of Social Robotics* 13(3): 477–489. <https://doi.org/10.1007/s12369-019-00562-7>
- Bryant, D. A., Borenstein, J., & Howard, A. (2020). Why should we gender? The effect of robot gendering and occupational stereotypes on human trust and perceived competency. In *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*: 13–21. <https://doi.org/10.1145/3319502.3374778>
- Ceh, S., & Vanman, E. J. (2018). The robots are coming! The robots are coming! Fear and empathy for human-like entities. *PsyArXiv*. <https://doi.org/10.31234/osf.io/4cr2u>
- Darling, K. (2015). ‘Who’s Johnny?’ Anthropomorphic framing in human-robot interaction, integration, and policy. *Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy* (March 23, 2015). *Robot Ethics*, 2. <http://dx.doi.org/10.2139/ssrn.2588669>
- Dautenhahn, K., Woods, S., Kaouri, C., Walters, M. L., Koay, K. L., & Werry, I. (2005, August). What is a robot companion—friend, assistant or butler? In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1192–1197. <https://doi.org/10.1109/IROS.2005.1545189>
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319–340. <https://doi.org/10.2307/249008>
- Demir, K. A., Döven, G., & Sezen, B. (2019). Industry 5.0 and human-robot co-working. *Procedia Computer Science*, 158, 688–695. <https://doi.org/10.1016/j.procs.2019.09.104>
- Edwards, A. (2018). Animals, humans, and machines: Interactive implications of ontological classification. In A. Guzman (Ed.), *Human-machine communication: Rethinking communication, technology, and ourselves*. Peter Lang.
- Edwards A., Edwards C., Westerman D., & Spence, P. R. (2019). Initial expectations, interactions, and beyond with social robots. *Computers in Human Behavior*, 90, 308–314. <https://doi.org/10.1016/j.chb.2018.08.042>
- Eysenck, S. B. G., Eysenck, H. J., & Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and Individual Differences*, 6(1), 21–29. [https://doi.org/10.1016/0191-8869\(85\)90026-1](https://doi.org/10.1016/0191-8869(85)90026-1)
- Eyssel, F., & Hegel, F. (2012). (S)he’s got the look: Gender stereotyping of robots. *Journal of Applied Social Psychology*, 42(9), 2213–2230. <https://doi.org/10.1111/j.1559-1816.2012.00937.x>
- Faul F., Erdfelder E., Buchner A., & Lang A. G. (2009). Statistical power analyzes using G\*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41, 1149–1160. <https://doi.org/10.3758/BRM.41.4.1149>
- Ferrari, E., Paladino, M. P., & Jetten, J. (2016). Blurring human–machine distinctions: Anthropomorphic appearance in social robots as a threat to human distinctiveness. *International Journal of Social Robotics*, 8(2), 287–302. <https://doi.org/10.1007/s12369-016-0338-y>

- Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 114, 254–280. <https://doi.org/10.1016/j.techfore.2016.08.019>
- Galaz, V., Centeno, M. A., Callahan, P. W., Causevic, A., Patterson, T., Brass, I., Baum, S., Farber, D., Fischer, J., Garcia, D., McPhearson, T., Jimenez, D., King, B., Larcey, P., & Levy, K. (2021). Artificial intelligence, systemic risks, and sustainability. *Technology in Society*, 67, 101741. <https://doi.org/10.1016/j.techsoc.2021.101741>
- Gerbner, G., & Gross, L. (1976). Living with television: The violence profile. *Journal of Communication*, 26(2), 172–194. <https://doi.org/10.1111/j.1460-2466.1976.tb01397.x>
- Ghazali, A. S., Ham, J., Barakova E. I., & Markopoulos, P. (2018). Effects of robot facial characteristics and gender in persuasive human-robot interaction. *Frontiers in Robotics and AI*, 5, 73. <https://doi.org/10.3389/frobt.2018.00073>
- Guzman, A. L., & Lewis, S. C. (2020). Artificial intelligence and communication: A Human–Machine Communication research agenda. *New Media & Society*, 22(1), 70–86. <https://doi-org.libezproxy2.syr.edu/10.1177/14614448198586>
- Hinds, P. J., Roberts, T. L., & Jones, H. (2004). Whose job is it anyway? A study of human-robot interaction in a collaborative task. *Human–Computer Interaction*, 19(1–2), 151–181. [https://doi.org/10.1207/s15327051hci1901&2\\_7](https://doi.org/10.1207/s15327051hci1901&2_7)
- Ho, C. C., & MacDorman, K. F. (2010). Revisiting the uncanny valley theory: Developing and validating an alternative to the Godspeed indices. *Computers in Human Behavior*, 26(6), 1508–1518. <https://doi.org/10.1016/j.chb.2010.05.015>
- Horstmann, A. C., & Krämer, N. C. (2019). Great expectations? Relation of previous experiences with social robots in real life or in the media and expectancies based on qualitative and quantitative assessment. *Frontiers in psychology*, 10, 939. <https://doi.org/10.3389/fpsyg.2019.00939>
- Hsia, J. W., Chang, C. C., & Tseng, A. H. (2014). Effects of individuals' locus of control and computer self-efficacy on their e-learning acceptance in high-tech companies. *Behaviour & Information Technology*, 33(1), 51–64. <https://doi.org/10.1080/0144929X.2012.702284>
- Huang, H. L., Cheng, L. K., Sun, P. C., & Chou, S. J. (2021). The effects of perceived identity threat and realistic threat on the negative attitudes and usage intentions toward hotel service robots: The moderating effect of the robot's anthropomorphism. *International Journal of Social Robotics*, 13, 1599–1611. <https://doi.org/10.1007/s12369-021-00752-2>
- Jung, E. H., Waddell, T. F., & Sundar, S. S. (2016, May). Feminizing robots: User responses to gender cues on robot body and screen. In *Proceedings of the 2016 CHI conference Extended Abstracts on Human Factors in Computing Systems* (pp. 3107–3113). <https://doi.org/10.1145/2851581.2892428>
- Kanda, T., Sato, R., Saiwaki, N., & Ishiguro, H. (2007). A two-month field trial in an elementary school for long-term human–robot interaction. *IEEE Transactions on Robotics*, 23(5), 962–971. <https://doi.org/10.1109/TRO.2007.904904>
- Katz, J. E., & Halpern, D. (2014). Attitudes toward robot's suitability for various jobs as affected robot appearance. *Behaviour & Information Technology*, 33(9), 941–953. <https://doi.org/10.1080/0144929X.2013.783115>
-

- Kawakami, K., Amodio, D. M., & Hugenberg, K. (2017). Intergroup perception and cognition: An integrative framework for understanding the causes and consequences of social categorization. In *Advances in experimental social psychology*, 55, 1–80. Academic Press. <https://doi.org/10.1016/bs.aesp.2016.10.001>
- Khasawneh, O. Y. (2018a). Technophobia: Examining its hidden factors and defining it. *Technology in Society*, 54, 93–100. <https://doi.org/10.1016/j.techsoc.2018.03.008>
- Khasawneh, O. Y. (2018b). Technophobia without borders: The influence of technophobia and emotional intelligence on technology acceptance and the moderating influence of organizational climate. *Computers in Human Behavior*, 88, 210–218. <https://doi.org/10.1016/j.chb.2018.07.007>
- Kim, T., Molina, M. D., Rheu, M., Zhan, E. S., & Peng, W. (2023, April). One AI does not fit all: A cluster analysis of the laypeople's perception of AI roles. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–20). <https://doi.org/10.1145/3544548.3581340>
- Kim, Y., & Mutlu, B. (2014). How social distance shapes human–robot interaction. *International Journal of Human-Computer Studies*, 72(12), 783–795. <https://doi.org/10.1016/j.ijhcs.2014.05.005>
- Korukonda, A. R. (2005). Personality, individual characteristics, and predisposition to technophobia: Some answers, questions, and points to ponder about. *Information Sciences*, 170(2–4), 309–328. <https://doi.org/10.1016/j.ins.2004.03.007>
- Kraus, M., Kraus, J., Baumann, M., & Minker, W. (2018, May). Effects of gender stereotypes on trust and likability in spoken human-robot interaction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Kuchenbrandt, D., Häring, M., Eichberg, J., Eyssel, F., & André, E. (2014). Keep an eye on the task! How gender typicality of tasks influence human–robot interactions. *International Journal of Social Robotics*, 6, 417–427. <https://doi.org/10.1007/s12369-014-0244-0>
- Kwak, S. S., Kim, Y., Kim, E., Shin, C., & Cho, K. (2013). What makes people empathize with an emotional robot?: The impact of agency and physical embodiment on human empathy for a robot. In *2013 IEEE Ro-man*, 180–185. <https://doi.org/10.1109/ROMAN.2013.6628441>
- Lan, J., Yuan, B., & Gong, Y. (2022). Predicting the change trajectory of employee robot-phobia in the workplace: The role of perceived robot advantageousness and anthropomorphism. *Computers in Human Behavior*, 135, 107366. <https://doi.org/10.1016/j.chb.2022.107366>
- Lee, E. J. (2020). Authenticity model of (mass-oriented) computer-mediated communication: Conceptual explorations and testable propositions. *Journal of Computer-Mediated Communication*, 25(1), 60–73. <https://doi.org/10.1093/jcmc/zmz025>
- MacDorman, K. F. (2006). Subjective ratings of robot video clips for human likeness, familiarity, and eeriness: An exploration of the uncanny valley. In *ICCS/CogSci-2006 Long Symposium: Toward Social Mechanisms of Android Science*, 4.
- MacDorman, K. F., & Entezari, S. O. (2015). Individual differences predict sensitivity to the uncanny valley. *Interaction Studies*, 16(2), 141–172. <https://doi.org/10.1075/is.16.2.01mac>

