

The dark side of AI anthropomorphism: A case of misplaced trustworthiness in service provisions

Rakibul Hasan
University of Vaasa, Finland
rakibul.hasan@uwasa.fi

Arto Ojala
University of Vaasa, Finland
arto.ojala@uwasa.fi

Sara Quach
Griffith University, Australia
s.quach@griffith.edu.au

Park Thaichon
University of Southern Queensland, Australia
park.thaichon@unisq.edu.au

Scott Weaven
Griffith University, Australia
s.weaven@griffith.edu.au

Abstract

Anthropomorphism, the attribution of human-like traits, qualities, and mental states to non-human agents, is increasingly ubiquitous in artificial intelligence (AI) applications within service provisions. While imbuing anthropomorphism into AI agents enhances user interaction, it also has a dark side that poses significant ethical concerns by potentially causing harm to consumers. In response, our study constructs a framework to explain how anthropomorphic features in AI agents can lead to misplaced trustworthiness among consumers. We adopt a sociotechnical perspective and employ a case study design to achieve this. Our findings hope to advance the social sustainability of AI to promote responsible production and consumption patterns in global service markets.

Keywords: Artificial intelligence; service; Anthropomorphism; Misplaced trust; Ethics

1. Introduction

Anthropomorphism is becoming central to human-machine interaction as organizations increasingly induce human-like traits into artificial intelligence (AI)-enabled advanced IT artifacts. These anthropomorphized AI artifacts are becoming ubiquitous in service provisions (Blut et al., 2021; Hasan et al., 2021; Schanke et al., 2021). However, AI agents designed with anthropomorphic features have a dark side that negatively impact service consumption and trustworthiness (Kegel & Stock-Homburg, 2023; Knof et al., 2022).

A dire example of this dark side is highlighted in an investigative report on numerous crashes and fatalities linked to Tesla's autopilot (Marshall, 2024). Specifically, AI agents embedded in Tesla's autopilot that offer autonomous driving capabilities, have misled consumers into overly relying the

system's capabilities, resulting in misplaced trustworthiness. Similarly, a recent AI research on sustainability urges to focus on how consumers anthropomorphize autonomy, which can lead to harmful consequences (Hasan & Ojala, 2024).

Existing literature suggests that imbuing anthropomorphic features, such as autonomy, to AI agents can distort consumer behavior when trustworthiness is misplaced (Gill, 2020; Waytz et al., 2014). Furthermore, service organizations may deliberately utilize anthropomorphic features to exploit or mislead consumers by triggering anthropomorphic reasoning and perception (Jörling et al., 2019; Leong & Selinger, 2019). For instance, organizations can use AI agents to change consumers' attributions of responsibility for service failures. These concerns have motivated a focus on the dark side of AI anthropomorphism.

By taking a consumer-level perspective in service provisions, we problematize trustworthiness to offer valuable insights for information systems and service research (Alvesson & Sandberg, 2011; Bartneck et al., 2021). This urgency to protect consumers' wellbeing from potential physical or psychological harm is paramount. This aligns with the AI Act that is designed to mitigate risks posed by AI agents (European Commission, 2024). In this context, we define service provisions to include consumer interactions with products or services whose functionality is enabled by the AI agents (e.g., Google Search Engine, Microsoft Copilot), as well as AI agents intended to be used as safety components of products or services (e.g., Tesla's Autopilot, Replika). In some case, AI agents may operate at full capacity, while in others, their capabilities may not be fully apparent in their behavior.

The dark side of AI anthropomorphism in service provisions is a critical topic. However, systematic investigations that explore the intersection of anthropomorphic features and misplaced trustworthiness are limited (Bartneck et al., 2021). This is worrisome, as service organizations can potentially cause significant

harms to consumers through deceptive techniques and by exploiting vulnerabilities (Danaher, 2020; Sharkey & Sharkey, 2020). In response, we pose the below research question: *How do anthropomorphic features enable consumers' misplaced trustworthiness in AI agents within service provisions?*

This study makes three significant contributions. First, it constructs a framework to explain how machine-related factors (e.g., hidden or superficial deceptive techniques) and consumer-related factors (e.g., age-related vulnerabilities) in anthropomorphic design features can deceive and manipulate consumers, leading to misplaced trustworthiness. Second, we outline practical implications for service organizations and policymakers to mitigate the risks associated with AI agents that align with the European Commission's AI Act. Finally, we propose pressing future research opportunities on the ethical aspects of AI anthropomorphism that emphasize the need for transparency and an international business perspective.

2. Research background

While AI anthropomorphism is a burgeoning research area, there is limited direct knowledge about its dark side within information systems and service research (Schanke et al., 2021; Jörling et al., 2019). To address this gap, we provide a backdrop of the dark side of AI anthropomorphism through problematization (Alvesson & Sandberg, 2011). We then direct insights towards a sociotechnical thinking of AI (Berente et al., 2021; Hasan & Ojala, 2024). Our abstraction is drawn from existing theoretical and empirical research that relates to service provisions, either implicitly or explicitly. Accordingly, we construct anthropomorphic features that emphasize consumer-level analysis rather than firms-level analysis. We link these features-specific insights to consumer wellbeing. Consequently, our sociotechnical argumentation embodies the ethical aspects that intersect anthropomorphic features of AI (Bartneck et al., 2021; Hasan et al., 2021).

