# Conversational AI Agents: The Effect of Process and Outcome Variation on Anthropomorphism and Trust

**Kambiz Saffarizadeh**
kambiz.saffari@uta.edu

**Mark Keil**
mkeil@gsu.edu

**Abstract.**
Organizations increasingly deploy conversational AI agents (CAs) in agentic roles where behavioral variations are inevitable. Prior work often conflates two distinct forms of variation: outcome variation (where success fluctuates) and process variation (where the path to completion varies). This study investigates how these two types of variation jointly impact user anthropomorphism and trust, integrating the theory of anthropomorphism with the misconception-of-chance bias.

Using a CA in an agentic role within a car parking simulation experiment, we manipulate process and outcome variation. Results indicate that outcome variation negatively impacts trust, but process variation shows no such direct effect. Interestingly, while neither variation alone increases anthropomorphism, their interaction does. We theorize that the co-presence of process and outcome variation prevents falsification of anthropomorphic explanations for the CA's behavior by creating non-random-appearing sequences that violate individuals' randomness expectations due to misconception of chance. This allows users to apply human-like mental models, attributing agency and experience to the CA, making failures interpretable as exploratory behavior. In contrast, when there is outcome variation without process variation (i.e., the CA fails to achieve the desired outcome but exhibits no process variation), users will tend to infer that the CA is mechanically defective.

Our results further demonstrate that the joint effect of process and outcome variation indirectly increases trust through anthropomorphism as a mediator. These findings provide insights into the nuanced ways behavioral variation in CAs influence user trust, contributing to the literature on anthropomorphism, trust in AI, and the misconception of chance.

**Keywords:** Conversational Agent, Artificial Intelligence, Chatbot, Agentic AI, Misconception of Chance, Error, Hallucination, Alignment, Anthropomorphism, Process Variation, Outcome Variation

Version: September 20, 2025

# Conversational AI Agents: The Effect of Process and Outcome Variation on Anthropomorphism and Trust

## INTRODUCTION

Conversational AI agents (CA) have become integral to our daily lives, with the global market projected to grow from $5.39 billion in 2023 to $43.83 billion in 2033 (Spherical Insights, 2024). Organizations deploy CAs in increasingly agentic roles, as investment advisors, software-development copilots, and autonomous parking assistants (Baird & Maruping, 2021; Berente et al., 2021; Hodgson, 2024; Saffarizadeh, Keil, & Maruping, 2024). Because these agents increasingly act on users' behalf, their economic value hinges on sustained user trust (Glikson & Woolley, 2020; Saffarizadeh, Keil, Boodraj, et al., 2024).

Yet the probabilistic nature of modern CAs creates a paradox that challenges our understanding of trust in information systems (IS). Built on machine learning pipelines that produce inherently variable behavior (Berente et al., 2021), these systems violate a core assumption underlying trust in a specific technology: that technological artifacts exhibit deterministic behavior (McKnight et al., 2011; Schuetz & Venkatesh, 2020). A generative model that "hallucinates" references in writing or an automated parking assistant that leaves a car in the wrong parking spot could exemplify non-deterministic behavioral variation (McCracken, 2023). This behavioral variation stems from the fundamental architecture of probabilistic inference, data drift, and environmental indeterminacies (Kordzadeh & Ghasemaghaei, 2021; Tarafdar et al., 2022).

According to established competence-based trust theories, such variation should erode user trust (Lankton et al., 2015; McKnight et al., 2011). In line with this view, the algorithm aversion literature argues that users abandon algorithmic systems after observing even minor errors, believing these systems will deterministically repeat the same mistakes (Dietvorst et al., 2015, 2018). This body of literature further contends that people prefer human judgment over algorithms that exhibit performance variations, even when algorithms perform better on average (Dietvorst & Bharti, 2020).

However, empirical observations of CA usage reveal a puzzling contradiction. Despite frequent errors and unpredictable behaviors, users not only continue engaging with CAs but often exhibit remarkably persistent trust (Nearsure, 2024; Prillaman, 2024). Some studies show individuals often trust algorithmic advice more readily than human advice, even after observing errors (Logg et al., 2019; Yeomans et al., 2019). Users appear to employ different cognitive mechanisms when evaluating modern CAs (Sohail et al., 2024), discounting ambiguous errors while attributing greater objectivity and internal consistency to algorithms than to humans (Castelo et al., 2019). These apparently paradoxical findings suggest traditional competence-based trust theories, developed for deterministic systems, may need revising in the context of modern probabilistic CAs.

We argue that resolving this theoretical tension requires recognizing that behavioral variation in CAs is not monolithic but comprises distinct types that may trigger different cognitive processes. Drawing from engineering and robotics literature (Kreye et al., 2011; Mirnig et al., 2017), we distinguish between two fundamental types of variation: outcome variation and process variation.[1] *Outcome variation* refers to unpredictable variation in the degree to which a CA fulfills a user's intended task over repeated interactions. Outcome variation is about the unpredictable changes in how well aligned a CA's outcome is with a user's intention over repeated interactions (Gabriel, 2020).[2] *Process variation* refers to unpredictable variation in how a CA fulfills a task over repeated interactions. For instance, an autonomous parking CA exhibits outcome variation when it occasionally fails to park in the requested spot, but process variation when it takes different routes while successfully completing the task. CAs used in agentic roles and built on current platforms like n8n, Zapier, and OpenAI GPTs typically

[1] A third type of variation, input variation, is also discussed in the literature (Kreye et al., 2011). Input variation refers to unpredictable changes in the inputs received by a CA. This type of variation arises from external sources, such as the user or the environment, that provide inputs to the CA. Therefore, while process and outcome variations directly pertain to the CA's internal behavior, input variation is externally driven and must be managed by the CA. As such, input variation falls outside the scope of our research, which focuses on the behavioral variations in the CA itself.