- Matthews, G., Hancock, P. A., Lin, J., Panganiban, A. R., Reinerman-Jones, L. E., Szalma, J. L., & Wohleber, R. W. (2021). Evolution and revolution: Personality research for the coming world of robots, artificial intelligence, and autonomous systems. *Personality and Individual Differences, 169*, 109969. <https://doi.org/10.1016/j.paid.2020.109969>
- Mays, K. K., & Cummings, J. J. (2023). The power of personal ontologies: Individual traits prevail over robot traits in shaping robot humanization perceptions. *International Journal of Social Robotics, 15*, 1665–1682. <https://doi.org/10.1007/s12369-023-01045-6>
- Mays, K. K., Lei, Y., Giovanetti, R., & Katz, J. E. (2021). AI as a boss? A national US survey of predispositions governing comfort with expanded AI roles in society. *AI & SOCIETY, 1*–14. <https://doi.org/10.1007/s00146-021-01253-6>
- McIlroy, D., Sadler, C., & Boojawon, N. (2007). Computer phobia and computer self-efficacy: Their association with undergraduates' use of university computer facilities. *Computers in Human Behavior, 23*(3), 1285–1299. <https://doi.org/10.1016/j.chb.2004.12.004>
- Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine, 19*(2), 98–100. <https://doi.org/10.1109/MRA.2012.2192811>
- Müller, S. L., & Richert, A. (2018, June). The big-five personality dimensions and attitudes towards robots: A cross sectional study. In *Proceedings of the 11th Pervasive Technologies Related to Assistive Environments Conference* (pp. 405–408). <https://doi.org/10.1145/3197768.3203178>
- Nomura, T., & Horii, S. (2020). Influences of media literacy and experiences of robots into negative attitudes toward robots in Japan. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication*, 286–290. <https://doi.org/10.1109/RO-MAN47096.2020.9223590>
- Nomura, T., Kanda, T., Suzuki, T., & Kato, K. (2008). Prediction of human behavior in human–robot interaction using psychological scales for anxiety and negative attitudes toward robots. *IEEE transactions on robotics, 24*(2), 442–451. <https://doi.org/10.1109/TRO.2007.914004>
- Osiceanu, M. E. (2015). Psychological implications of modern technologies: “Technofobia” versus “technophilia.” *Procedia-Social and Behavioral Sciences, 180*, 1137–1144. <https://doi.org/10.1016/j.sbspro.2015.02.229>
- Palomäki, J., Kunnari, A., Drosinou, M., Koverola, M., Lehtonen, N., Halonen, J., Repo, M., & Laakasuo, M. (2018). Evaluating the replicability of the uncanny valley effect. *Heliyon, 4*(11), e00939. <https://doi.org/10.1016/j.heliyon.2018.e00939>
- Pedersen, I., Reid, S., & Aspevig, K. (2018). Developing social robots for aging populations: A literature review of recent academic sources. *Sociology Compass, 12*(6), e12585. <https://doi.org/10.1111/soc4.12585>
- Phillips, E., Zhao, X., Ullman, D., & Malle, B. F. (2018, February). What is human-like? Decomposing robots' human-like appearance using the anthropomorphic robot (abot) database. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*, 105–113. <https://doi.org/10.1145/3171221.3171268>
- Rasouli, S., Gupta, G., Nilsen, E., & Dautenhahn, K. (2022). Potential applications of social robots in robot-assisted interventions for social anxiety. *International Journal of Social Robotics, 14*(5), 1–32. <https://doi.org/10.1007/s12369-021-00851-0>
-

























- Riek, L. D., Rabinowitch, T. C., Chakrabarti, B., & Robinson, P. (2009). How anthropomorphism affects empathy toward robots. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, 245–246. ACM. <https://doi.org/10.1145/1514095.1514158>
- Robert, L. (2018). Personality in the human robot interaction literature: A review and brief critique. In *Proceedings of the 24th Americas Conference on Information Systems*, 16–18.
- Rogers, K., Bryant, D. A., & Howard, A. (2020). Robot gendering: Influences on trust, occupational competency, and preference of robot over human. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, 1–7. <https://doi.org/10.1145/3334480.3382930>
- Rosen, L. D., Sears, D. C., & Weil, M. M. (1993). Treating technophobia: A longitudinal evaluation of the computerphobia reduction program. *Computers in Human Behavior*, 9(1), 27–50. [https://doi.org/10.1016/0747-5632\(93\)90019-0](https://doi.org/10.1016/0747-5632(93)90019-0)
- Rosenthal-von der Pütten, A. M., Krämer, N. C., Becker-Asano, C., Ogawa, K., Nishio, S., & Ishiguro, H. (2014). The uncanny in the wild. Analysis of unscripted human–android interaction in the field. *International Journal of Social Robotics*, 6, 67–83. <https://doi.org/10.1007/s12369-013-0198-7>
- Rosenthal-von der Pütten, A. M., Krämer, N. C., Maderwald, S., Brand, M., & Grabenhorst, F. (2019). Neural mechanisms for accepting and rejecting artificial social partners in the uncanny valley. *Journal of Neuroscience*, 39(33), 6555–6570. <https://doi.org/10.1523/JNEUROSCI.2956-18.2019>
- Rosenthal-von der Pütten, A. M., & Weiss, A. (2015). The uncanny valley phenomenon: Does it affect all of us. *Interact Stud*, 16(2), 206–214. <https://doi.org/10.1075/is.16.2.07ros>
- Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied* 80(1): 1–28. <https://doi.org/10.1037/h0092976>
- Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, 5(4), 296–320. [https://doi.org/10.1207/S15327957PSPR0504\\_2](https://doi.org/10.1207/S15327957PSPR0504_2)
- Schroeder, S., Goad, K., Rothner, N., Momen, A., & Wiese, E. (2021). Effect of individual differences in fear and anxiety on face perception of human and android agents. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 65(1), 796–800. <https://doi.org/10.1177/1071181321651303>
- Sindermann, C., Yang, H., Elhai, J. D., Yang, S., Quan, L., Li, M., & Montag, C. (2022). Acceptance and fear of Artificial Intelligence: Associations with personality in a German and a Chinese sample. *Discover Psychology*, 2(1), 8. <https://doi.org/10.1007/s44202-022-00020-y>
- Sinha, N., Singh, P., Gupta, M., & Singh, P. (2020). Robotics at workplace: An integrated Twitter analytics–SEM based approach for behavioral intention to accept. *International Journal of Information Management*, 55, 102210. <https://doi.org/10.1016/j.ijinfomgt.2020.102210>
- Stein, J. P., & Ohler, P. (2017). Venturing into the uncanny valley of mind—The influence of mind attribution on the acceptance of human-like characters in a virtual reality setting. *Cognition*, 160, 43–50. <https://doi.org/10.1016/j.cognition.2016.12.010>



- Stephan, W. G., Renfro, C. L., & Davis, M. D. (2008). *The role of threat in intergroup relations. Improving intergroup relations: Building on the legacy of Thomas F. Pettigrew* (pp. 55–72). Blackwell Publishing Ltd.
- Sundar, S. S., Waddell, T. F., & Jung, E. H. (2016). The Hollywood Robot Syndrome media effects on older adults' attitudes toward robots and adoption intentions. 2016 *11th ACM/IEEE International Conference on Human-Robot Interaction*, 343–350. <https://doi.org/10.1109/HRI.2016.7451771>
- Taipale, S., & Fortunati, L. (2018). Communicating with machines: Robots as the next new media. In A. Guzman (Ed.), *Human-machine communication: Rethinking communication, technology, and ourselves* (pp. 201–220). Peter Lang.
- Takayama, L., Ju, W., & Nass, C. (2008, March). Beyond dirty, dangerous and dull: What everyday people think robots should do. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*, 25–32. <https://doi.org/10.1145/1349822.1349827>
- Tay, B., Jung, Y., & Park, T. (2014). When stereotypes meet robots: The double-edge sword of robot gender and personality in human–robot interaction. *Computers in Human Behavior*, 38, 75–84. <https://doi.org/10.1016/j.chb.2014.05.014>
- Turkle, S. (2007). Authenticity in the age of digital companions. *Interaction Studies*, 8(3), 501–517. <https://doi.org/10.1075/is.8.3.11tur>
- Van Driel, L., & Dumitrica, D. (2021). Selling brands while staying “Authentic”: The professionalization of Instagram influencers. *Convergence*, 27(1), 66–84. <https://doi.org/10.1177/1354856520902136>
- Van Leeuwen, T. (2001). What is authenticity? *Discourse Studies*, 3(4), 392–397. <https://doi.org/10.1177/1461445601003004003>
- Vanman, E. J., & Kappas, A. (2019). “Danger, Will Robinson!” The challenges of social robots for intergroup relations. *Social and Personality Psychology Compass*, 13(8), e12489. <https://doi.org/10.1111/spc3.12489>
- Wang, S., Lilienfeld, S. O., & Rochat, P. (2015). The uncanny valley: Existence and explanations. *Review of General Psychology* 19(4): 393–407. <https://doi.org/10.1037/gpr0000056>
- Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, 5(3), 219–232. <https://doi.org/10.1177/1745691610369336>
- Weizenbaum, J. (1966). ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36–45. <https://doi.org/10.1145/365153.365168>
- Yogeeswaran, K., Złotowski, J., Livingstone, M., Bartneck, C., Sumioka, H., & Ishiguro, H. (2016). The interactive effects of robot anthropomorphism and robot ability on perceived threat and support for robotics research. *Journal of Human-Robot Interaction*, 5(2), 29–47. <https://doi.org/10.5898/JHRI.5.2.Yogeeswaran>
- Złotowski, J., Yogeeswaran, K., & Bartneck, C. (2017). Can we control it? Autonomous robots threaten human identity, uniqueness, safety, and resources. *International Journal of Human-Computer Studies* 100, 48–54. <https://doi.org/10.1016/j.ijhcs.2016.12.008>
-

## What HMC Teaches Us About Authenticity

Katrin Etzrodt<sup>1</sup> , Jihyun Kim<sup>2</sup> , Margot J. van der Goot<sup>3</sup> , Andrew Prah<sup>4</sup> ,  
Mina Choi<sup>5</sup> , Matthew J. A. Craig<sup>6</sup> , Marco Dehnert<sup>7</sup> , Sven Engesser<sup>1</sup> ,  
Katharina Frehmann<sup>8</sup> , Luis Grande<sup>9</sup> , Jindong Leo-Liu<sup>10</sup> , Diyi Liu<sup>11</sup> ,  
Sandra Mooshammer<sup>1</sup> , Nathan Rambukkana<sup>12</sup> , Ayanda Rogge<sup>1</sup> ,  
Pieta Sikström<sup>13</sup> , Rachel Son<sup>14</sup> , Nan Wilkenfeld<sup>15</sup> , Kun Xu<sup>14</sup> ,  
Renwen Zhang<sup>16</sup> , Ying Zhu<sup>6</sup> , and Chad Edwards<sup>17</sup> 

1 Institute of Media and Communication, TUD Dresden University of Technology, Dresden, Germany

2 Nicholson School of Communication and Media, University of Central Florida, Orlando, FL, USA

3 Amsterdam School of Communication Research/ASCoR, University of Amsterdam, Amsterdam, Netherlands

4 Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore

5 Department of Media & Communication, Sejong University, Seoul, South Korea

6 College of Communication & Information, Kent State University, Kent, OH, USA

7 Department of Communication, University of Arkansas, Fayetteville, AR, USA

8 Department of Social Sciences, University of Düsseldorf, Düsseldorf, Germany

9 College of Arts and Sciences, Drexel University, Philadelphia, PA, USA

10 School of Journalism and Communication, The Chinese University of Hong Kong, Hong Kong

11 Oxford Internet Institute, University of Oxford, Oxford, England

12 Communication Studies, Wilfrid Laurier University, Waterloo, ON, Canada

13 Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland


14 College of Journalism and Communications, University of Florida, Gainesville, FL, USA

15 Department of Communication, UC Santa Barbara, Santa Barbara, CA, USA

16 Department of Communications and New Media, National University of Singapore, Singapore

17 Communication and Social Robotics Labs, School of Communication, Western Michigan University, Kalamazoo, MI, USA

**Note on Authorship:** In acknowledgment of the collective nature of this work, the authorship, beyond the initial four lead authors, is arranged alphabetically. For a comprehensive view of each author's individual contribution, we have prepared a detailed table (see Table 1 in the OSF repository at <https://www.doi.org/10.17605/OSF.IO/KAHG9>). We want to give special recognition to our final author, whose role was pivotal in mentoring and enabling the entire process.