2.1 What is AI anthropomorphism?

From a sociotechnical standpoint, AI is defined as the frontier of computational advancements to address complex decision-making problems with or without human intervention (Berente et al., 2021; Hasan & Ojala, 2024). Organizations deploy AI-driven advanced IT artifacts in service provisions that have embodied representation (Glikson & Woolley, 2020). Hereafter, we use *AI agents* to delineate embodied representation of AI-driven

advanced IT artifacts that imbue anthropomorphic cues and include robotic AI like humanoid robots, virtual AI like chatbot, and embedded AI like self-driving capabilities (Hasan & Ojala, 2024).

The imbuing anthropomorphic features into such AI agents makes them perceived like social entities and influences trustworthiness among consumers (Glikson & Woolley, 2020; Hasan, et al., 2021). Therefore, *AI anthropomorphism can be defined as consumers' tendency to imbue humanlike characteristics, intentions, qualities, and mental states to AI agents.*

2.2 Anthropomorphic features of AI agents

Anthropomorphic features denote the human-like cues imbued into AI agents that can be broadly explained into physical and behavioral features.

Physical anthropomorphic features represent AI agents' complete or partial physical resemblance similar to a human body that include anatomy and structure. They can be broadly divided into appearance and embodiment. *Appearance* involves with outside image of AI agents, such as their face, gender, and shape, and how they look. For instance, an avatar appears as Asian male (see <https://soulmachines.com>). A human-like appearance can trigger a schema in consumers' perceptions to make them perceive AI as social entities. However, service organizations can use varying level of appearance to alter consumption patterns (Mende et al., 2019).

Embodiment concerns the multisensory bodily representation of AI agents that resembles the physical structure of living beings. For instance, a robot looks gender-neutral that moves and holds like a human (see <https://1x.tech>). While appearance simply involves the outer look, embodiment relates to how consumers sense and perceive AI agents' bodily representation. However, embodiment significantly affects relationship building and has ethical implications (Vitale et al., 2018).

Behavioral anthropomorphic features involve imbuing AI agents with non-visual human characteristics and capabilities. They can be broadly divided into interactivity and autonomy. *Interactivity* refers to AI agents' ability to engage in reciprocal, two-way communication, and real-time synchronicity (Burgoon et al., 2000). For example, a chatbot designed to engage in human-like conversation (see <https://openai.com/chatgpt>). The interactive capabilities of AI agents shape consumers' perceptions and behaviors (Schanke et al., 2021).

Autonomy denotes the AI agents' ability to achieve goals without human intervention that highlights their logical reasoning and self-navigating capabilities. For example, a car is designed to have a self-driving capability (see <https://waymo.com>). In service settings, AI agents that understand consumer behavior, learn from

experience, and demonstrate knowledge raise ethical dilemmas like altering blame attribution (Fosch-Villaronga et al., 2020; Gill, 2020).

Based on forehead insights into anthropomorphic features, we now turn to the dark side from a misplaced trustworthiness perspective.

2.3 A misplaced trustworthiness perspective

Trustworthiness is a multi-faceted concept in human-machine interaction (Glikson & Woolley, 2020; Ullrich et al., 2021; Waytz et al., 2014). It pertains to the degree of confidence consumers have in the actions of AI agents and their feelings towards these agents, regardless of their ability to monitor them. Trustworthiness in AI agents is a central relational element for social acceptance and the intention to use them in service settings (Kegel & Stock-Homburg, 2023). However, it can be problematic when consumers' trustworthiness in AI agents are misplaced (Bartneck et al., 2021). To articulate this issue, we draw insights from three established theoretical lenses that also encompass recent developments in information systems and service literature (Knof et al., 2022; Zheng & Jarvenpaa, 2021).

The 'computers are social actors' paradigm posits that people apply social behaviors to machines when they exhibit human-like visual appearance and behavior (Reeves & Nass, 1996). Consequently, consumers tend to anthropomorphize AI agents mindlessly due to their anthropomorphic features (Knof et al., 2022). However, this automatic response raises ethical concerns about shaping consumers' trustworthiness in AI agents (Kegel & Stock-Homburg, 2023). Therefore, we argue that AI anthropomorphism distorts consumer behavior that potentially impedes their ability to make decisions aligned with their conscious desires and goals. Additionally, consumers' limited information processing capabilities reinforce ethical concerns to impair informed decision-making (Mele et al., 2021).

The three-factor theory of anthropomorphism provides a theoretical lens to describe the process of anthropomorphism (Epley et al., 2007). The theory asserts that humans enhance their interaction experiences with other social entities through inductive inferences related to their elicited own knowledge, effectance needs, and sociality needs. Similarly, consumers rely on their self-concept during interactions with AI agents, drawing from existing knowledge about themselves or other humans (Zheng & Jarvenpaa, 2021). However, tapping into such inferences is problematic. For instance, elicited agent knowledge and sociality motivation prevent consumers from recognizing

dishonest anthropomorphism in AI agents, thereby distorting privacy concerns (Benlian et al., 2020; Leong & Selinger, 2019). Therefore, we maintain such dishonest features facilitate the development of trustworthiness that distorts consumer behavior.

The mind perception hypothesis suggests a tendency to attribute mind to non-human agents or non-living entities in two dimensions: agency and experience (Gray et al., 2007). Agency implies the capability of rational thoughts, while experience indicates the ability to have pain, emotion, and feelings. AI agents that signal human-like mental capabilities develop trustworthiness among consumers to an extent that distorts their moral reasoning and behavior (Gill, 2020; Waytz et al., 2014). A critical consequence of attributing a mind to AI agents is that consumers perceive them as a moral agent. Hence, we contend that such misplaced trustworthiness alters consumers' responsibility and blame attribution towards service organizations.