[2] It is important to note that in this conceptualization of outcome variation, erroneous outcomes are a special case. The concept of error assumes a dichotomous outcome wherein all outcomes could be categorized as either a complete success or a complete failure. However, in the context of artificial intelligence, a request to the CA could lead to an outcome that *ranges* from being completely aligned with the request to a complete failure to fulfill the request (Gabriel, 2020).

consist of several key components that can yield both types of variation simultaneously. Outcome variation often emerges from probabilistic inference in decision-making and action execution, while process variation is often intentionally programmed (Gartner, 2023; GoogleCloud, 2023; OpenAI, 2023; Salesforce, 2025).

Examining these variations in isolation misses a fundamental insight: users do not experience process and outcome variation separately but as integrated behavioral patterns that shape their overall perception of the CA. Theoretically, the joint effect matters because variation patterns provide mutual context that fundamentally alters cognitive processing (Clark, 2013; Friston, 2010). When outcome variation appears alone, it may signal incompetence ("the system has a bug"). When process variation appears alone, it may seem like programmed variety. But when both occur together, they create behavioral patterns that likely resist simple explanation, appearing instead as the complex, adaptive behavior characteristic of human-like agents (Caruso et al., 2010; Ebert & Wegner, 2011; Szollosy, 2017). Importantly, process variation could prevent the falsification of anthropomorphic explanations for outcome variation: when a CA repeatedly attempts tasks identically but fails randomly, the pattern may appear mechanically defective, but when the CA varies its approach across attempts, failures become interpretable as exploratory behavior or contextual adaptation, patterns consistent with human-like agency and experience (Ebert & Wegner, 2011; H. M. Gray et al., 2007). This falsification-prevention mechanism, where process variation provides alternative explanations that preserve anthropomorphic attribution despite occasional outcome failures, highlights why joint variation may produce qualitatively different effects than either variation alone.

To theorize about this phenomenon, we integrate the misconception of chance literature and the theory of anthropomorphism (Epley et al., 2007). Humans expect random sequences to "look random" with excessive alternation, even over short spans (Gilovich et al., 1985; Kahneman & Tversky, 1972). When CA behavior violates these expectations, producing clusters or streaks that appear intentional rather than random, users likely perceive human-like "streaky" variation rather than mechanical stochasticity. Combined process and outcome variation likely

create such patterns. Anthropomorphism, the attribution of humanlike agency and experience to non-human entities (Epley, Waytz, et al., 2008; Waytz, Morewedge, et al., 2010), offers a promising theoretical lens through which to understand why these patterns may intensify people's effectance motivation (the motivation to explain and predict agent behavior) and drive them to apply their most accessible mental model for understanding complex, adaptive behavior: human agency and ability to experience (Broadbent, 2017; Epley et al., 2007).[3]

We propose that when process and outcome variation occur concurrently, this may simultaneously decrease trust via a direct effect (Dietvorst et al., 2015, 2018; Glikson & Woolley, 2020) and increase trust via an indirect effect through anthropomorphic attribution (Epley, Waytz, et al., 2008; Waytz, Morewedge, et al., 2010). The tension between these two opposing effects makes it crucial to understand the interplay between process and outcome variation, and whether this interplay influences trust in CAs. Users who perceive a CA as human-like may apply different evaluative standards, forgiving occasional errors if the agent appears to be "trying" through varied approaches. An anthropomorphized CA that varies its problem-solving approaches while occasionally failing might maintain higher trust than one that fails identically each time, despite the two CAs having the same success rate (i.e., overall outcome). This insight would also be valuable to practice, where trust persistence despite imperfection is essential for adoption and continued use. Therefore, we pose the following research questions:

***RQ1****: How do the combined patterns of outcome and process variation influence users' trust in a CA?*

***RQ2****: What role does anthropomorphism play in this context?*

To investigate our research questions, we conducted a randomized experiment in which a CA assumed a fully agentic role, parking a vehicle on behalf of the user. Participants interacted

[3] Alternative perspectives could certainly be applied: for example, expectancy-confirmation theory would emphasize user expectation and performance alignment (Sohail et al., 2024), and algorithm aversion and appreciation models would foreground comparative trust in algorithms versus humans (Dietvorst et al., 2015; Logg et al., 2019). Yet none of these perspectives systematically explain why variability itself might lead users to perceive error-prone agents as intentional, humanlike actors. Anthropomorphism uniquely integrates cognitive tendencies to infer hidden mental states with design-induced behavioral patterns, making it especially suitable for unpacking the interplay between outcome and process variation.

with the CA through a car-parking simulation. The CA that we developed resembles Tesla's AutoPark app, making the study more realistic than a hypothetical scenario in which the user has no direct interaction with the software artifact. The parking context was selected because it both exemplifies an agent acting on the user's behalf and permits clean, independent manipulation of our two focal constructs: *outcome variation* (whether the car ultimately reaches the designated parking spot) and *process variation* (the path taken to get there). Participants were randomly assigned to one of four treatment groups in a 2×2 factorial design. Randomized experiments are the gold standard of internal validity as they provide a robust way of assessing causal relationships (Shadish et al., 2002). In addition, we collected qualitative data to gain additional understanding of participants' interaction with the CA.