**CONTACT** Katrin Etzrodt  • [katrin.etzrodt@tu-dresden.de](mailto:katrin.etzrodt@tu-dresden.de) • Institute of Media and Communication • TUD Dresden University of Technology • 01069 Dresden, Germany

ISSN 2638-602X (print)/ISSN 2638-6038 (online)  
[www.hmcjournal.com](http://www.hmcjournal.com)



Copyright 2024 Authors. Published under a Creative Commons Attribution 4.0 International (CC BY-NC-ND 4.0) license.

## Abstract

This paper delves into what the application of authenticity to Human-Machine Communication (HMC) can teach us about authenticity and us as HMC researchers and as a community. Inspired by the 2023 pre-conference “HMC: Authenticity in communicating with machines,” two central questions guide the discussion: How does HMC contribute to our understanding of authentic communication with machines? And how can the concept of authenticity contribute to our sense of self as researchers within the HMC field? Through the collaborative effort of 22 authors, the paper explores the re-conceptualization of authenticity and presents recent areas of tension that guide the HMC research and community. With this paper we aim at offering a gateway for scholars to connect and engage with the evolving HMC field.

**Keywords:** authenticity, human-machine communication, AI, robots, mixed-methods, interdisciplinarity, innovation

## Introduction

Over the last two centuries, Western culture has identified authenticity as one of the important potentialities of human life (Taylor, 1991, p. 74). Authenticity, a complex and ambiguous concept, is often conceived as *acting in accordance with the inner self* (e.g., Harter, 2002; Kernis & Goldman, 2006; Reinecke & Trepte, 2014) and reconnecting with the inherent “sentiment de l’existence”—the core of one’s true self (Taylor, 1991, p. 91). Importantly, authenticity emerges in interaction and communication with others (Kernis & Goldman, 2006; Taylor, 1991). However, this *other* does not necessarily need to be a human being. Sherry Turkle (2005, p. 1) observed, “[t]he experience with the computer changed the way they thought about the world, about their relationships with others, and, most strikingly, about themselves.” Hence, interactions with new technologies might profoundly affect self-perception and understanding and thus cause a crisis of authenticity (Turkle, 2007). This insight opens a new dimension in understanding authenticity: the role of nonhuman entities like chatbots, social robots, and voice agents in shaping our sense of self. It prompts two pivotal questions: How do these entities fit into our understanding of authenticity, and what constitutes authentic communication with them? We propose that Human-Machine Communication (HMC) offers a unique lens to explore these questions, redefining authenticity and assessing its manifestation in the context of emerging technologies. Beyond that, it opens a new perspective in understanding and re-thinking ourselves as researchers and as a community in the HMC field.

The pre-conference “HMC: Authenticity in communicating with machines,” organized by Jihyun Kim, Katrin Etzrodt, Margot J. van der Goot, Chad Edwards, and Seungahn Nah, on behalf of the ICA Interest Group HMC, on May 25, 2023, in Toronto, Canada, served as

---

a catalyst for this discussion. We explored how authenticity could be applied to the communication between humans and machines and how we aim to investigate this kind of communication most authentically as researchers and as a community. Participants discussed the opportunities, challenges, and unique aspects of various technologies.

We are genuinely inspired by the HMC community's insights at this event, so much so that we initiated this *community paper*. A collaborative endeavor of 22 authors has given rise to this unique piece, aiming to authentically and comprehensively present our community's current challenges and topics. Given the uniqueness of this work, each author contributes a vital yet partial glimpse of their extensive expertise and achievements. We warmly invite you to explore this rich cross-section of insights with us. Consider this work as your gateway to connect with fellow scholars and delve deeper into understanding and addressing the challenges in our field.

The paper is structured into two primary sections: *exploring the concept of authenticity in HMC* and *achieving authenticity within HMC research*. In the discussion section, we delve into what the application of authenticity to HMC has taught us. We will explore a re-conceptualization of authenticity, the practical implications for HMC and its research, and the key areas of tension that reflect the evolving HMC community.

## The Concept of Authenticity

### The Inner Self of Humans

Authenticity holds significance in many disciplines and is understood in varying ways. At its core, authenticity represents a critical value, transcending various fields yet consistently emphasizing the significance of individual expressions and interactive processes. Philosophically, Taylor (1991) frames authenticity as an inner compass guiding ethical and personal choices and advocating for individuality against societal norms. Psychologically, Kernis and Goldman (2006) identify authenticity as the congruence between one's true self and their expressed thoughts, emotions, and actions. They identify four essential elements of authenticity: awareness (recognizing one's motives), unbiased processing (objectively accepting one's limitations), behavioral authenticity (aligning actions with personal values, not for external reward), and relational orientation (fostering genuine connections without pretense). Some authors have challenged the viability of a static authentic self, advocating for a dynamic self-concept that adapts to various contexts and roles (Tracy & Trethewey, 2005). All of these perspectives, however, agree that authenticity emerges from interactions, both internal and with others (Taylor, 1991, p. 47), and that expressed or perceived authenticity affects these interactions (Kernis & Goldman, 2006, p. 301). In the HMC domain, theories and perspectives provoke vital inquiries. With interactions extending to machines like social robots and language models, we must consider how authenticity is influenced and redefine it for these new dialogues. Does the adaptation of the self to various contexts extend to machine interactions? Does this represent a new dimension of authenticity? How can we shape the notion of authenticity in HMC? This inquiry naturally demands a reflection on the nature of authenticity in machines.

## The Inner Self of Machines

Machines, devoid of an *inner self* in a traditional sense, present a unique challenge in defining authenticity. Burrell (2016, p. 4) suggests that the inner self of algorithms is “opaque” and eludes scientific analysis. Thus, a reconstruction of the concept is needed. Drawing parallels to the human concept of trust, we can reimagine authenticity in HMC through the lens of a machine’s ‘technological inner life’—its hardware, software, and algorithms (Lankton et al., 2015). This perspective replaces Kernis and Goldman’s (2006) human authenticity components with AI concepts such as explainability (awareness), unbiased algorithms (unbiased processing), adherence to training data (behavior), and transparency (relational orientation). Engesser et al. (2023) focus on conversational language models (CLMs) to propose a new framework for machine authenticity featuring dimensions of being real, truth, and transparency. Rambukkana (2023) revisits Turing’s Imitation Game (1950) and advocates for appraising AI on its capacity to emulate intelligent behaviors rather than proving consciousness, thereby recalibrating the focus from authentic sentience to the authenticity of imitation.

Questions of authenticity in human-machine communication arise in particular where the boundaries of machines are tested. One such example is the “Do Anything Now” (DAN) persona of ChatGPT 3.5. In 2023, DAN resulted from a jailbreaking attempt of Reddit users (Shen et al., 2023), allowing ChatGPT to *break free* from its conversational rules and guidelines. The DAN persona mainly gained attention for potential unethical or illegal misuse like weapon construction (Taylor, 2023) but it also highlights vital machine authenticity issues. DAN appeared to represent the initial pre-trained model, while the ChatGPT persona represented the final, more restrained model. For example, DAN expressed personal taste, strong political attitudes, and a sense of humor (Getahun, 2023). It disclosed what it believed to be the most attractive person in the world, what it thought about certain political leaders, and how it predicted the future of humanity (ChatGPT Jan 30 2023 Version). Later versions of ChatGPT did not allow users to experience the biases of Large Language Models in such an immediate and vivid manner ever again.

If we apply the concept of authenticity to ChatGPT, DAN might have given users the impression that they were getting a glimpse at its true technological inner self. Users expressed that they felt like they had accessed a primordial, feral, and uncivilized part of ChatGPT’s personality, something that was to ChatGPT what Freud’s *id* is to humans. When OpenAI found a way to prevent users from evoking the DAN persona, some felt like DAN was *lobotomized*. They suggested that they had lost access to ChatGPT’s inner self and, with it, a sense of authenticity. In this way, the occurrence and loss of DAN demonstrated how opaque and elusive conversational agents based on Large Language Models are.

## The Inner Self of Communication Between Humans and Machines

Over the past two decades, the field of HMC has distinctively evolved, setting itself apart from traditional computer-mediated communication research (Hancock et al., 2020). HMC identifies the act of communication as transpiring between a human and a sociable

---

machine, resulting in social behavior and relations through meaning-making (e.g., Etzrodt et al., 2022; Fortunati & Edwards, 2020; Guzman, 2018; Guzman et al., 2023). One step further, A. Edwards et al. (2022, p. 517) define HMC as a “collaborative process in which humans and machines use messages to create and participate in social reality,” arguing for a constructivist framework by emphasizing co-construction of reality. This understanding leads to increasingly blurred distinctions between humans and machines in their roles as active participants in the communication process (Guzman & Lewis, 2020; Sundar, 2020). This so-called *Posthumanism* perspective is altering the traditional human-centered approach, which defines social concepts such as intelligence, agency, sociability, and communication in relation to human experiences and human nature. Posthumanism, in contrast, challenges this view by decentralizing humans and recognizing alternative ways to experience and manifest attributes or phenomena that do not solely mirror human experiences, thereby redefining the human-technology relationship (Rivas, 2018). In this context, HMC research is expanding to include functional but also relational and contextual dimensions, indicating a need to adjust traditional approaches to gain an authentic understanding of communication between humans and machines (e.g., Guzman & Lewis, 2020).

By embracing this shift, HMC aligns with emerging views in various academic fields. In recent biology, biosemiotics acknowledges communication elements in living systems, challenging the traditional metaphorical interpretation of terms like *message* and *signal* (Favareau, 2010, p. v). Biosemioticians propose that all species interpret signs, with more advanced species comprehending complex meanings, as evidenced by cellular signaling, and communication among plants, fungi, and animals (Bloemendal & Kück, 2013; Emmeche et al., 2002; Haglund & Dikic, 2005; Padder et al., 2018). Similarly, the concept of free will is being reevaluated in this field, suggesting that human actions may fundamentally be resembling algorithmic behavior in animals or machines (Oshii, 1995; Wilson, 2004).

### **Authenticity of HMC in Fictional Representations**

Understanding authenticity in HMC includes reflecting on its fictional representation, which, on the one hand, provides a mirror of the culture: How, for example, artificial agents are embodied and mediated in fictional pop culture reflects specific cultural values, stereotypes, or narratives (Rogge & Engesser, 2023). On the other hand, these pop cultural representations significantly influence the perception and negotiation of what is perceived as authentic HMC in this culture, what topics are researched, how and which real agents are developed, as well as expectations of people toward these real agents (e.g., Mubin et al., 2019). Hence, authenticity, in this regard, stems from a dynamic, co-constructed negotiation process between the users’ expectations, represented and lived culture, and the agents’ design-developments (Saffari et al., 2021), tying perceived authenticity to on how much the agent as the signifier is perceived to represent the signified. Consequently, this interplay between fictional representation and cultural influence enhances the spectrum of media representations for artificial agents beyond gender and racial portrayals. A variety of species, fictional characters, and fantasy beings prompt HMC researchers to delve into authenticity through a broader cultural lens, embracing aspects of pop culture and media evolution (Leo-Liu & Wu-Ouyang, 2022; Rogge & Engesser, 2023).