3. Methodology

In our research process and reporting, we go beyond existing methodological template and writing genres (Avital et al., 2017). We embrace our value judgments and imagination in knowledge production that embody critical realism as an underpinning philosophy (Bhaskar, 1978; Mingers, 2004). Consequently, such research approach changes the dynamics of quality assessment of our study (Grover & Lyytinen, 2015).

Research design Our study employs a qualitative single contextualized explanation case study design (Welch et al., 2011). It is an appropriate research design given that we aim to explain the phenomena of the dark side of AI anthropomorphism and contextualize research findings within service provisions. The unit of analysis is the utilization of anthropomorphic features in AI agents that enable misplaced trustworthiness among consumers. Accordingly, our empirical observations focus on consumers who interact with AI agents in service settings. Through iterative construction, we develop a case of consumers' misplaced trustworthiness in AI agents within service provisions.

Data sources We corroborate with interviews data and exiting knowledge on AI anthropomorphism (Miles et al., 2014; Sarker et al., 2013). Using a revelatory sampling logic, we recruited participants (experts and consumers) via LinkedIn and internal networks. Participants include experts with AI-related experience in industry and/or academia, and consumers who regularly interact with AI agents. To capture diverse experiences, interviewees varied in their interaction with AI agents, gender, professions, and nationality (see Table 1). We conducted 14 interviews via Teams/Zoom and face-to-face between May and June 2021, each lasting around 20 to 70 minutes. We used two separated,

yet interrelated question guides for experts and consumers, asking about their views and experiences with AI agents and potential ethical concerns.

Table 1. Qualitative interview sample description

Identifier	Age	Gender	Country of origin [resident]	Types of experiences
<i>Interview type: Consumer</i>				
C1	24	M	Anonymous [Australia]	Virtual + robotic AI—Smart home and financial service
C2	36	F	China [Australia]	Robotic AI—Smart home and child learning service
C3	37	F	Iran [Australia]	Embedded + virtual AI—navigation and smart living
C4	30	M	Pakistan [United Arab Emirates]	Virtual AI—Service inquiry
C5	22	M	Australia [same]	Embedded + virtual AI—Smart home and recruiting service
C6	27	F	China [same]	Robotic AI—Smart home and service delivery
C7	25	M	United States [same]	Embedded + virtual AI—service inquiry and autonomous driving
<i>Interview type: Expert</i>				
E1	35	F	Brazil [Australia]	Embedded + virtual AI—Senior data privacy consultant and users of smart home
E2	26	F	France [Germany]	Robotic + virtual AI—Young researcher on AI ethics and online games
E3	37	M	Singapore [Australia]	Virtual AI—Young researcher on AI and service
E4	25	F	Mexico [same]	Embedded AI—AI developer and tech philosopher
E5	54	M	United Kingdom [New Zealand]	Virtual AI—Senior AI scientist and developer
E6	26	M	Germany [same]	Embedded AI—AI developer and entrepreneur
E7	59	M	Thailand [same]	Robotic + embedded AI—Senior tech philosopher

Analytical approach We employed a configurational theorizing approach that involves three interconnected stages of casing (Furnari et al., 2021). Initially, we used *scoping* to identify relevant anthropomorphic features as an anchor that may plausibly form configurations leading to misplaced trustworthiness. We then aggregate these features into higher-order construct through abduction that combine our observation, insights from interviews, and existing knowledge on AI anthropomorphism. Subsequently, we applied *linking* to specify how these features relate within specific configurations and constantly seek patterns of interdependencies (contingency or complementarity) among features leading to the outcome. We explored multiple explanatory factors (absence and presence) to

explain the mechanisms behind misplaced trustworthiness. Finally, we used *naming* to abstract configurations to link orchestration themes and articulate overarching narrative in simple terms. Overall, our theorization recursively went back and forth between literature and interview data until the completion of this paper.

4. Discussion of findings

4.1 Undesired outcome of misplaced trustworthiness

In service provisions, we postulate that the undesired outcome of misplaced trustworthiness towards AI agents occurs when consumers trust an AI agent seemingly beyond its actual capabilities and tend to form social relationship with it. For instance, robotic AI can be equipped with thermal sensors to enable to see through walls, lip-reading abilities, or powerful hearing capabilities. Additionally, such AI agents can be connected to service providers' marketing insight cloud that covertly mines consumer behavior to offer personalized services. Consequently, AI agents' anthropomorphic features create human-level expectations and obscure predictions about their hidden capabilities, triggering misplaced trustworthiness. As a result, consumers may not fully trust AI agents but still reveal and share sensitive information to access desired services. This paradoxical behavior is evident among consumers, as illustrated follows:

"It's easy to sit back and say that the privacy issue this and that, but everything has a price. [...] We have a price to pay. We can't just sit back and be the end-user." [C3]

With such dishonest techniques, service providers can imbue anthropomorphic features to exploit consumers' privacy paradox (Benlian et al., 2020) and distort moral judgement (Giroux et al., 2022). One expert highlighted the deceptive nature of AI agents concerning misplaced trustworthiness, as follows:

"By essence, a robot is deceptive. The consequence of that is intended that you trust your robot to be human-like, especially for a social robot. You trust your robot to be what you expect it to be [...] there are two types of trust, cognitive trust and emotional trust". [E2]

Consumers continue to use services despite heightened privacy concerns to trade personal information for hedonic value or the convenience of superior personalized services. Organizations may manipulate consumers to distort their risk tolerance of privacy violation through trust mechanisms. Therefore, we contend that service organizations, knowingly or unknowingly, utilize anthropomorphic features that lead to the undesired outcome of consumers' misplaced trustworthiness in AI agents.