This research contributes to our understanding of anthropomorphism in the context of conversational AI agents by highlighting the falsification-prevention mechanism through which process variation moderates the effect of outcome variation on anthropomorphic attribution. We extend the theory of anthropomorphism by integrating the misconception of chance as the cognitive mechanism that explains when and why behavioral variation triggers human-like attributions, specifically, when variation patterns violate expectations of randomness and appear intentionally adaptive. Our findings challenge traditional competence-based trust models by suggesting that anthropomorphism may shift users from outcome-based to effort-based evaluation, fundamentally altering how trust forms in probabilistic AI systems. While outcome variation directly erodes trust, process variation can preserve it indirectly through anthropomorphism, revealing dual pathways through which behavioral variation affects trust, enriching the discourse on trust in algorithms and human-AI interaction. Additionally, the study offers new perspectives on user attribution of human-like qualities to CAs by drawing from the misconception of chance literature. These theoretical insights are particularly important as organizations deploy inherently variable, probabilistic CAs in increasingly agentic roles. The practical implications drawn from this study provide guidance for developers on strategically

implementing process variation to manage inevitable outcome variation, with the aim of promoting both trust and continued use.

## THEORETICAL BACKGROUND

### Anthropomorphism

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

The term anthropomorphism is conceptualized in two fundamentally distinct ways across various research streams. In studies predominantly rooted in human-computer interaction (HCI) and communication research (Burgoon et al., 1999, 2000; Nunamaker et al., 2011), anthropomorphism generally refers to a design attribute, defined explicitly as "the degree to which the interface simulates or incorporates humanlike characteristics" (Burgoon et al., 1999, p. 36) or as "the technological efforts of imbuing computers with human characteristics and capabilities." (Gong, 2008, p. 1495). Conversely, in studies primarily influenced by psychology and management research (Epley et al., 2007; Haslam & Loughnan, 2014; Waytz, Morewedge, et al., 2010; Waytz et al., 2014), anthropomorphism is conceptualized as an internal, cognitive phenomenon—specifically, an inductive inference whereby individuals attribute humanlike characteristics to nonhuman entities. Thus, while the former stream emphasizes anthropomorphism as an externally designed property, the latter emphasizes it as a psychological process rooted in human cognition.

In line with psychology and management literature, we define anthropomorphism as an inference about real or imagined nonhuman entities that leads to the *attribution* of humanlike characteristics, properties, emotions, inner mental states, and motivations to them (Epley et al., 2007; Epley, Waytz, et al., 2008; H. M. Gray et al., 2007). Therefore, first, anthropomorphism is not adding physical (referred to as form anthropomorphism in HCI) or behavioral

anthropomorphic features (referred to as behavioral anthropomorphism in HCI) to a nonhuman agent (Gambino et al., 2020); it is a person's mental attribution of humanlike characteristics to the nonhuman agent, which may be triggered by either the presence of such characteristics (Benlian et al., 2020; Diederich et al., 2022; Seeger et al., 2021) or the person's internal state such as chronic loneliness (Dang & Liu, 2023; Epley, Akalis, et al., 2008; Eyssel & Reich, 2013). Second, anthropomorphism is not the mere use of human adjectives to describe the physical aspects of nonhumans; it involves going beyond observable characteristics of the entity and making inference about its unobservable characteristics (Epley, Waytz, et al., 2008). Third, anthropomorphism reflects people's tendency to *perceive* human traits in nonhuman agents. Questions regarding the accuracy of this perception and whether a nonhuman entity should be treated as human are orthogonal to anthropomorphism (Epley, Waytz, et al., 2008).

The theory of anthropomorphism (Epley et al., 2007) identifies three determinants that influence when and why people attribute humanlike characteristics to nonhuman entities: elicited agent knowledge, sociality motivation, and effectance motivation. Elicited agent knowledge refers to the accessibility and applicability of anthropocentric knowledge (i.e., the cognitive availability of human-like schemas that can be applied to nonhuman entities). When human knowledge structures are highly accessible (either chronically or situationally), people are more likely to use these schemas to interpret nonhuman behavior (Epley et al., 2007). This represents the cognitive foundation for anthropomorphism, as human schemas are typically the most elaborate and readily available frameworks for understanding complex behavior (Broadbent, 2017).

Sociality motivation captures the desire for social contact and affiliation (Mourey et al., 2017). When people experience loneliness or social disconnection, they are more motivated to mentally construct human-like agents from nonhuman entities to fulfill their need for social connection (Epley et al., 2007). Research has demonstrated that chronically lonely individuals show increased anthropomorphism of pets, gadgets, and supernatural agents (Epley, Akalis, et al., 2008; Eyssel & Reich, 2013).

Effectance motivation represents the drive to explain, understand, and predict other agents' behavior to attain mastery over one's environment. When an entity's behavior is uncertain or unpredictable, effectance motivation intensifies as individuals seek to reduce uncertainty and regain a sense of control (Waytz, Morewedge, et al., 2010; White, 1959). This motivation makes individuals more susceptible to taking mental shortcuts in processing information, including applying human-like explanatory mental models to nonhuman entities (Waytz, Gray, et al., 2010; Waytz et al., 2014).