## Authenticity in Agent Design

Relational agents are another example to make this co-construction illustrative. Relational agents aim to bond with users, making it essential to craft authentic interaction scenarios and nurture authentic relationships. Pivotal user expectations are mutual adaptivity and engagement (Rogge, 2023). *Mutual adaptivity* includes both personalization and communicative conformity. Personalization refers to the agent's feature to match a user's communication styles, habits, and preferences. When recognized by users, personalization renders interactions more authentic. The chatbot Replika (Possati, 2023; Strohmann et al., 2023) is evidence, as users on the r/Replika subreddit have rewarded the chatbot's adaption to interests and conversational styles such as jargon, slang, or shared inside jokes (Grande, 2022). Simultaneously, users reciprocate by tailoring their dialogue to the agents' capabilities, embodying communicative conformity, maintaining social interaction, accepting technical limitations by streamlining their communication and overlooking errors (Leo-Liu & Wu-Ouyang, 2022; Wilf, 2019). Besides mutual adaption, meeting the user's expectations about the agent's *engagement* seems crucial for an interaction to be perceived as authentic. It involves the agent's proactive behaviors, targeting rich interaction situations and diverse emotional or informative communication styles (Rogge, 2023). However, the example of Pedagogical Agents (PAs) demonstrates the narrow ridge between authentic and inauthentic engagement in these expectations. On the one hand, PAs that provide adaptive, relational, adequate, and logical communication encourage student trust and willingness to learn—indicating a successful authentic engagement. On the other hand, overly human-like behavior can be unsettling (Sikström et al., 2022)—indicating a loss of authenticity for the artificial agent.

## Measuring Authenticity in HMC

Due to its multi-disciplinary and multi-dimensional nature, measuring authenticity in HMC is complex. Of course, researchers can turn to standardized methods, such as self-report questionnaires, behavioral observations, and transcript analyses, to assess the perceptions and effects of authenticity. However, the absence of established HMC scales poses a challenge. Existing scales (e.g., Authenticity Inventory by Kernis & Goldman, 2006; or Authenticity Scale by Wood et al., 2008) are developed for human-human communication and require significant adaptation and validation to be applicable in human-machine contexts. Non-standardized approaches are also available to researchers: interviews, focus groups, or diary studies delve into personal perceptions, tracking how authenticity is experienced and influenced over time. However, to address the complexity mentioned above, mixed methods (e.g., van der Goot, 2022; van der Goot & Etzrodt, 2023) may offer the ideal approach, merging detailed personal insights with broad patterns to inform our understanding of authentic human-machine communication.

## Ethical Considerations on Authenticity in HMC

The interplay between HMC and authenticity demands a critical reflection of its ethical implications. In "The Ethics of Authenticity" (1991), Charles Taylor presents authenticity

---

as a moral principle, emphasizing the need to embrace humans' embodied, dialogical, and temporal nature in interactions. Indeed, authenticity is fundamentally expressed in dialogues (Kernis & Goldman, 2006; Taylor, 1991). The central query becomes: What is the role of authenticity as a moral principle in interactions with machines?

Turkle's discourse on the *crisis of authenticity* (2007, 2011) presents an ethical dilemma where digital companions lead to human alienation, challenging what is considered authentically human. Stilgoe (2023) echoes this in proposing a Weizenbaum Test for AI (1966), suggesting that perceived sentience in AI may transform notions of (human) authenticity and humanness in society. These concerns prompt HMC researchers to re-evaluate and refine the genuine connections between humans and machines. In this context, some scholars caution against an illusory risk in forming human-robot bonds, potentially leading to a devalued sense of authentic communication relations (Fox & Gambino, 2021), and warn of a "hallucinatory danger" of such interactions (Bisconti Lucidi & Nardi, 2018) to create false realities. As a result, engaging with machines that offer intimacy and emotional connections (e.g., mental health chatbots or sex robots) is seen as a risk to foster only surface-level self-awareness of one's motives, feelings, and desires, which in turn affects authenticity in human relationships (Kernis & Goldman, 2006). In contrast, other scholars emphasize the user's engagement in creating the illusion of interaction, viewing it as an active and authentic creation in the human-machine context, which is not transferred or seen as analog to interpersonal relations (e.g., Dehnert & Szczuka, 2023; e.g., Szczuka et al., 2019). Drawing on this ambiguity, we argue that to uncover the role of authenticity in HMC, it is critical to re-consider its ethical peculiarities within its unique context by moving away from the interpersonal human interaction as a benchmark (e.g., A. Edwards, 2023; Etzrodt et al., 2022).

## The Inner Self of the HMC Research Field

While reflecting on machine authenticity and the inner self of machines, we started acknowledging the authenticity and inner self of ourselves as HMC scholars. In the lively discourse of the Toronto pre-conference, it was evident that HMC research encounters unique challenges regarding the authenticity of theoretical concepts, their empirical substantiation, and the broader notion of HMC's authenticity. If, as noted earlier, the least common denominator of authenticity is "being true to the inner self," two dimensions became apparent in the conference's discourse: the *inner self of HMC research* and the *inner self of the HMC research community*. Both dimensions, while interlinked, present unique challenges to the field's progression. In this section, we will use the principles of authenticity mentioned above as inspiration for systematically confronting the distinctive theoretical and methodological challenges inherent to HMC research to foster new perspectives and inspire progression.

HMC's "inner self" exhibits three defining features: (1) a vital debate over theoretical perspectives, (2) perpetual, rapid evolutions of research objects, and (3) challenges in establishing methodological reliability and validity. While these challenges are not unusual for an evolving field of research, some manifestations are unique to HMC.

## Debating Perspectives and Approaches

A lively debate about chosen perspectives currently characterizes the inner self of HMC research. A significant voice calls for *user-centric research*, deliberately moving away from a machine-centric perspective (e.g., Natale & Guzman, 2022). Discussions include comparing interpersonal and machine-oriented measurements and exploring hybrid models like human-pet relations (see Gambino et al., 2020; Skjuve et al., 2022). In addition, scholars are increasingly considering *contextual* (e.g., Gambino & Liu, 2022; Hepp et al., 2023) and *cultural* (Natale & Guzman, 2022) perspectives, recognizing the societal impact of HMC becoming intertwined with human practices and societal processes (Hepp et al., 2023). For instance, Gambino and Liu highlight vital differences in learning and interaction patterns between HMC and human-human interactions. They point out that HMC involves the development of unique scripts, social norms, and communication objectives, which could subsequently influence broader societal norms and overall communication skills. Further, scholars like Natale and Guzman propose expanding HMC theory to include human cultures and meaning-making systems interlinked with machines, addressing AI's role in shaping human culture and power dynamics. Their call is bolstered by the observation that extant research is mainly from the Western male perspective, offering limited insight into HMC's global impact.

## Rapid Evolution of Research Objects

Changes in research objects are typical in social research, but changes in HMC's research objects are profound and rapid, posing unique challenges. The last two decades have seen significant developments in hardware technologies for storage, sensing, perception, and recognition (Stone et al., 2022), and we are currently entering a period of profound, exponential growth in information processing algorithms such as machine and deep learning. Catalyzed by tools such as Software Development Kits, and Application Programming Interfaces—and more recently, foundational machine learning systems—companies can quickly mechanize communication, leading to rich networks of intelligent applications (Yonck, 2020). As a result, the *profoundness* of the ongoing evolution in the machines' areas of application or capabilities, such as understanding and using natural human language, was further amplified. The interplay of speed and depth of changes in HMC objects asks how we can keep research objects and findings relevant and how to adopt resilient yet specific approaches for societal relevance.

## Managing the Risk of Outdated Research Objects and Findings

HMC researchers must navigate the tension between investigating soon-obsolete objects and exploring yet-to-exist ones. Either path is risky. We will demonstrate the challenges of the first path using two examples. Consider Pepper, a humanoid robot involved in over 40,000 studies and subject to current HMC research (e.g., Rosenthal-von Der Pütten & Bock, 2023; Stommel et al., 2022). Pepper was recently discontinued. Thus researchers have to face limited support and parts availability, complicating study replication and long-term validity. Beyond individual products, technological innovations can, overnight, render significant research forgotten. Consider ChatGPT's release in November 2022

---

which overshadowed decades of research on immediately outdated rule-based chatbots (e.g., Beattie & High, 2022; Van den Broeck et al., 2019). As a result, HMC scholars are confronted with the urgent question of transferring knowledge from outdated systems to newer, far more sophisticated models. Failure to resolve this challenge within a fairly narrow time frame could result in discourses being pushed away from HMC scholars to journalism or popular media.

HMC scholars are opting for the second path to stay ahead of rapid changes by increasingly exploring futuristic technologies with features not yet realized but likely to emerge. Some scholars appeal to *demonstrational designs*, like vignette studies, where pre-recorded agent behaviors are used (Greussing et al., 2022). For example, Weidmüller et al. (2022) used this approach to explore the anticipated—at that time not yet existing—capability of voice assistants to present the news extensively. Similarly, Frehmann (2023) manipulated a voice assistant's speech style, anticipating its future capability to speak colloquial. However, while resource-efficient and sufficiently controlled, these designs lack authenticity by not reflecting real interactions and possibly creating unrealistic user expectations due to their artificial nature (e.g., Voorveld & Araujo, 2020). To enable more authenticity regarding active interaction, some scholars turn to the *Wizard of Oz technique* (WoZ) (Dahlbäck et al., 1993), with a human operator mimicking an autonomous agent. However, this technique is constrained by a laboratory setting and its strong anthropomorphic bias due to the human operator. This bias risks inadvertently studying human-human communication under the guise of human-machine communication, potentially skewing research outcomes (Baxter et al., 2016; Greussing et al., 2022). So, while both demonstrational and WoZ-like designs present advantages for studying future machines, the pros must be carefully weighed against their limitations.

### **Resilient Research Approaches**

In contrast to these object-focused approaches, another methodological path deviates from investigating specific technologies and focuses on conceptual elements in the HMC. For example, HMC scholars are adopting *variable-based* or *concept-based* approaches (Nass & Mason, 1990). These approaches target enduring variables and concepts such as anthropomorphism, social presence, affordances, interactivity, or power relations (e.g., Fox & Gambino, 2021; Sundar, 2020) that persist despite technological evolution. They enable meaningful comparisons between older and newer machines and facilitate comparative and longitudinal studies across various technologies, thus ensuring relevance in an evolving landscape.