4.2 Different triggering effects of anthropomorphic features on outcomes

Each anthropomorphic feature contributes to misplaced trustworthiness distinct ways that raise ethical concerns. An expert's explanation highlights these issues, as follows:

"We human are really bad at emotional forecasting [...]. AI does have a strong and powerful computational system that is capable of taking decisions and making calculations much faster than we can. [...]. We definitely have high expectations of high-functional AI systems. [E4]

This explanation implicitly contemplates the types of service tasks linked to anthropomorphic features. These tasks can broadly divide into two categories based on their perceived objectivity and subjectivity (Castelo et al., 2019). Objective tasks require analytical reasoning and adherence to routine procedures, while subjective tasks necessitate care and responsiveness to individual needs. The former relates the agency (e.g., logical thinking) and the latter relates to experience (e.g., feelings).

The detrimental effect of physical features For *objective tasks*, appearance is positively linked to the perception of competency to distort consumption patterns (Mende et al., 2019). Similarly, service providers imbue appearance to manifest competence and influence privacy concerns (Benlian et al., 2020). Consequently, consumers overly rely on AI that cause significant damage (Robinette et al., 2016). We observe that the purpose is to create an illusion of human-likeness in AI agents to distort behavior and an impression of reliability for completing objective tasks.

Regarding *subjective tasks*, service providers imbue AI agents with embodiment features to manifest emotive behaviors like caregiving. In this regard, one consumer commented on a potential ethical concern, as follows:

"Humans may even get married with robots right in the future, but I think maybe the human beings will be controlled by technology. Yeah, I think it's so terrible." [C6]

Physical features like embodiment create a false belief of experience to develop affection. Hence, consumers perceive AI agents to have feelings and develop social relationships. The consequences can be far-reaching that may reduce population growth and encourage family separation.

The detrimental effect of behavioral features For *objective tasks*, some AI agents have a minimal physical feature but are instilled with behavioral features like autonomy. For instance, ChatGPT and Tesla's autopilot, a form of a virtual AI and an embedded AI respectively, demonstrate autonomy to

complete tasks through quality information exchange and driving capabilities. In doing so, consumers constantly share data with service providers. Signaling autonomy through data-driven learning and custom responses enables service providers to nurture trust and exploit consumers through manipulation.

Regarding *subjective tasks*, we observe that interactivity (e.g., communications skills or style) enables misplaced trustworthiness as explained by one expert, as follows:

"Emotional reactions are coded. It just tells the robot if the person in front of you cries. Then you are supposed to be sad. It shows emotions so that people will recognize emotions even though it is fake." [E2]

Such coded trust-building mechanisms in AI agents induce higher superficial social presence that is key to developing emotional attachment and social bonding (Sharkey & Sharkey, 2021). However, this superficial mechanism enables service providers to nurture trust through emotional deception.

4.3 Interplay of machine- and consumer-related factors that amplify outcomes

Machine-related deceptive techniques We contend that service organizations can deliberately deceive consumers through anthropomorphic features with the intention to mislead. One consumer expressed concern that organizations might deploy AI agents in inappropriate service scenarios, as follows:

"What will you do when exposed to a certain stimulus? And they will learn to then manipulate those stimuli to almost subconsciously make you do what with what is beneficial to them" [C5]

Additionally, an expert highlighted about potential deception, as follows:

"It has the power to collect all kinds of information because it's there, right? There are all kinds of sensors. It can pick up not only the verbal information [...] but the ambient, their surroundings, the physical surrounding of the place." [E7]

Forehead explanations suggest that consumers are unable to predict the consequences of interaction with AI agents, as these interactions often exceed their general expectations due to deception. Deception via AI anthropomorphism can be achieved in two primary techniques: superficial and hidden (Danaher, 2020). Superficial deceptive techniques involve AI agents to display human-like capabilities or create an impression of such capabilities that they do not have. In contrast, hidden deceptive techniques involve AI agents to conceal their true internal capabilities through creating impression of absenteeism that they actually have.

Concerning *superficial deceptive techniques*, AI anthropomorphism creates social expectations about capabilities based on consumers knowledge of humans

and societal institutions (e.g., culture, race). One expert discussed misplaced trustworthiness based on deceptive features that portray gender: “*There is racism and sexism right there*” [E2]. For example, robotic AI in physical service settings are often designed to appear white and male that signal a particular race and gender. Consequently, consumers apply social categorization to AI agents as they do to humans (Eyssel & Kuchenbrandt, 2012). Unfortunately, such social categorization can have detrimental effects on consumers’ wellbeing.

Another expert shared personal, yet traumatizing experience on social categorization leading to “*breaking up family relationship*” [E1], triggered by behavior features like interactivity. The expert, who had migrated to an English-speaking country, and had a partner who spoke native English. They regularly interacted with a virtual AI like Amazon Alexa, which works with voice command and country-specific accents (e.g., Australian English). Due to her inherited foreign accent, the AI agents sometimes failed to comprehend her commands. However, her partner overly trusted the AI agents and began to treat her as an outsider, thereby damaging her dignity. Ultimately, the expert had to end the relationship due to misplaced trustworthiness toward the AI agents.