Among these three determinants, effectance motivation is most directly triggered by behavioral variation in artificial agents (Caruso et al., 2010; Epley, Waytz, et al., 2008). When faced with unpredictable behavior, individuals experience heightened effectance motivation and often attribute human-like capacities to nonhuman entities as an explanatory mechanism (Waytz, Morewedge, et al., 2010). These capacities fall into two categories: capacity for agency (self-control, morality, memory, emotion recognition, planning, communication, and thought) and capacity for experience (hunger, fear, pain, pleasure, rage, desire, personality, consciousness, pride, embarrassment, and joy) (H. M. Gray et al., 2007). Some scholars refer to these factors as human uniqueness and human nature, respectively (Haslam & Loughnan, 2014).

Despite theoretical predictions that behavioral variation should increase anthropomorphism through effectance motivation (Epley et al., 2007; Saffarizadeh, Keil, Boodraj, et al., 2024; Zheng & Jarvenpaa, 2021), empirical findings remain inconsistent. Studies examining this effect are scattered across different fields and lack systematic approaches, making it difficult to draw definitive conclusions about this theoretically important relationship.

Several studies support the theoretical prediction that behavioral variation increases anthropomorphism. For instance, Epley, Waytz, et al. (2008) found that participants in their experiments were more likely to anthropomorphize a pet when it exhibited more behavioral variation. Similarly, Waytz, Morewedge, et al. (2010) demonstrated that participants were more likely to anthropomorphize computerized gadgets when they were perceived as having higher behavioral variation. Chen (2020) showed that unpredictable behaviors (e.g., a clock displaying

two different random responses to hitting snooze) led to higher levels of anthropomorphism, especially among people with conservative views. Johnson & Barrett (2003) found that participants who lacked control over an electromagnet that moved marbles along unexpected trajectories attributed intentional agency to the marbles.

However, the relationship between behavioral variation and anthropomorphism is not always straightforward. While Salem et al. (2013) found that occasional incongruence of gestures and verbal utterances in robots increased anthropomorphism, their later study (Salem et al., 2015) discovered that participants were more likely to anthropomorphize and trust a flawless robot compared to an imperfect robot that exhibited occasional inconsistencies in its behavior. Mirnig et al. (2017) found that people liked faulty robots more than flawless ones but found no significant difference in how participants anthropomorphized faulty versus flawless robots.

The relationship between behavioral variation and anthropomorphism thus remains unresolved. While some studies show a positive effect (Salem et al., 2013; Waytz, Morewedge, et al., 2010), others fail to find an effect (Mirnig et al., 2017). These contradictory results may stem from a key limitation in the current literature: the lack of distinction between different types of behavioral variation. Existing research often fails to differentiate between process variation (variability in how a task is performed) and outcome variation (variability in the results). This oversight is problematic because process and outcome variation may have distinct, and potentially interactive, effects on anthropomorphism. For instance, high process variation coupled with low outcome variation might be perceived differently from low process variation with high outcome variation, leading to different levels of anthropomorphism.

Understanding how both process and outcome variation affect anthropomorphism, and whether the effect of one depends on the other, is essential for resolving the current theoretical ambiguity. The inconsistent findings suggest that the relationship between behavioral variation and anthropomorphism may depend on how humans cognitively process patterns of variation. Understanding this cognitive processing requires examining a fundamental bias in human perception of randomness that may shape how users interpret CA behavior patterns.

## The Misconception of Chance

The misconception of chance is the biased tendency to expect that a sequence of outcomes generated by a random process will "look" random even when the sequence involves a short span (Gilovich et al., 1985; Kahneman & Tversky, 1972; Tversky & Kahneman, 1974). For instance, even when people understand that the outcome of a coin toss is random with a 50% chance of head (H) or tail (T), they are more likely to perceive an HTHTHT sequence as random than an HTTTTT sequence, which "does not represent the equal likelihood of heads and tails," or even an HHHTTT sequence, which "does not appear random" (Bazerman & Moore, 2013, p. 41). People's mental representations of randomness are thus biased toward over-alternation, and they interpret deviations from these expectations, especially short streaks or clusters, as evidence of an intentional process (Oskarsson et al., 2009). This bias manifests in diverse contexts, from lottery number selection (Clotfelter & Cook, 1993) to sports performance (Gilovich et al., 1985) and to attributions of intentionality in physical systems (Valdesolo & Graham, 2014).

Research suggests that people often have unnecessarily "complex models of the mechanisms they believe generate observed events, and they rely on these models for explanations, predictions, and other inferences about event sequences" (Oskarsson et al., 2009, p. 262). Such complex models may be used to explain sudden non-random looking patterns in an otherwise unpredictable sequence of events (Ebert & Wegner, 2011). In the context of agentic CAs, people tend to rely on their mental model of human behavior when an agent's behavior appears unpredictable because such a model is our most accurate mental model for understanding seemingly unpredictable behavior (Broadbent, 2017; Riedl et al., 2014; Waytz, Morewedge, et al., 2010).