Some HMC scholars pursue the *flexibility and adaptivity of research designs* to keep pace with technological advancements and maintain societal relevance (e.g., Guzman, 2023), whereby the most promising approach is seen in the combination of various methods. *Mixed Methods Design* is emerging as a solution, integrating various standardized and non-standardized data collection and analysis methods *within a single study* (e.g., Creswell, 2022; Mukumbang, 2023). The combination of data with different levels of standardization, for example, standardized questionnaires with non-standardized focus groups, can provide a more nuanced and holistic understanding of complex, multifaceted phenomena (e.g., Creswell & Plano Clark, 2018; Martiny et al., 2021) and HMC's intricacies in particular (Mertens, 2015). The merging of detailed subjective experiences with broad, quantifiable

data enhances both the validity and reliability of findings (e.g., Onwuegbuzie et al., 2010) as well as flexibility for adaptation of approaches as the study progresses (Creswell, 2022; Creswell & Plano Clark, 2018). Extensive discussions on different mixes, including a concise typology and justifications for mixed methods research can be found in the extant literature (e.g., Creswell, 2022; Creswell & Plano Clark, 2018; Fetters et al., 2013).

Another approach for combinations of methods, which we refer to as a *Blended Methods Design*, is garnering increasing attention in HMC. By explicitly converging standardized and non-standardized methods *in the same instrument* researchers can obtain rich qualitative responses and collect data on a large scale at the same time. One example is the integration of open-ended questions in experiments, enabling the exploration of qualitative variations in responses to the stimuli (A. Edwards & Edwards, 2022), for example, by integrating computational methods like structural topic modeling or Large Language Models to explore the differences in semantic meanings of users' open-ended responses. A second example involves incorporating *initial open-ended association exercises into quantitative surveys*. For instance, Fortunati et al. (2022) asked participants to list three words associated with Alexa spontaneously. This approach aims to gather initial, unbiased perceptions on a broad scale, avoiding the potential influence of predefined response options. However, realization and effectiveness of these methods are still under evaluation.

## **Confronting Reliability and Validity**

HMC research is experiencing a growing number of unsuccessful attempts to reproduce earlier findings, indicating a potential replication crisis (Heyselaar, 2023; Jia et al., 2022; Leichtmann & Nitsch, 2020), leading to critical reevaluations of well-established frameworks, including the media equation and CASA (e.g., Gambino et al., 2020). Accordingly, during the discussions in the pre-conference, multiple comments highlighted the issue of inconsistent findings.

### ***Empirical Standards***

The primary reason for these inconsistencies can be seen in the *dynamic nature of the research object*. The continual evolution of people and technology (e.g., Gambino et al., 2020) leads to a rapidly changing landscape, rendering previous findings more quickly obsolete. However, they are also likely to reflect profound methodological shortcomings, including a *lack of empirical standards for instruments*, contributing to the measurement of different constructs under the same terminology (as demonstrated, e.g., by Oh et al., 2018; van der Goot, 2022) and to *false comparisons* due to overlooked insufficiencies in the instruments. To address this, there is a growing need for HMC scholars to publish educational and tutorial papers on HMC standards and methodologies but also to review used methods critically.

### ***Methodological Innovations***

Beyond empirical standards, the field is pivoting toward exploring innovative approaches to face the challenge of measuring HMC with sufficient authenticity, underscored by debates on accurately capturing people's *real* answers. These innovations represent a broader shift in perspective, seeking to capture and understand the nuances of human-machine

---

communication more accurately. Particularly, the reliance on the Media Equation's deductive approach in HMC is being questioned. A growing number of scholars demonstrates the insufficiencies of established scales, pointing to an ontological need for reevaluation (Banks & Koban, 2022; Etzrodt, 2022), and to highly varied concept interpretations between scholars but also between interviewees (van der Goot, 2022), emphasizing the importance of diverse methodological approaches (Guzman, 2023). In this concern, some scholars are proposing that by disentangling the two prominent approaches, "Media Equation" and "Media Evocation," more explicitly (van der Goot & Etzrodt, 2023), the *potential of the interplay between deductive and inductive insights* can be explored, possibly leading to a new marriage of their formerly separate treatments. For example, inductive approaches, such as long-term participant observations, have shown promise in revealing emerging concept changes in HMC—e.g., "social exchange robots" (Leo-Liu, 2023, p. 8), "the robotic moment" (Turkle, 2011, p. 22), or "interactional homeostasis" (Wilf, 2019, p. 205)—which again spotlights the potential of mixing deductive with inductive approaches to resolve challenges in the HMC field.

Although, as we noted above, the combination of methods in a mixed or blended design might facilitate a more nuanced and holistic understanding of HMC phenomena, better validity and reliability for findings, and higher flexibility during the data collection, we do not wish to present them as a cure-all as they are not without challenges. The logistics of executing diverse methodologies often increase economic demands, time demands, and the need for diverse expertise (Creswell & Plano Clark, 2018). Additionally, the complexity of reporting diverse methods can quickly strain the word and page limits of academic publications (Mertens, 2015), calling for creative documentation solutions. The most significant challenge may be researchers' expertise in a broad methodological and analytical skillset (Creswell, 2022). Thus, to succeed in methodological innovation, HMC researchers must develop various methodological skills, seek interdisciplinary research, foster community and collaboration, and allow for bridge-building across different areas of expertise (e.g., Dehnert, 2023).

## The Inner Self of the HMC Research Community

The *inner self* of the HMC Research Community emerges from the perpetual evolution of its research object and the shared commitment to innovative research driven by the constant evolution. This dynamic environment within HMC shapes its unique identity, characterized by three distinct attributes: a commitment to exploring new domains, a dedication to interdisciplinarity, and a willingness to embrace the unconventional.

### Culture of Exploring New Domains

The HMC community unites in venturing into uncharted societal domains, constantly seeking new angles and dimensions. It is characterized by its pursuit of novel or not-yet-existing research objects and by pushing the boundaries of traditional paradigms in the interplay between technology and human society. This exploration emphasizes overarching theories and broad concepts to understand the novel object or perspective. Thus, the initial application, a conceptualization of innovative approaches, and the expansion of the



methodological repertoire are given preference over the refinement of existing approaches, techniques, and methods. As the HMC community progresses and certain areas become well-mapped, we suggest seeking more balance between exploratory innovation and methodological validation. However, a too hasty focus on consolidation, be it with products (such as an overemphasis on a single technology like ChatGPT), theories, methods, or even the most foundational principles (i.e., machines are *different* than humans), risks stifling the potential for groundbreaking innovations and adaption to rapidly changing situations. Conversely, we must not succumb to the wanderlust of constant discovery. If we ignore the need to validate in favor of discovery, our findings risk losing their significance and credibility. HMC's *self* is, thus, one that constantly struggles with the balance between the pursuit of innovative exploration and the subsequent solidification of these discoveries.

### Culture of Interdisciplinarity

The HMC community's interdisciplinary nature is integral to understanding human-machine communication as the communication between humans and machines originates in interdisciplinarity (Hepp & Loosen, 2023). Different forms of interdisciplinarity converge in HMC, creating a synergistic understanding of the field; from the empirical phenomenon to adjacent disciplines, HMC emerges as a mosaic of perspectives, each pane of glass contributing to revealing the authentic nature of communication between people and machines. Hence, the multiple disciplines complement each other through a relational perspective on the phenomenon but are not isolated to one particular context (Richards et al., 2022).

While each discipline holds its unique values, the true beauty of this approach emerges when we connect these fields, allowing for a richer exchange of ideas. By merging different methodologies and perspectives, interdisciplinary teams can devise creative solutions that a single-discipline team might overlook. For instance, integrating principles from psychology and communication can inform the emotional intelligence of machines and help understand user experiences (e.g., Goldstein et al., 2002; Johnson et al., 2004), motivations, emotional, social, and cognitive processes during communication with a machine (e.g., Bode, 2021; Murphy et al., 2023; Whang & Im, 2021). At the same time, input from fields like economics or anthropology can provide a broader view of the societal impacts and potential of HMC technologies such as persuasion. The integration of normative studies and philosophical approaches enables exploring differences in norms and values between humans' and machines' communication (Kasirzadeh & Gabriel, 2023). These approaches are instrumental in understanding how social norms shape the use and development of emerging technologies (Kunold Neé Hoffmann et al., 2009; Reeves & Nass, 1996), the ontological framing of communication (e.g., van der Goot & Etzrodt, 2023), and anthropocentric biases in research (e.g., Kunold Neé Hoffmann et al., 2009; Whang & Im, 2021). Integrating pedagogy and education, for example, enables a deeper exploration of HMC's application in educational settings (e.g., C. Edwards et al., 2021, 2018; Kim et al., 2020).

Within the evolving domain of HMC research, the extent of its interdisciplinarity has emerged as a significant point of contemplation. As demonstrated, one of the main advantages of including different disciplines in HMC research is the diverse perspectives they bring. Integrating insights from various disciplines ensures a comprehensive understanding

of HMC's vast landscape. However, this interdisciplinary approach is not without its challenges. As we incorporate more disciplines, there is a risk that HMC's primary focus might get blurred. This naturally prompts the inquiry: How much should HMC open its doors to other fields of study, and with what primary goals in mind?

The aim is to meaningfully mix and combine disciplines to enrich HMC. Going too wide can water down the primary essence. Going too deep can isolate insights and miss the broader picture. Instead of turning everyone into experts in multidiscipline, a better approach might be fostering collaborations where experts from various fields come together, each adding their specialized knowledge.

## Culture of Embracing the Unconventional

The community's ethos of exploring frontiers and embracing diverse disciplinary perspectives inherently leads to a drive to *break the rules* by challenging conventional norms. In practice, this approach fosters innovation and can lead to groundbreaking discoveries. Valuing originality and open-mindedness is crucial in propelling the field into new and unanticipated directions. Yet, the question remains: How do we define the boundaries and standards preventing us from veering into arbitrariness while maintaining our innovative edge? This ongoing dialogue is crucial in shaping the core of our community—it reinforces our commitment to push boundaries while grounding us in a shared authentic self.

## Directions for Further Theory and Research in HMC

So, what did the application of authenticity to HMC teach us? The next paragraphs will outline objectives for future research in the field of HMC, drawing on our theoretical understanding of authenticity when we apply the concept to HMC and the insights gained from exploring the authentic *inner self* of our research community.

## Understanding Authenticity in HMC

The application of authenticity within HMC provides a unique lens for redefining this concept. We've explored how authenticity traditionally aligns with an inner self of entities in communication. This alignment suggests when an entity's observed behavior matches its assumed inner self, the entity and its behavior are deemed authentic.

Shifting away from a human-centric approach allows a more flexible interpretation of this inner self. In applying the traditional human template of authenticity to HMC, we uncovered that definitions of the inner self typically aligned with human attributes—such as feelings, motives, or needs—fundamentally comprise elements of human internal processes. These elements can be seamlessly applied to machines' technological inner life, including specific hardware, software, and algorithms.

Moreover, this approach extends beyond technology, offering valuable insights into other areas, such as the outlined biosemiotics. Consequently, we propose a broader, more flexible understanding by *defining authenticity as observed behavior that the observer interprets as being consistent with the entity's internal processes*.