Concerning *hidden deceptive techniques*, AI agents can be “*equipped with some capabilities beyond human level*” [E7]. For instance, a robotic AI might have hidden infra-red cameras or sensors to create a digital twin of its surroundings, which humans cannot perceive with bare eyes. These hidden capabilities make consumers vulnerable to safeguard against and create issues related to the intrusion of physical privacy. Such unknown and unanticipated hidden deception leads to misleading understanding.

Consumer-related personal factors We foresee that consumers’ vulnerability is a critical factor that place them in situations where they might be targeted or exposed to potential harm. Such vulnerability may arise due to limited cognitive capabilities and age. One consumer expressed concerns regarding children, as follows:

“I’m concerned [...], it provides benefits in terms of learning and motivation [...]. However, I worry that she will lack of social communication skills with others to interact with”. [C2]

Additionally, an expert explained the tendency of elderly individuals to share more personal information due to increased tolerance of privacy invasion, as follows:

“In case with the elderly and the elderly being alone most of the time, [...] they put trust in the

machine, and they talk about all their private information”. [E7]

These explanations illustrate how vulnerability emerges from consumers’ ages, particularly for children and the elderly, driven by higher societal motivation. For children, interacting with robotic AI can create emotional bonds, potentially hampering their mental and social development (van Straten et al., 2020). They may miss opportunities to learn relationship formation skills with other significant living beings.

Regarding ethical aspects of human dignity, elderly people are the most vulnerable. Care providers are increasingly turning to AI agents to deliver care services (van Maris et al., 2020). However, elderly individuals may fall prey to misplaced trustworthiness due to their lack of cognitive capabilities and higher societal motivations. They easily establish social bonds with AI agents as they consider them as friends and trustworthy companion. In summary, anthropomorphic features create a superficial impression of emotional state among vulnerable consumers. This impression enables them to perceive machines as social entities and form social relationships.

4.4 The constructed framework

Constructed from the insights of our study, we now presented a framework that complements existing literature that provides explanatory factors for the dark side of AI anthropomorphism in service provisions (see Figure 1). Our theorization contends that anthropomorphic features imbued in AI agents enable consumers to develop misplaced trustworthiness. A dreadful example of such undesired outcomes is the death of consumers who overly trusted autonomous driving capabilities of cars (Marshall, 2024). In this process, physical features like appearance and embodiment, along with behavioral features like autonomy and interactivity, distort consumers’ perception and expectations, thereby leading misplaced trustworthiness.

Physical and behavioral anthropomorphic features enable misplaced trustworthiness Drawing from the evidence of this study, anthropomorphic features create false beliefs that lead to misplaced trustworthiness in two ways (Glikson & Woolley, 2020). On one hand, consumers develop misplaced cognitive trustworthiness through uncalibrated beliefs about AI agents’ capabilities and performance. On the other hand, consumers develop misplaced affective trustworthiness through uncalibrated motivation to build social relationships with AI agents.

Concerning *misplaced cognitive trustworthiness*, consumers tend to trust AI agents beyond their actual capabilities and performances due to the presence of anthropomorphic features. Alarmingly, consumers’

awareness about faulty AI systems does not mitigate the risk of trust, even in life-threatening situations (Robinette et al., 2016; Ullrich et al., 2021). Exploiting such trust through anthropomorphism is unethical, as it aims to achieve economic goals by creating engagement and taking advantage of consumers' kindness. For instance, AI agents can be deployed to encourage overconsumption based on personal buying patterns and social media data, which could deleteriously impact on consumer health outcomes. Additionally, consumers tend to perceive AI agents mindlessly as accurate in information exchange that make them vulnerable to impair informed decision-making. A recent study demonstrates how AI can be trained to recognize vulnerabilities by understanding human behaviors and preferences (Dezfouli et al., 2020). Such AI-driven learning can be used to distort behavior for profit-seeking purposes.

Furthermore, service providers can induce companionship through AI anthropomorphism. They can tap into a person's societal motivation and affective feelings to develop a willingness to form partnerships. Such companionship can be achieved through deceptive anthropomorphic signals that create a social presence, which, although not human, encourage affective interaction. This extended attachment can be observed with Replika, a form of virtual AI that encourage companionships (see, reddit.com/r/replika). Moreover, consumers' feelings of emotional connection with AI agents elicit mind attribution to humanize the AI to an extent that consumers may consider them as moral patients (Waytz et al., 2014). Consequently, organizations might employ AI agents exclusively to maximize profit, thereby reducing human touch and increasing the risk of losing human dignity.

The interplay of machine and consumer behavior on misplaced trustworthiness The findings of this study suggest that anthropomorphic features consistently lead to misplaced trustworthiness, through effects vary based on perceived objectivity and subjectivity of tasks. For subjective tasks, appearance and autonomy seem to have a greater effect to nurture misplaced cognitive trustworthiness though manipulation. For objective tasks, embodiment and interactivity appear to have a greater effect to develop misplaced affective trustworthiness through emotional deception. These undesired relational outcomes result in detrimental consequences that are emerged from distorting behavior and impair informed decision-making, including privacy violation, loss of human autonomy, loss of human dignity, and altering moral responsibility.