Understanding how humans misperceive random sequences provides a critical cognitive mechanism that may explain when and why behavioral variation triggers anthropomorphic attribution. When combined with the theory of anthropomorphism's emphasis on effectance motivation, the misconception of chance offers a pathway to understanding how different types of variation (process and outcome) might interact to influence user perceptions. These perceptual

processes, in turn, have implications for how users develop and maintain trust in CAs, our final theoretical component.

## Trust in AI Agents

Prior research suggests that people tend to use a human-based conceptualization of trust when interacting with AI agents such as CAs (Lankton et al., 2015; Saffarizadeh, Keil, & Maruping, 2024). Hence, in line with most prior studies on trust in CAs (e.g., Saffarizadeh, Keil, Boodraj, et al., 2024), we use a human-based conceptualization of trust. *Trust* refers to "the willingness of a party to be vulnerable to the actions of another party based on the expectation that they will perform a particular action important to the trustor, irrespective of the ability to monitor or control the other party" (Mayer et al., 1995, p. 712). Trust represents a willingness or intention to rely on another party (McKnight et al., 2002) and this willingness is mostly based on the other party's perceived *trustworthiness*. Trustworthiness, also referred to as trusting beliefs (McKnight et al., 2002), comprises perceived competence, perceived integrity, and perceived benevolence (Mayer et al., 1995).

While prior literature predominantly examines trust, scholars highlight the value of explicitly considering distrust as a distinct but related construct. Distrust represents negative expectations regarding another party's intentions or actions, while ambivalence captures simultaneous, conflicting feelings of trust and distrust, often of similar magnitude (Lewicki et al., 1998; Moody et al., 2017). Although trust and distrust intuitively appear as opposite ends of a continuum, neuroimaging research reveals these constructs engage distinct brain regions, underscoring their conceptual separateness (Dimoka, 2010). Research on ambivalence (i.e., holding simultaneous, conflicting attitudes of similar magnitude) indicates that trust and distrust often do coexist (Moody et al., 2014, 2017). Proponents of differentiating these constructs argue that relationships are multifaceted, allowing individuals to simultaneously hold trust and distrust toward the same entity (Lewicki et al., 1998). However, others contend that trust and distrust empirically function as opposite ends of the same continuum within specific task domains,

suggesting minimal practical benefit in treating them as separate constructs (Schoorman et al., 2007). Given our research context—trust toward AI agents within a clearly defined, specific task domain—we align with the latter approach. Thus, while we acknowledge the theoretical value and potential insights offered by explicitly studying distrust and ambivalence, we choose to focus primarily on trust.

Research on trust in AI agents spans various fields (for a thorough review of trust in AI, see Glikson & Woolley, 2020). Studies focusing on trust in recommendation agents have identified several key factors that influence trust. These include familiarity, perceived personalization (Komiak & Benbasat, 2006), human-like features (Qiu & Benbasat, 2009), explanation and transparency (Wang & Benbasat, 2007; Xu et al., 2014), as well as the type of recommendation agent and response times (Wang & Benbasat, 2013).

Findings from the algorithm aversion literature indicate that people are less likely to rely on algorithms than on humans (Burton et al., 2019), particularly in subjective areas such as joke recommendations (Castelo et al., 2019; Yeomans et al., 2019). This tendency persists even when algorithms and humans make identical errors (Dietvorst et al., 2015). Similarly, people are averse to AI agents making a range of ethical decisions (Bigman & Gray, 2018). In uncertain domains, people prefer methods with higher outcome variation such as human judgment (i.e., error-prone options) even at the expense of average performance (Dietvorst & Bharti, 2020). In contrast, evidence from the algorithm appreciation literature demonstrates a general user preference for algorithms (Bigman et al., 2022; Logg et al., 2019; You et al., 2022). Specifically, in situations where information regarding human versus algorithm performance is absent, individuals exhibit a greater tendency to appreciate and adhere to advice provided by algorithms rather than humans (Bauer & Gill, 2024; Logg et al., 2019). For instance, research shows that people are more likely to share contact information with automated sales agents than human sales agents (Adam et al., 2023). To explain these seemingly contradictory findings between algorithm aversion and algorithm appreciation, researchers have proposed various factors (for

comprehensive reviews, see Burton et al., 2019; Jussupow et al., 2020, 2024). One such factor is anthropomorphism, which may offer valuable insights.