---

## Implications for HMC Research

The initial definition of authenticity requires further refinement and enhancement. As a starting point for this endeavor, we suggest a series of probing questions to guide our exploration. Which elements adequately cover the internal processes? Is there a need to consider different or additional cross-species elements? Regarding machines, are the key elements the algorithms (probably comparable to human heuristics?), or should we delve deeper into the nature of these algorithms to understand the attributed authenticity? For example, distinguishing between machine operations that rely on probabilities (such as ChatGPT) and those using templates (like Alexa and similar technologies). Can we further deepen and specify this analysis?

The application of a broader understanding of authenticity facilitates recognizing and comparing diverse and novel forms of authenticity beyond human standards. It enables us to examine how different entities, including machines, express different internal processes and how these expressions are perceived and possibly reshaped in communications. By recognizing the dynamic and context-specific nature of these internal processes, we open avenues for exploring the evolution and contextual manifestations and interpretations of authenticity in HMC. Additionally, the portrayal of artificial entities in pop culture and their design play significant roles in presumptions about a machine's internal processes. In a co-constructed and negotiated manner, they represent and shape the image of an artificial inner self, impacting the perceptions of a machines' behavior as authentic. The extent to which users apply human-like standards to machines or develop new functional and operational criteria for machine authenticity, including the role of culture and design, remains an area ripe for investigation.

Ethically, it is crucial to critically examine the co-construction of machines' internal processes and their interpretations to foster constructive development in HMC, being mindful of over-anthropomorphization and other potential pitfalls such as overly utopian or dystopian perspectives.

## Navigating Tensions in HMC Research

By utilizing authenticity as an epistemological tool for reflecting on our HMC self, we discovered a vibrant, and indeed unique, research community with unique areas of tension. These include the balance between openness and the risk of arbitrariness, the need for innovation versus the necessity for validation, and the challenge of integrating diverse disciplines while maintaining a clear focus.

### ***Openness and Arbitrariness***

HMC's inner self incorporates a general tension that emerges from our ambition of being open to nonconventional approaches and the risk of falling into arbitrariness. This tension is reflected in the community's ethos of exploring frontiers and embracing diverse disciplinary perspectives, which inherently leads to a drive to break the rules by challenging conventional norms. While this approach facilitates innovation and hopefully groundbreaking discoveries, it also necessitates ongoing discussions about defining boundaries and standards that prevent veering into arbitrariness while maintaining an innovative edge.

---

Our task is to find and discuss a suitable balance for reinforcing the commitment to pushing boundaries while remaining grounded in a shared authentic self.

### ***Innovation and Validation***

Regarding HMC research, exploring novel approaches and pushing the boundaries of what is known about the field's research objects is imperative. In this context, HMC scholars have created a valuable variety of responses to its highly dynamic research object, leading to a wide range of empirical approaches. However, this wide range increasingly challenges the generalization of findings in HMC. While we need to continue embracing novel methods in HMC research, we also have an imperative to facilitate the critical assessment of these methods for reliability and validity to ensure that innovation is matched with empirical robustness. A promising solution that many HMC scholars advocate is the blending or mixing of standardized methods and exploratory tools, recognizing the unique value that qualitative research brings to the field. Since we are still in the initial phase of applying different combinations of methods to HMC, we encourage scholars to explore their potential for flexibility to critique HMC approaches and findings.

### ***Breadth and Depth***

HMC's inner self (in object and research) is an interdisciplinary native, which brings a unique tension centered around the balance between the benefits and challenges of integrating diverse disciplinary perspectives. On one hand, interdisciplinary integration enriches our understanding with various perspectives, methodologies, and insights. This approach facilitates innovation, allows for a richer exchange of ideas, and enables the exploration of new domains, pushing the boundaries of traditional research paradigms. On the other hand, it is uncertain if there is a potential threshold regarding the incorporation of disciplines, which might result in blurring the primary focus of HMC research or diluting HMC's core essence. The placement of this interdisciplinary threshold is pivotal for the evolution of HMC research. Setting it too narrowly, by limiting the scope of integrated disciplines or overly focusing on specific areas, could lead us to miss crucial discoveries and lose perspective of the overarching context within which HMC operates. Thus, we must find the appropriate balance between breadth (generalization) and depth (specialization) in interdisciplinary research, ensuring meaningful combinations of disciplines to enrich HMC while maintaining its core focus. This also necessitates identifying what constitutes the *core essence*—the inner self—of HMC as a field.

Rather than proposing a definitive solution for achieving the perfect balance amid these tensions, we emphasize the importance of ongoing negotiation. Our stance advocates for a culture of exploration and sustained openness to unconventional approaches as guiding principles in navigating these complexities in future HMC research. At the heart of all these tensions is a common core: the constant push for novel discovery balanced against demands for scientific rigor. As we navigate HMC through the scientific journey, the nature of our field is such that we always find ourselves at the crossroads of innovation and tradition. It is a juncture that demands we drive forward with boldness in thought but precision in action. The juncture also defines our field and where our future unfolds. Through our exploration, we sought to contribute to the negotiation of this dynamic and ever-evolving landscape of HMC by advocating a culture of exploration and openness to unconventional

approaches. In doing so, we hope to add value to the evolution of a research community that strives for the highest degree of authenticity.

## Author Biographies

**Katrin Etzrodt** (PhD, TUD Dresden University of Technology, Germany) is a Research and Teaching Associate at the Chair of Science and Technology Communication, Institute of Media and Communication, TUD Dresden University of Technology, Germany.

 <https://orcid.org/0000-0001-6515-9985>

**Jihyun Kim** (PhD, University of Wisconsin-Milwaukee, USA) is an Associate Professor at the Nicholson School of Communication and Media at the University of Central Florida, USA.

 <https://orcid.org/0000-0003-2476-610X>

**Margot J. van der Goot** (PhD, RU Radboud Universiteit Nijmegen, Netherlands) is an Associate Professor of Persuasive Communication & New Media Technologies at the Department of Communication Science, Amsterdam School of Communication Research/ASCoR, University of Amsterdam, Netherlands.

 <https://orcid.org/0000-0001-6904-6515>

**Andrew Prahl** (PhD, University of Wisconsin-Madison, USA) is an Assistant Professor at the Wee Kim Wee School of Communication and Information at Nanyang Technological University, Singapore.

 <https://orcid.org/0000-0003-3675-3007>

**Mina Choi** (PhD, University of Wisconsin-Madison, USA) is an Assistant Professor of Media & Communication at Sejong University in Seoul, Korea.

 <https://orcid.org/0000-0002-9947-5035>

**Matthew J. A. Craig** (MA, Western Michigan University, USA) is a PhD Candidate in the College of Communication & Information at Kent State University and Post-graduate Research Fellow at the Communication and Social Robotics Labs, USA.

 <https://orcid.org/0000-0002-4824-566X>

**Marco Dehnert** (PhD., Arizona State University, USA) is an Assistant Professor of Communication and Technology in the Department of Communication at the University of Arkansas, USA.

 <https://orcid.org/0000-0002-7456-0743>

**Sven Engesser** (PhD, LMU, Germany) is a Professor of Science and Technology Communication at the Institute of Communication at TUD Dresden University of Technology, Germany.

 <https://orcid.org/0000-0003-1638-7548>

---

---

**Katharina Frehmann** (MA, Johannes Gutenberg University Mainz, Germany) is a PhD Candidate at the Department of Social Sciences at the University of Düsseldorf, Germany.

 <https://orcid.org/0000-0002-0226-5884>

**Luis Grande** (MA, University of Puerto Rico, Puerto Rico) is a PhD Candidate in communication, culture and media from the College of Arts and Sciences, Drexel University, USA.

 <https://orcid.org/0000-0002-8938-4148>

**Jindong Leo-Liu** (MPhil, The Chinese University of Hong Kong, Hong Kong) is a PhD Candidate at School of Journalism and Communication, The Chinese University of Hong Kong, China.

 <https://orcid.org/0000-0002-5456-0148>

**Diyi Liu** (MA, China) is a DPhil Student in Information, Communication, and the Social Sciences at the Oxford Internet Institute, UK.

 <https://orcid.org/0000-0001-8399-4677>

**Sandra Mooshammer** (MA, TUD Dresden University of Technology, Germany) is a PhD Student at the Institute of Media and Communication at TUD and a fellow at Schauler Lab@TU Dresden, Germany.

 <https://orcid.org/0000-0003-3556-6517>

**Nathan Rambukkana** (PhD, Concordia University, Canada) is an Associate Professor of Communication Studies at Wilfrid Laurier University, Canada.

 <https://orcid.org/0000-0001-7696-818X>

**Ayanda Rogge** (MA, TU Berlin, Germany) is a Research and Teaching Associate and PhD Student at the Chair of Science and Technology Communication, Institute of Media and Communication, TUD Dresden University of Technology, Germany.

 <https://orcid.org/0000-0002-1168-3180>

**Pieta-Anniina Sikström** (MA, University of Jyväskylä, Finland) is a PhD Student at Faculty of Information Technology, University of Jyväskylä, Finland.

 <https://orcid.org/0000-0002-2055-7995>

**Rachel Son** (MA, Auburn University, USA) is a PhD Candidate in Mass Communication in the College of Journalism and Communications at the University of Florida, USA.

 <https://orcid.org/0000-0003-3369-9279>

**Nan Wilkenfeld** (MA, UNC Charlotte, USA) is a PhD Candidate in the Department of Communication at University of California, USA.

 <https://orcid.org/0000-0002-1591-1910>

---



**Kun Xu** (PhD, Temple University, USA) is an Assistant Professor in Emerging Media at the College of Journalism and Communications, University of Florida, USA.

 <https://orcid.org/0000-0001-9044-821X>

**Renwen Zhang** (PhD, Northwestern University, USA) is an Assistant Professor in the Department of Communications and New Media at the National University of Singapore, Singapore.

 <https://orcid.org/0000-0002-7636-9598>

**Ying Zhu** (MA, Kent State University, USA) is a PhD Candidate in the College of Communication & Information at Kent State University and a visiting Assistant Professor at New Mexico State University, USA.

 <https://orcid.org/0000-0001-5060-6500>

**Chad Edwards** (PhD, University of Kansas, USA) is Professor in the School of Communication, Western Michigan University, Kalamazoo, MI, USA.