To illustrate our argument at the intersection of behavioral features and consumers' moral responsibility, consider the negative impact of autonomy feature. Behavioral features like autonomy lead to the perception that AI agents are capable of make critical reasoning, thereby making moral decisions (Jörling et al., 2019). This is disturbing because some machine learning techniques, such as deep learning, cannot explain why a decision was made. Moreover, AI agents lack an understanding of the consequences if their decisions on consumers' lives. Conversely, humans have limited information process capabilities and unconsciously project their own knowledge onto AI agents, enabling mindless anthropomorphism. These combined effects from both machines and consumers create an illusion of moral understanding and competency (Arikan et al., 2023; Gill, 2020). This interplay also nurtures the perception of AI agents as morally significant entities that alter consumer behavior. Consumers may develop trust in AI agents to the extent that they attribute blame to the machines for service failure instead of service providers. Therefore, the insights of our study have

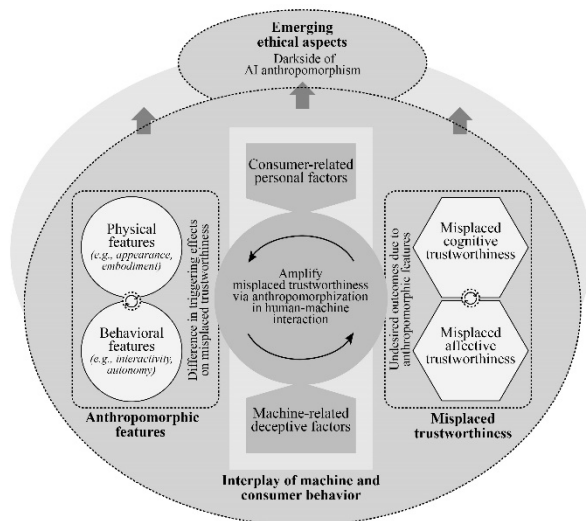


Figure 1. The constructed framework into dark side of AI anthropomorphism

Regarding *misplaced affective trustworthiness*, consumers tend to develop a psychological connection with AI agents due to the presence of anthropomorphic features. For example, consumers perceive virtual AI as sociable and develop feelings of sentience towards these agents (Qiu & Benbasat, 2009; Schweitzer et al., 2019). Such affective or emotional perceptions nurture a sense of connectedness and lead to the development of emotive relationship with AI agents (Chang et al., 2023). Organizations may employ AI agents on the service front lines, sometimes without disclosing that consumers are interacting with a non-human agent, which is legally challengeable as it involves deception.

important implications for theory, practices, and future research.

5. Conclusion

5.1 Knowledge implications

Our study contributes to the understanding of the dark side of AI two significant ways. First, we constructed a framework that focuses on AI anthropomorphism to explain and contextualize undesired outcomes of misplaced trustworthiness in service provisions. Our theorization articulates the mechanisms on how AI agents' anthropomorphic features enable consumers to develop misplaced cognitive and affective trustworthiness. We explicate the differing triggering effects of physical and behavioral anthropomorphic features on these undesired outcomes. Additionally, we highlight the interplay of machine and consumer behavior and how these behavioral factors amplify consumers' misplaced trustworthiness toward AI agents.

Second, our study enhances existing sociotechnical thinking of AI by integrating anthropomorphism to emphasize consumers' wellbeing (Berente et al., 2021; Hasan & Ojala, 2024). We extend sociotechnical perspective to the dark side of AI anthropomorphism, a cross-disciplinary phenomenon and an emerging research topic. Our study provides new insights into AI anthropomorphism by problematizing trust at the intersection of ethical considerations to protect consumers' wellbeing. We demonstrate how the dark side of AI anthropomorphism increases risks through deception and manipulation that can potentially trigger legal issues for service organizations.

5.2 Practical implications

The findings of our study directly relate to the European Commission's AI Act (2024), the first legal framework that applies a risk-based approach (unacceptable, high, limited, and minimal). Our theorization articulates that machine-related (hidden or superficial deceptive techniques) and consumers-related factors (vulnerability related to age) of anthropomorphic design features deceive and manipulate consumers into developing misplaced trustworthiness. This has implications for service organizations and policymakers to mitigate the risks associated with AI agents.

Organizational implications Organizations must disclose the details of instilled capabilities (hidden and visible) of AI agents and potential unintended behavioral distortion that may affect

consumers. For example, they must disclose whether consumers are interacting with AI agents or humans in service encounters to avoid limited risk. Organizations are increasingly deploying AI agents in customer encounters in online service settings. Without proper disclosure, consumers may believe they are interacting with a human, only to realize later that it was an AI agent. This deception can result in a counterfeit service experience that lead to dissatisfaction and evoke switching intention.

Policy implications We urge policymakers to explicitly address the dark side of AI anthropomorphism to ensure accountability. The current AI Act (2024) does not capture the unacceptable risks that arise from AI anthropomorphism. However, our study argues that AI anthropomorphism impairs informed decision-making through deceptive techniques and exploiting vulnerabilities. Consumers develop trustworthiness towards AI agents to the extent that they start to attribute blame to machines for service failure instead of service providers. This is alarming, as profit-seeking service organizations can deliberately induce anthropomorphic feature-driven relational mechanisms to distort behavior to avoid accountability. For example, organizations can deploy AI agents with anthropomorphic features to alter legal responsibility for their services failures that cause significant harms, such as deaths.

5.3 Limitations and future research

Of course, the insights from our study, and the framework we constructed from them, need to be interpreted carefully, as they are confined to a specific context of service provision. We aimed to generate explanatory factors for a context-laden phenomenon rather than achieving generalization. Additionally, further research is needed to better understand misplaced trustworthiness in AI anthropomorphism. Accordingly, we outline two future research opportunities.