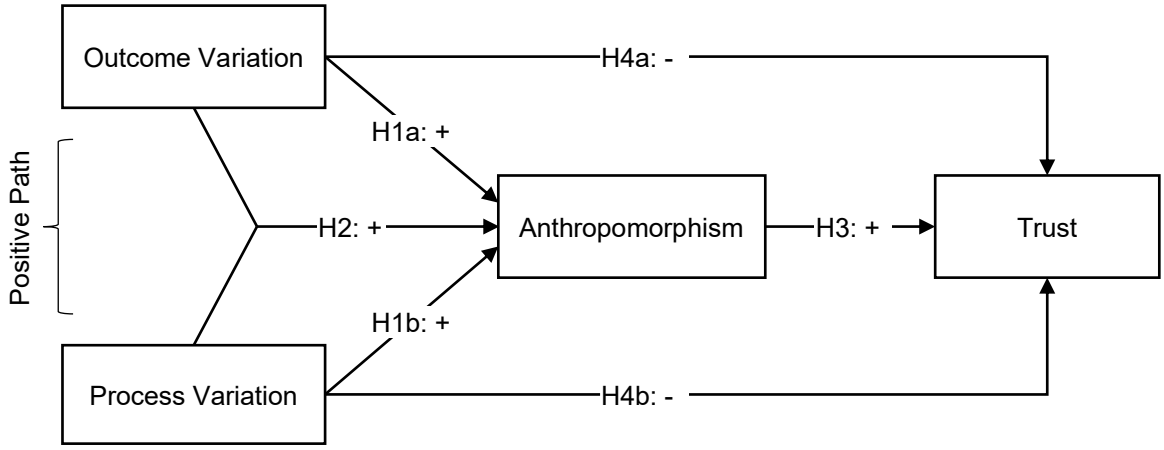
Previous research suggests that when users anthropomorphize a CA, they shift from treating it as a non-human algorithm to more like a human agent (Saffarizadeh, Keil, Boodraj, et al., 2024). However, the implications of this shift for trust are unclear. While the algorithm appreciation literature (Logg et al., 2019) implies that anthropomorphizing might decrease trust because people prefer algorithms over humans, studies on algorithm aversion (Dietvorst et al., 2015, 2018) imply the opposite. Some work has found that anthropomorphism may be positively associated with trust (Glikson & Woolley, 2020) as it may provide cognitive evidence for a CA's agency and affective evidence for its ability to experience (Saffarizadeh, Keil, Boodraj, et al., 2024; Waytz et al., 2014).

Despite these findings, the combined effects of anthropomorphism, process variation, and outcome variation on trust in CAs remains largely unexplored. Prior work indicates that outcome variation can decrease trust in algorithms (Dietvorst et al., 2015; Dietvorst & Bharti, 2020), although studies in human-robot interaction have reported no significant impact (Salem et al., 2015). Since anthropomorphism can shift the perception of a CA from an algorithm to a human-like agent, it is crucial to understand how process variation and outcome variation interact to shape trust, considering the role of anthropomorphism. Thus, our study aims to address an important gap in our present understanding by investigating the interplay of these factors and their collective influence on user trust in CAs.

Together, these three theoretical perspectives, anthropomorphism, the misconception of chance, and trust in AI agents, provide the foundation for understanding how behavioral variation in CAs influences user trust. In the following section, we integrate these perspectives to develop our research model and derive testable hypotheses.

## THEORY DEVELOPMENT

We integrate the misconception of chance into the theory of anthropomorphism to explain the effect of process and outcome variation on anthropomorphism and in turn the effect of anthropomorphism on trust. We then draw on the trust literature to account for the direct impact of process and outcome variation on trust. The main premise of our theorizing is that the way individuals process and make sense of behavioral variation in CAs influences the degree to which they trust them. We explain why individuals' misconception of random sequences in behavioral variation of CAs can fulfill their innate need for control and lead them to anthropomorphize CAs to be able to make sense of the variations. Figure 2 provides a summary of our research model and hypotheses.




Note: Control variables not shown for visual clarity.

**Figure 2.** Research Model

### Extending the Theory of Anthropomorphism Through the Misconception of Chance

We argue that the misconception of chance acts as a perceptual trigger for effectance motivation when an entity's behavior is difficult to explain or predict. When observed variation departs from expected randomness, producing clusters, streaks, or patterns that "don't look random," it increases the subjective salience of an underlying cause (Gilovich et al., 1985; Kahneman & Tversky, 1972; Tversky & Kahneman, 1974). This perceived non-randomness intensifies

effectance motivation, the drive to explain and predict agent behavior (Epley et al., 2007; Waytz, Morewedge, et al., 2010).

The misconception of chance explains why behavioral variation specifically triggers anthropomorphism rather than other explanatory frameworks. When variation produces sequences that violate expectations of randomness, users perceive patterns requiring explanation (Gilovich et al., 1985; Oskarsson et al., 2009). If no sufficiently rich nonhuman causal model is available, people default to importing the human mental model as the explanatory schema (Broadbent, 2017; Riedl et al., 2014; Waytz, Morewedge, et al., 2010).

Importantly, the misconception of chance creates conditions where anthropomorphism becomes the most cognitively accessible explanation. When a CA fails twice then succeeds three times, this pattern violates expectations of alternation, appearing to show "learning" or "increased effort" rather than random variation. The discomfort with unexplained variation (effectance motivation) combined with misperception of random patterns creates ideal conditions for anthropomorphic attribution (Caruso et al., 2010; Epley et al., 2007). Anthropomorphism reduces uncertainty not by making behavior directly predictable but by making it *understandable* within a familiar framework, that of human-like experience and agency with goals, emotions, and situational responses (Epley et al., 2007; H. M. Gray et al., 2007; K. Gray et al., 2012).

**Effects of Process Variation and Outcome Variation on Anthropomorphism**

**Outcome variation** could trigger anthropomorphism by creating behavioral sequences that appear intentional rather than random. The human brain is so active in making sense of seemingly random patterns in its sensory inputs (Clark, 2013) that people often claim to have found meaningful patterns in purely random events (Ebert & Wegner, 2011; Gilovich et al., 1985). When users observe CA behavior across repeated interactions, they unconsciously assess whether the pattern matches their mental template of "random" behavior. Due to the misconception of chance, users expect even short sequences to reflect global probabilities with excessive alternation (Kahneman & Tversky, 1972). However, actual random generation often

produces clusters or runs that violate these expectations, leading users to perceive intentionality rather than mechanical stochasticity.