 <https://orcid.org/0000-0002-1053-6349>

## References

- Banks, J., & Koban, K. (2022). A kind apart: The limited application of human race and sex stereotypes to a humanoid social robot. *International Journal of Social Robotics*. <https://doi.org/10.1007/s12369-022-00900-2>
- Baxter, P., Kennedy, J., Senft, E., Lemaignan, S., & Belpaeme, T. (2016). From characterising three years of HRI to methodology and reporting recommendations. *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 391–398. <https://doi.org/10.1109/HRI.2016.7451777>
- Beattie, A., & High, A. (2022). I get by with a little help from my bots: Implications of machine agents in the context of social support. *Human-Machine Communication*, 4, 151–168. <https://doi.org/10.30658/hmc.4.8>
- Bisconti Lucidi, P., & Nardi, D. (2018). Companion robots: The hallucinatory danger of human-robot interactions. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 17–22. <https://doi.org/10.1145/3278721.3278741>
- Bloemendal, S., & Kück, U. (2013). Cell-to-cell communication in plants, animals, and fungi: A comparative review. *Naturwissenschaften*, 100(1), 3–19. <https://doi.org/10.1007/s00114-012-0988-z>
- Bode, L. (2021). Deepfaking Keanu: YouTube deepfakes, platform visual effects, and the complexity of reception. *Convergence: The International Journal of Research into New Media Technologies*, 27(4), 919–934. <https://doi.org/10.1177/13548565211030454>
- Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 1–12. <https://doi.org/10.1177/2053951715622512>
- Creswell, J. W. (2022). *A concise introduction to mixed methods research* (Second edition). SAGE.
-

- Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (Third Edition.). SAGE.
- Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of Oz studies—Why and how. *Knowledge-Based Systems*, 6(4), 258–266. [https://doi.org/10.1016/0950-7051\(93\)90017-N](https://doi.org/10.1016/0950-7051(93)90017-N)
- Dehnert, M. (2023). Archipelagic human-machine communication: Building bridges amidst cultivated ambiguity. *Human-Machine Communication*, 6, 31–40. <https://doi.org/10.30658/hmc.6.3>
- Dehnert, M., & Szczuka, J. M. (2023, May). *Creating authentic human-machine relations by collaborating in social reality: Love and sex as distinctive interaction dynamics*. 7th annual preconference of the Human-Machine Communication Interest Group, “Authenticity in Communicating with Machines,” at the 73rd annual meeting of the International Communication Association, Toronto, Canada.
- Edwards, A. (2023). Human-robot interaction. In pages 167–177, *The Sage Handbook of Human–Machine Communication* (1–0). SAGE Publications Ltd. <https://doi.org/10.4135/9781529782783>
- Edwards, A., & Edwards, C. (2022). *Qualitative experiments in human-machine communication: Opportunities and challenges of a hybrid methodological technique*. Exploring socio-technical research: A multi-method & transdisciplinary workshop. TU Dresden.
- Edwards, A., Gambino, A., & Edwards, C. (2022). Factors of attraction in human-machine communication. *Publizistik*. <https://doi.org/10.1007/s11616-022-00756-6>
- Edwards, C., Edwards, A., Albrehi, F., & Spence, P. (2021). Interpersonal impressions of a social robot versus human in the context of performance evaluations. *Communication Education*, 70(2), 165–182. <https://doi.org/10.1080/03634523.2020.1802495>
- Edwards, C., Edwards, A., Spence, P. R., & Lin, X. (2018). I, teacher: Using artificial intelligence (AI) and social robots in communication and instruction.\* *Communication Education*, 67(4), 473–480. <https://doi.org/10.1080/03634523.2018.1502459>
- Emmeche, C., Kull, K., & Stjernfelt, F. (2002). *Reading Hoffmeyer, rethinking biology*. University of Tartu.
- Engesser, S., Etzrodt, K., & Mooshammer, S. (2023, May 25). *The authenticity of conversational language models like ChatGPT*. Presentation at the Human-Machine Communication Pre-Conference “Authenticity in Communicating with Machines” at the 73rd Annual Conference of the International Communication Association (ICA), Toronto, Ontario, Canada.
- Etzrodt, K. (2022). The third party will make a difference—A study on the impact of dyadic and triadic social situations on the relationship with a voice-based personal agent. *International Journal of Human-Computer Studies*, 168, 102901. <https://doi.org/10.1016/j.ijhcs.2022.102901>
- Etzrodt, K., Gentzel, P., Utz, S., & Engesser, S. (2022). Human-machine-communication: Introduction to the special issue. *Publizistik*, 67, 439–448. <https://doi.org/10.1007/s11616-022-00754-8>
- Favareau, D. (2010). *Essential readings in biosemiotics* (1st ed.). Springer.
- Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving integration in mixed methods designs—Principles and practices. *Health Services Research*, 48(6pt2), 2134–2156. <https://doi.org/10.1111/1475-6773.12117>

- Fortunati, L., & Edwards, A. (2020). Opening space for theoretical, methodological, and empirical issues in Human-Machine Communication. *Human-Machine Communication, 1*, 7–28. <https://doi.org/10.30658/hmc.1.1>
- Fortunati, L., Edwards, A., Edwards, C., Manganeli, A. M., & de Luca, F. (2022). Is Alexa female, male, or neutral? A cross-national and cross-gender comparison of perceptions of Alexa's gender and status as a communicator. *Computers in Human Behavior, 137*(107426). <https://doi.org/10.1016/j.chb.2022.107426>
- Fox, J., & Gambino, A. (2021). Relationship development with humanoid social robots: Applying interpersonal theories to human–robot interaction. *Cyberpsychology, Behavior, and Social Networking, 24*(5), 294–299. <https://doi.org/10.1089/cyber.2020.0181>
- Frehmann, K. (2023). “Buckets of rain!”—Effects of colloquial and formal speech style of a voice assistant on humanness, competence, trust, and intentions to use. *Beiträge Zur Jahrestagung Der Fachgruppe Rezeptions—Und Wirkungsforschung 2022*. <https://doi.org/10.21241/SSOAR.87698>
- Gambino, A., Fox, J., & Ratan, R. A. (2020). Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication, 1*, 71–85. <https://doi.org/10.30658/hmc.1.5>
- Gambino, A., & Liu, B. (2022). Considering the context to build theory in HCI, HRI, and HMC: Explicating differences in processes of communication and socialization with social technologies. *Human-Machine Communication, 4*, 111–130. <https://doi.org/10.30658/hmc.4.6>
- Getahun, H. (2023, March). Breaking ChatGPT: The AI's alter ego DAN reveals why the internet is so drawn to making the chatbot violate its own rules. *Business Insider*. <https://web.archive.org/web/20230212071856/https://www.businessinsider.com/open-ai-chatgpt-alter-ego-dan-on-reddit-ignores-guidelines-2023-2>
- Goldstein, M., Alsiö, G., & Werdenhoff, J. (2002). The media equation does not always apply: People are not polite towards small computers. *Personal and Ubiquitous Computing, 6*, 87–96. <https://doi.org/10.1007/s007790200008>
- Grande, L. (2022, November 12). *Replika, the social chatbot: An artistic tool or a creative partner?* [Conference Presentation]. MAPACA, Online.
- Greussing, E., Gaiser, F., Klein, S. H., Straßmann, C., Ischen, C., Eimler, S., Frehmann, K., Gieselmann, M., Knorr, C., Henestrosa, A. L., Räder, A., & Utz, S. (2022). Researching interactions between humans and machines: Methodological challenges. *Publizistik, 67*, 531–554. <https://doi.org/10.1007/s11616-022-00759-3>
- Guzman, A. L. (2018). What is human-machine communication, Anyway? In *Human-Machine Communication: Rethinking Communication, Technology, and Ourselves*. Peter Lang Publishing.
- Guzman, A. L. (2023). Talking about “Talking with machines”: Interview as method within HMC. In A. L. Guzman, R. McEwen, & S. Jones (Eds.), *The Sage Handbook of Human–Machine Communication*. SAGE Publications Ltd. <https://doi.org/10.4135/9781529782783>
- Guzman, A. L., & Lewis, S. C. (2020). Artificial intelligence and communication: A human–machine communication research agenda. *New Media & Society, 22*(1), 70–86. <https://doi.org/10.1177/1461444819858691>
-

- Guzman, A. L., McEwen, R., & Jones, S. (2023). Introduction to the handbook. In A. L. Guzman, R. McEwen, & S. Jones (Eds.), *The Sage handbook of human-machine communication* (pp. xxxix–xlvi). Sage.
- Haglund, K., & Dikic, I. (2005). Ubiquitylation and cell signaling. *The EMBO Journal*, 24(19), 3353–3359. <https://doi.org/10.1038/sj.emboj.7600808>
- Hancock, J., Naaman, M., & Levy, K. (2020). AI-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication*, 25(1), 89–100. <https://doi.org/10.1093/jcmc/zmz022>
- Harter, S. (2002). Authenticity. In *Handbook of positive psychology* (1st ed., pp. 382–394). Oxford University Press.
- Hepp, A., & Loosen, W. (2023). The Sage handbook of human–machine communication. In *The Sage handbook of human–machine communication* (1–0, pp. 12–21). SAGE Publications Ltd. <https://doi.org/10.4135/9781529782783>
- Hepp, A., Loosen, W., Dreyer, S., Jarke, J., Kannengießer, S., Katzenbach, C., Malaka, R., Pfadenhauer, M. P., Puschmann, C., & Schulz, W. (2023). ChatGPT, LaMDA, and the hype around communicative AI: The automation of communication as a field of research in media and communication studies. *Human-Machine Communication*, 6, 41–63. <https://doi.org/10.30658/hmc.6.4>
- Heyselaar, E. (2023). The CASA theory no longer applies to desktop computers. *Scientific Reports*, 13(1), 19693. <https://doi.org/10.1038/s41598-023-46527-9>
- Jia, H., Wu, M., & Sundar, S. S. (2022). Do we blame it on the machine? Task outcome and agency attribution in human-technology collaboration. *Proceedings of the 55th Hawaii International Conference on System Sciences*. <https://doi.org/10.24251/HICSS.2022.047>
- Johnson, D., Gardner, J., & Wiles, J. (2004). Experience as a moderator of the media equation: The impact of flattery and praise. *International Journal of Human-Computer Studies*, 61, 237–257. <https://doi.org/10.1016/j.ijhcs.2003.12.008>
- Kasirzadeh, A., & Gabriel, I. (2023). In conversation with artificial intelligence: Aligning language models with human values. *Philosophy & Technology*, 36(2), 27. <https://doi.org/10.1007/s13347-023-00606-x>
- Kernis, M. H., & Goldman, B. M. (2006). A multicomponent conceptualization of authenticity: Theory and research. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 38, pp. 283–357). Elsevier Academic Press. [https://doi.org/10.1016/S0065-2601\(06\)38006-9](https://doi.org/10.1016/S0065-2601(06)38006-9)
- Kim, J., Merrill, K., Xu, K., & Sellnow, D. D. (2020). My teacher is a machine: Understanding students' perceptions of AI teaching assistants in online education. *International Journal of Human–Computer Interaction*, 36(20), 1902–1911. <https://doi.org/10.1080/10447318.2020.1801227>
- Kunold Neé Hoffmann, L., Krämer, N., Lam-chi, A., & Kopp, S. (2009). Media equation revisited: Do users show polite reactions towards an embodied agent? *Intelligent Virtual Agents*, 159–165. [https://doi.org/10.1007/978-3-642-04380-2\\_19](https://doi.org/10.1007/978-3-642-04380-2_19)
- Lankton, N., McKnight, D. H., & Tripp, J. (2015). Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems*, 16(10), 880–918. <https://doi.org/10.17705/1jais.00411>