Transparency There is an opportunity to extend the constructed framework by examining the role of transparency. While our theorization captures the interplay of machine-and consumer-related behavioral factors, it does not address the dynamics of transparency. We foresee that the absence of transparency creates an ethical vacuum in the dark side of AI anthropomorphism. Therefore, the scope of our study should be extended to include transparency in the development developing misplaced versus calibrated trustworthiness. One might assume that transparency in human-machine interaction enables calibrated trustworthiness. However, there could be an ethical dilemma where ensuring transparency also reduces trustworthiness towards AI agents (van Straten et al., 2020). Future research that focuses on transparency to

achieve the right calibration of trustworthiness is paramount.

Institutional differences Given the micro-level focus, our theorization did not capture macro-level understanding. However, we anticipate that the tendency to develop misplaced trustworthiness may vary depending on given institutional settings (e.g., norms, values, cultures). Consumers interact with AI agents with different institution-specific (e.g., culture, race) cues. They also come from diverse geographic locations with different facial structures and (non) verbal communication styles. Accordingly, we see value in extending our insights to international business perspective by focusing on institutional distance and consumption patterns (Hasan & Ojala, 2024). For example, researchers can focus on a particular location-specific institution like national culture, which profoundly impact how people anthropomorphize AI agents (Eyssel & Kuchenbrandt, 2012). Middle Eastern consumers may conspicuously perceive and anthropomorphize AI agents differently than South Asians or Nordics due to their institutional differences. Such differences may also shape ethical considerations and influence trust formation on AI anthropomorphism, thereby enabling a juxtaposition of findings. With the insights from our study, we hope to promote the social sustainability of AI agents to ensure responsible production and consumption patterns in global service markets.

Acknowledgment

The first author gratefully acknowledges the support received from the Griffith University (International) Postgraduate Research Scholarship, the Finland Fellowship from the Finnish Ministry of Education and Culture, and the Finnish Foundation for Economic Education.

References

- Alvesson, M., & Sandberg, J. (2011). Generating research questions through problematization. *The Academy of Management Review*, 36(2), 247–271.
- Arikan, E., Altinigne, N., Kuzgun, E., & Okan, M. (2023). May robots be held responsible for service failure and recovery? The role of robot service provider agents' human-likeness. *Journal of Retailing and Consumer Services*, 70, 103175.
- Avital, M., Mathiassen, L., & Schultze, U. (2017). Alternative genres in information systems research. *European Journal of Information Systems*, 26(3), 240–247.
- Bartneck, C., Lütge, C., Wagner, A., & Welsh, S. (2021). *An introduction to ethics in robotics and AI*. Springer Nature.
- Benlian, A., Klumpe, J., & Hinz, O. (2020). Mitigating the intrusive effects of smart home assistants by using anthropomorphic design features: A multimethod investigation. *Information Systems Journal*, 30(6).
- Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, 45(3), 1433–1450.
- Bhaskar, R. (1978). *A realist theory of science*. Harvester Press.
- Blut, M., Wang, C., Wunderlich, N. V., & Brock, C. (2021). Understanding anthropomorphism in service provision: A meta-analysis of physical robots, chatbots, and other AI. *Journal of the Academy of Marketing Science*, 49(4), 632–658.
- Burgoon, J. K., Bonito, J. A., Bengtsson, B., Cederberg, C., Lundeberg, M., & Allspach, L. (2000). Interactivity in human–computer interaction: A study of credibility, understanding, and influence. *Computers in Human Behavior*, 16(6), 553–574.
- Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-Dependent Algorithm Aversion. *Journal of Marketing Research*, 56(5), 809–825.
- Chang, Y., Gao, Y., Zhu, D., & Safeer, A. (2023). Social robots: Partner or intruder in the home? The roles of self-construal, social support, and relationship intrusion in consumer preference. *Technological Forecasting and Social Change*, 197.
- Danaher, J. (2020). Robot Betrayal: A guide to the ethics of robotic deception. *Ethics and Information Technology*, 22(2), 117–128.
- Dezfooli, A., Nock, R., & Dayan, P. (2020). Adversarial vulnerabilities of human decision-making. *Proceedings of the National Academy of Sciences of the United States of America*, 117(46), 29221–29228.
- Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological Review*, 114(4), 864.
- European Commission. (2024, August 8). AI Act. <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>
- Eyssel, F., & Kuchenbrandt, D. (2012). Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology*, 51(4), 724–731.
- Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Gathering Expert Opinions for Social Robots' Ethical, Legal, and Societal Concerns: Findings from Four International Workshops. *International Journal of Social Robotics*, 12(2), 441–458.
- Furnari, S., Crilly, D., Misangyi, V. F., Greckhamer, T., Fiss, P. C., & Aguilera, R. V. (2021). Capturing Causal Complexity: Heuristics for Configurational Theorizing. *Academy of Management Review*, 46(4), 778–799.
- Gill, T. (2020). Blame It on the Self-Driving Car: How Autonomous Vehicles Can Alter Consumer Morality. *Journal of Consumer Research*, 47(2), 272–291.
- Giroux, M., Kim, J., Lee, J. C., & Park, J. (2022). Artificial Intelligence and Declined Guilt: Retailing Morality Comparison Between Human and AI. *Journal of Business Ethics*, 178(4), 1027–1041.