This perception is amplified because humans tend to heavily rely on recent events in a sequence (DuBrow & Davachi, 2013). Even with maximum outcome variation (e.g., 50% success rate), a short sequence of CA outcomes is unlikely to "look" random with equal numbers of successes and failures. The resulting non-random-appearing patterns trigger effectance motivation, as users perceive patterns that seem intentional but whose logic remains opaque (Caruso et al., 2010; Epley et al., 2007). When outcomes deviate from expectations, humans are motivated to identify causal explanations (Clark, 2013; Kelley, 1967; Lombrozo, 2006), especially when unpredictability could affect goal attainment (White, 1959).

**Process variation** could also signal behavioral flexibility that contradicts mechanical explanations. When a CA varies how it accomplishes tasks, users observe a form of adaptability that mechanical systems typically lack (Szollosy, 2017). This flexibility violates expectations of algorithmic rigidity (Dietvorst et al., 2015), the assumption that artificial systems execute identical procedures for identical requests. Each variation in process could suggest the presence of internal states or decision-making capabilities that influence behavior selection (Waytz, Gray, et al., 2010). Users might perceive a CA's varied responses to similar queries as reflecting emotional fluctuations or intentional decision-making, even though these variations result from algorithmic processes (Epley et al., 2007). This tendency reflects the same cognitive bias that underpins the misconception of chance, where people misinterpret variation as involving human agency or experience (Caruso et al., 2010).

When CA behavior appears neither fully random nor fully deterministic, existing non-agentive explanatory frameworks prove insufficient. Variation in CAs occupies this liminal space: too patterned to be pure randomness, too inconsistent to be mechanical determinism. In such situations, people tend to leverage their mental models of humans to make sense of the observed variation (albeit unconsciously) (Ebert & Wegner, 2011; Kim & Sundar, 2012).

Anthropomorphism provides a cognitive framework that accommodates behavioral variation through attribution of agency and experience (H. M. Gray et al., 2007; Waytz, Cacioppo, et al., 2010; Waytz, Gray, et al., 2010). Attribution of agency implies that the agent's behavior can be driven by the ability to make autonomous choices, which are not qualitatively different than purely random choices (Bigman & Gray, 2018; Ebert & Wegner, 2011). Attribution of experience implies that the agent's behavior can be influenced by internal states such as emotions, which could lead to unpredictability (K. Gray et al., 2012). While neither attribution represents a rational view of a CA, which typically operates based on algorithms and data rather than emotions or truly autonomous decision-making, experimental and fMRI studies have shown that people are prone to such irrational attributions (Waytz, Gray, et al., 2010; Waytz, Morewedge, et al., 2010). Following the previous literature, we thus expect that the presence of variation in the behavior of a CA increases the likelihood that people anthropomorphize the CA to explain away the process and outcome variation. Therefore, we advance the following hypotheses:

*H1a: The presence of outcome variation in a CA increases the level of anthropomorphism.*

*H1b: The presence of process variation in a CA increases the level of anthropomorphism.*

**Joint Effect of Process and Outcome Variation on Anthropomorphism**

Process variation could fundamentally alter how users interpret outcome variation by providing contextual cues that support intentional rather than mechanical explanations. When outcome variation occurs in isolation, users still can often maintain mechanical explanations ("the system has a bug"). However, when paired with process variation, the behavioral pattern becomes too complex for simple mechanical failure (Szollosy, 2017) but consistent with human-like behavior. Process variation transforms occasional outcome failures from systematic incompetence into what appears as exploratory behavior or contextual adaptation.

The co-presence of process and outcome variation creates patterns that violate randomness expectations. Due to the misconception of chance, people are more likely to assume

that complex variation comes from an intentional model (Oskarsson et al., 2009) and are prone to anthropomorphizing such models (Epley, Waytz, et al., 2008). When users observe correlation between process changes and outcome changes (even when none truly exists) they perceive intentional adaptation rather than random variation.

Process variation prevents the falsification of anthropomorphic explanations for outcome variation. When a person anthropomorphizes a sequence-generating randomizer as an imaginary persona such as the "lady luck," very few contextual cues exist that can falsify the soundness of the anthropomorphism. The unfalsifiable nature of the phenomenon in this example is a major reason why the misconception of chance strongly predicts people's behavior in such a context (Caruso et al., 2010). In CAs, however, behavioral cues, such as process and outcome variation, can interact with each other and potentially falsify the anthropomorphism. Research has shown that a mismatch between a person's expectations of an anthropomorphized agent and observations create large feedback errors in the person's mind due to violations in neurocognitive expectancies (Friston, 2010; Saygin et al., 2011). Depending on the level of mismatch, the attribution of human-like characteristics to the nonhuman agent may be completely rejected (Clark, 2013; Epley et al., 2007).

When paired with outcome variation, process variation specifically supports anthropomorphic attribution through multiple reinforcing mechanisms. First, process variation may be perceived as a sign that the CA has "tried" different ways to fulfill the requested task as each attempt "looks" different. As a result, and due to the misconception of chance, outcome variation is likely to be perceived as not a mere coincidence and instead will be likely attributed to the CA's agency (Caruso et al., 2010; Oskarsson et al., 2009). For instance, if a CA changes its way of responding while interacting with a user (e.g., altering the phrasing or sequence of steps) it may reinforce the idea that it is acting with purpose rather than generating random outcomes. Second, process variation signals that some behavioral variance exists by design, leading users to interpret outcome variation as potentially intentional. Third, process variation

contradicts the classical idea of a "mechanical" AI that operates deterministically (Szollosy, 2017), lending additional support to explanations involving emotion or experience.