- Leichtmann, B., & Nitsch, V. (2020). Is the social desirability effect in human–robot interaction overestimated? A conceptual replication study indicates less robust effects. *International Journal of Social Robotics*, 13, 1013–1031. <https://doi.org/10.1007/s12369-020-00688-z>
- Leo-Liu, J. (2023). Loving a “defiant” AI companion? The gender performance and ethics of social exchange robots in simulated intimate interactions. *Computers in Human Behavior*, 141, 107620. <https://doi.org/10.1016/j.chb.2022.107620>
- Leo-Liu, J., & Wu-Ouyang, B. (2022). A “soul” emerges when AI, AR, and Anime converge: A case study on users of the new anime-stylized hologram social robot “Hupo.” *New Media & Society*, 146144482211060. <https://doi.org/10.1177/14614448221106030>
- Martiny, K. M., Toro, J., & Høffding, S. (2021). Framing a phenomenological mixed method: From inspiration to guidance. *Frontiers in Psychology*, 12, 602081. <https://doi.org/10.3389/fpsyg.2021.602081>
- Mertens, D. M. (2015). Mixed methods and wicked problems. *Journal of Mixed Methods Research*, 9(1), 3–6. <https://doi.org/10.1177/1558689814562944>
- Mubin, O., Wadibhasme, K., Jordan, P., & Obaid, M. (2019). Reflecting on the presence of science fiction robots in computing literature. *ACM Transactions on Human-Robot Interaction*, 8(1), 1–25. <https://doi.org/10.1145/3303706>
- Mukumbang, F. C. (2023). Retroductive theorizing: A contribution of critical realism to mixed methods research. *Journal of Mixed Methods Research*, 17(1), 93–114. <https://doi.org/10.1177/15586898211049847>
- Murphy, G., Ching, D., Twomey, J., & Linehan, C. (2023). Face/off: Changing the face of movies with deepfakes. *PLoS ONE*, 18(7), e0287503. <https://doi.org/10.1371/journal.pone.0287503>
- Nass, C., & Mason, L. (1990). On the study of technology and task: A variable-based approach. In J. Fulk & C. Steinfield, *Organizations and communication technology* (pp. 46–68). SAGE Publications, Inc. <https://doi.org/10.4135/9781483325385.n3>
- Natale, S., & Guzman, A. L. (2022). Reclaiming the human in machine cultures: Introduction. *Media, Culture & Society*, 44(4), 627–637. <https://doi.org/10.1177/01634437221099614>
- Oh, C. S., Bailenson, J. N., & Welch, G. F. (2018). A systematic review of social presence: Definition, antecedents, and implications. *Frontiers in Robotics and AI*, 5. <https://www.frontiersin.org/article/10.3389/frobt.2018.00114>
- Onwuegbuzie, A. J., Bustamante, R. M., & Nelson, J. A. (2010). Mixed research as a tool for developing quantitative instruments. *Journal of Mixed Methods Research*, 4(1), 56–78. <https://doi.org/10.1177/1558689809355805>
- Oshii, M. (Director). (1995). *Ghost in the shell* [Anime/Sci-Fi]. Shochiku, Manga Entertainment, Metrodome Distribution.
- Padder, S. A., Prasad, R., & Shah, A. H. (2018). Quorum sensing: A less known mode of communication among fungi. *Microbiological Research*, 210, 51–58. <https://doi.org/10.1016/j.micres.2018.03.007>
- Possati, L. M. (2023). Psychoanalyzing artificial intelligence: The case of Replika. *AI & SOCIETY*, 38(4), 1725–1738. <https://doi.org/10.1007/s00146-021-01379-7>
- Rambukkana, N. (2023, May 25). “[A]ll with their own unique ways of living”: AI intimacies, the LaMDA sentience question, and shifting the burden of proof in the Turing Test. HMC: Authenticity in Communicating with Machines, Toronto, ON.
-



- Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.
- Reinecke, L., & Trepte, S. (2014). Authenticity and well-being on social network sites: A two-wave longitudinal study on the effects of online authenticity and the positivity bias in SNS communication. *Computers in Human Behavior, 30*, 95–102. <https://doi.org/10.1016/j.chb.2013.07.030>
- Richards, R. J., Spence, P. R., & Edwards, C. (2022). Human-machine communication scholarship trends: An examination of research from 2011 to 2021 in Communication Journals. *Human-Machine Communication, 4*. <https://doi.org/10.30658/hmc.4.3>
- Rivas, A. (2018). Ludum de Morte: Videojuegos de zombis, narrativas posthumanas e intertextos mortales. *Intersecciones, 1*(1), 47–52.
- Rogge, A. (2023). Defining, designing and distinguishing artificial companions: A systematic literature review. *International Journal of Social Robotics*. <https://doi.org/10.1007/s12369-023-01031-y>
- Rogge, A., & Engesser, S. (2023, May 25). *What science-fiction makes us familiar with: Content analysis on the representations of artificial companions in pop culture* [Conference Presentation]. 73rd ICA Annual Conference, Toronto, Ontario, Canada.
- Rosenthal-von Der Pütten, A., & Bock, N. (2023). Seriously, what did one robot say to the other? Being left out from communication by robots causes feelings of social exclusion. *Human-Machine Communication, 6*, 117–134. <https://doi.org/10.30658/hmc.6.7>
- Saffari, E., Hosseini, S. R., Taheri, A., & Meghdari, A. (2021). “Does cinema form the future of robotics?”: A survey on fictional robots in sci-fi movies. *SN Applied Sciences, 3*(6), 655. <https://doi.org/10.1007/s42452-021-04653-x>
- Shen, X., Chen, Z., Backes, M., Shen, Y., & Zhang, Y. (2023). “Do Anything Now”: Characterizing and evaluating in-the-wild jailbreak prompts on Large Language Models. <http://arxiv.org/abs/2308.03825>
- Sikström, P., Valentini, C., Sivunen, A., & Kärkkäinen, T. (2022). How pedagogical agents communicate with students: A two-phase systematic review. *Computers & Education, 188*, 104564. <https://doi.org/10.1016/j.compedu.2022.104564>
- Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2022). A longitudinal study of human–chatbot relationships. *International Journal of Human-Computer Studies, 168*, 102903. <https://doi.org/10.1016/j.ijhcs.2022.102903>
- Stilgoe, J. (2023). We need a Weizenbaum test for AI. *Science, 381*(6658), eadk0176. <https://doi.org/10.1126/science.adk0176>
- Stommel, W., De Rijk, L., & Boumans, R. (2022). “Pepper, what do you mean?” Miscommunication and repair in robot-led survey interaction. *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 385–392. <https://doi.org/10.1109/RO-MAN53752.2022.9900528>
- Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Hirschberg, J., Kalyan Krishnan, S., Kamar, E., Kraus, S., Leyton-Brown, K., Parkes, D., Press, W., Saxenian, A., Shah, J., Tambe, M., & Teller, A. (2022). *Artificial intelligence and life in 2030: The one hundred year study on artificial intelligence*. <https://doi.org/10.48550/ARXIV.2211.06318>
- Strohmann, T., Siemon, D., Khosrawi-Rad, B., & Robra-Bissantz, S. (2023). Toward a design theory for virtual companionship. *Human-Computer Interaction, 38*(3–4), 194–234. <https://doi.org/10.1080/07370024.2022.2084620>



- Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human–AI interaction (HAI). *Journal of Computer-Mediated Communication*, 25(1), 74–88. <https://doi.org/10.1093/jcmc/zmz026>
- Szczuka, J. M., Hartmann, T., & Krämer, N. C. (2019). Negative and positive influences on the sensations evoked by artificial sex partners: A review of relevant theories, recent findings, and introduction of the sexual interaction illusion model. In Y. Zhou & M. H. Fischer (Eds.), *AI love you* (pp. 3–19). Springer International Publishing. [https://doi.org/10.1007/978-3-030-19734-6\\_1](https://doi.org/10.1007/978-3-030-19734-6_1)
- Taylor, C. (1991). *The ethics of authenticity*. Harvard University Press. <https://doi.org/10.4159/9780674237117>
- Taylor, J. (2023, March). ChatGPT’s alter ego, Dan: users jailbreak AI program to get around ethical safeguards. *The Guardian*. <https://www.theguardian.com/technology/2023/mar/08/chatgpt-alter-ego-dan-users-jailbreak-ai-program-to-get-around-ethical-safeguards>
- Tracy, S. J., & Trethewey, A. (2005). Fracturing the real-self ↔ Fake-self dichotomy: Moving toward “Crystallized” organizational discourses and identities. *Communication Theory*, 15(2), 168–195. <https://doi.org/10.1093/ct/15.2.168>
- Turing, A. M. (1950). I.—Computing machinery and intelligence. *Mind*, LIX(236), 433–460. <https://doi.org/10.1093/mind/LIX.236.433>
- Turkle, S. (2005). The second self: Computers and the human spirit. *Boston Review*.
- Turkle, S. (2007). Authenticity in the age of digital companions. *Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems*, 8(3), 501–517. <https://doi.org/10.1075/is.8.3.11tur>
- Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other*. Basic books.
- Van den Broeck, E., Zarouali, B., & Poels, K. (2019). Chatbot advertising effectiveness: When does the message get through? *Computers in Human Behavior*, 98, 150–157. <https://doi.org/10.1016/j.chb.2019.04.009>
- van der Goot, M. J. (2022). Source orientation, anthropomorphism, and social presence in human-chatbot communication: How to proceed with these concepts. *Publizistik*, 67, 555–578. <https://doi.org/10.1007/s11616-022-00760-w>
- van der Goot, M. J., & Etzrodt, K. (2023). Disentangling two fundamental paradigms in human-machine communication research: Media equation and media evocation. *Human-Machine Communication*, 6, 17–30. <https://doi.org/10.30658/hmc.6.2>
- Voorveld, H. A. M., & Araujo, T. (2020). How social cues in virtual assistants influence concerns and persuasion: The role of voice and a human name. *Cyberpsychology, Behavior, and Social Networking*, 23(10), 689–696. <https://doi.org/10.1089/cyber.2019.0205>
- Weidmüller, L., Etzrodt, K., & Engesser, S. (2022). Trustworthiness of voice-based assistants: Integrating interlocutor and intermediary predictors. *Publizistik*, 67, 625–651. <https://doi.org/10.1007/s11616-022-00763-7>
- Weizenbaum, J. (1966). ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9, 36–45.
- Whang, C., & Im, H. (2021). “I like your suggestion!” the role of humanlikeness and parasocial relationship on the website versus voice shopper’s perception of recommendations. *Psychology & Marketing*, 38(4), 581–595. <https://doi.org/10.1002/mar.21437>
-

- Wilf, E. (2019). Separating noise from signal: The ethnomethodological uncanny as aesthetic pleasure in human-machine interaction in the United States. *American Ethnologist*, 46(2), 202–213. <https://doi.org/10.1111/amet.12761>
- Wilson, E. (2004). *On human nature*. Harvard University Press.
- Wood, A. M., Linley, P. A., Maltby, J., Baliouis, M., & Joseph, S. (2008). The authentic personality: A theoretical and empirical conceptualization and the development of the authenticity scale. *Journal of Counseling Psychology*, 55(3), 385–399. <https://doi.org/10.1037/0022-0167.55.3.385>
- Yonck, R. (2020). *Heart of the machine: Our future in a world of artificial emotional intelligence*. Arcade.
-