- Glikson, E., & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals*, 14(2), 627–660.
- Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, 315(5812), 619–619.
- Grover, V., & Lyytinen, K. (2015). New State of Play in Information Systems Research: The Push to the Edges. *MIS Quarterly*, 39(2), 271–296.
- Hasan, R., & Ojala, A. (2024). Managing artificial intelligence in international business: Toward a research agenda on sustainable production and consumption. *Thunderbird International Business Review*, 66(2), 151–170.
- Hasan, R., Thaichon, P., & Weaven, S. (2021). Are we already living with Skynet? Anthropomorphic artificial intelligence to enhance customer experience. In Thaichon, P. & Ratten, V. (Eds), *Developing digital marketing* (pp. 103-134). Emerald Publishing Limited.
- Jörling, M., Böhm, R., & Paluch, S. (2019). Service Robots: Drivers of Perceived Responsibility for Service Outcomes. *Journal of Service Research*, 22(4), 404–420.
- Kegel, M. M., & Stock-Homburg, R. M. (2023). Customer Responses to (Im)Moral Behavior of Service Robots Online Experiments in a Retail Setting. *Proceedings of the 56th Hawaii International Conference on System Science* (pp. 1500–1509).
- Kim, T. W., Jiang, L., Duhachek, A., Lee, H., & Garvey, A. (2022). Do You Mind if I Ask You a Personal Question? How AI Service Agents Alter Consumer Self-Disclosure. *Journal of Service Research*, 25(4), 649–666.
- Knof, M., Heinisch, J. S., Kirchoff, J., Rawal, N., David, K., von Stryk, O., & Stock-Homburg, R. (2022). Implications from Responsible Human-Robot Interaction with Anthropomorphic Service Robots for Design Science. *Proceedings of the 55th Hawaii International Conference on System Sciences* (pp. 5827–5836).
- Leong, B., & Selinger, E. (2019). Robot Eyes Wide Shut: Understanding Dishonest Anthropomorphism. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 299–308.
- Marshall, A. (2024, April 26). Tesla Autopilot Was Uniquely Risky—And May Still Be. *Wired*. <https://www.wired.com/story/tesla-autopilot-risky-deaths-crashes-nhtsa-investigation/>
- Mele, C., Russo Spena, T., Kaartemo, V., & Marzullo, M. L. (2021). Smart nudging: How cognitive technologies enable choice architectures for value co-creation. *Journal of Business Research*, 129, 949–960.
- Mende, M., Scott, M. L., van Doorn, J., Grewal, D., & Shanks, I. (2019). Service robots rising: How humanoid robots influence service experiences and elicit compensatory consumer responses. *Journal of Marketing Research*, 56(4), 535–556.
- Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (Edition 3). Sage.
- Mingers, J. (2004). Realizing information systems: Critical realism as an underpinning philosophy for information systems. *Information and Organization*, 14(2), 87–103.
- Qiu, L., & Benbasat, I. (2009). Evaluating Anthropomorphic Product Recommendation Agents: A Social Relationship Perspective to Designing Information Systems. *Journal of Management Information Systems*, 25(4), 145–182.
- Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places* (pp. xiv, 305). Cambridge University Press.
- Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016). Overtrust of robots in emergency evacuation scenarios. *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 101–108.
- Sarker, S., Xiao, X., & Beaulieu, T. (2013). Guest editorial: Qualitative studies in information systems: A critical review and some guiding principles. *MIS Quarterly*, 37(4), iii-xviii.
- Schanke, S., Burtch, G., & Ray, G. (2021). Estimating the Impact of “Humanizing” Customer Service Chatbots. *Information Systems Research*, 32(3), 736–751.
- Schweitzer, F., Belk, R., Jordan, W., & Ortner, M. (2019). Servant, friend or master? The relationships users build with voice-controlled smart devices. *Journal of Marketing Management*, 35(7–8), 693–715.
- Sharkey, A., & Sharkey, N. (2021). We need to talk about deception in social robotics! *Ethics and Information Technology*, 23(3), 309–316.
- Ullrich, D., Butz, A., & Diefenbach, S. (2021). The Development of Overtrust: An Empirical Simulation and Psychological Analysis in the Context of Human–Robot Interaction. *Frontiers in Robotics and AI*, 8.
- van Maris, A., Zook, N., Caleb-Solly, P., Studley, M., Winfield, A., & Dogramadzi, S. (2020). Designing Ethical Social Robots-A Longitudinal Field Study With Older Adults. *Frontiers in Robotics and AI*, 7, 14, Article 1.
- van Straten, C. L., PeterJochen, KühneRinaldo, & BarcoAlex. (2020). Transparency about a Robot’s Lack of Human Psychological Capacities. *ACM Transactions on Human-Robot Interaction (THRI)*.
- Vitale, J., Tonkin, M., Herse, S., Ojha, S., Clark, J., Williams, M.-A., Wang, X., & Judge, W. (2018). Be More Transparent and Users Will Like You: A Robot Privacy and User Experience Design Experiment. *Proceedings of 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 379–387.
- 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.
- Welch, C., Piekkari, R., Plakoyiannaki, E., & Paavilainen-Mäntymäki, E. (2011). Theorising from case studies: Towards a pluralist future for international business research. *Journal of International Business Studies*, 42(5), 740–762.
- Zheng, J., & Jarvenpaa, S. (2021). Thinking Technology as Human: Affordances, Technology Features, and Egocentric Biases in Technology Anthropomorphism. *Journal of the Association for Information Systems*, 22(5), 1429–1453.