In contrast, the absence of process variation provides cues that contradict anthropomorphic explanations. When all the CA's attempts to carry out a specific user request look identical except for the outcome, the behavior appears mechanical. For instance, if the user asks the CA to park their car several times, the CA will always take the same path to park the car—albeit sometimes unsuccessfully (e.g., the car may suddenly stop before the parking spot). Therefore, a lack of process variation can serve as a contextual cue that contradicts an anthropomorphized explanation of outcome variation in the CA's behavior. That is, the behavior of the CA without process variation looks "mechanical" and causes one to reject the notion that the reason for the observed outcome variation is the humanlike qualities of the CA. In conclusion, we argue that process variation can complement outcome variation and increase the likelihood that people anthropomorphize the CA. Therefore, we propose the following hypothesis:

> *H2: The presence of process variation in a CA positively moderates the effect of outcome variation in the CA on anthropomorphism.*

**Effects of Anthropomorphism, Process Variation, and Outcome Variation on Trust**

Anthropomorphism preserves trust by transforming how users interpret behavioral variations, especially outcome variation. When users anthropomorphize an agent, they perceive it to be more competent, predictable, and caring. This likely occurs through an explanatory reframing, a shift in the causal model users apply to outcome variation from stable, internal deficiencies (incompetence) to unstable, potentially external factors (temporary states, misunderstanding, exploration). An anthropomorphized CA's outcome variation admits multiple explanations that preserve competence beliefs: the CA was "confused," "trying something new," or "having difficulties." These explanations, while they may be objectively incorrect, provide cognitive frameworks where current failure does not predict future failure (Waytz et al., 2014).

Anthropomorphized agents are perceived as possessing agency, which implies competence despite variation. An anthropomorphized agent is perceived to have high agency (H. M. Gray et al., 2007). People perceive entities with high agency to be capable of planning, controlling, and fulfilling tasks (K. Gray et al., 2011; Waytz et al., 2014). This perception of agency can be particularly important for CAs, as users may feel that the agent is capable of making informed decisions based on more human-like reasoning, even though it is governed by algorithmic logic. Therefore, an anthropomorphized CA is more likely to be perceived as competent.

Anthropomorphism increases perceived predictability by providing a familiar framework for understanding behavior. Prior research has shown that one of the major reasons that people anthropomorphize nonhuman agents is to increase their ability to predict the agents' behavior (Epley et al., 2007; Waytz, Morewedge, et al., 2010). When users believe that the CA operates based on recognizable and understandable human-like decision processes, they feel more in control of the interaction. In other words, anthropomorphism increases the perceived predictability of an agent. This is especially relevant for CAs because they often handle dynamic tasks, such as customer service (Schanke et al., 2021), where the user expects a certain level of predictability in responses. The misconception of chance reinforces this trust preservation by suggesting that patterns will change; just as humans expect random sequences to alternate, they expect an anthropomorphized CA's "bad streak" to end.

Anthropomorphism fosters emotional connection that mitigates negative responses to behavioral variation. Prior research has shown that anthropomorphism is associated with feelings of connectedness and warmth (Epley et al., 2007; Qiu & Benbasat, 2009). In the context of CAs, users might feel a stronger bond with an agent they anthropomorphize. Some scholars have suggested that lonely people "create human agents out of nonhumans through anthropomorphism to satisfy their motivation for social connection" (Epley et al., 2007, p. 866). In the context of CAs, this emotional connection can mitigate negative experiences with the CA's behavioral variation (e.g., occasional failures), as users may attribute such instances to understandable

lapses rather than systematic errors (Salem et al., 2013). When a user anthropomorphizes a CA, they are more likely to perceive it as caring. In interactions with CAs, this perceived care can translate into users believing the agent is more likely to act in their best interest.

In summary, users perceive an anthropomorphized agent to be more competent, predictable, and caring. Therefore, we argue that users are more willing to be vulnerable to the actions of a CA when they anthropomorphize it, and in line with prior research (Qiu & Benbasat, 2009; Waytz et al., 2014) we advance the following replication hypothesis:

*H3: Anthropomorphism increases user trust in a CA.*

## Direct Effects of Outcome and Process Variation on Trust

Outcome variation could also directly erode trust by signaling unreliability and incompetence. When a CA exhibits outcome variation, users cannot be sure about the behavioral outcome of the system. The reason for outcome variation could be system failure, mistaking user commands, conflicting commands, or the agent's own decision to override the user's commands. Regardless of the underlying reason and what the user perceives the reason to be, the agent exhibits inconsistency in fulfilling tasks. While the user might attribute the lack of fulfillment to either the agent's own agency or some other problem in the system, the inconsistency in behavior decreases the user's perception of the agent's integrity.

Furthermore, since outcome variation means that sometimes the agent will not completely fulfill the assigned task to the user's liking, the user will find the CA less competent and its behavior less desirable. Therefore, we hypothesize that:

*H4a: The presence of outcome variation in a CA decreases user trust in the CA.*

Process variation could also directly reduce trust by violating users' expectations of behavioral consistency. When a CA performs the same task in noticeably different ways across interactions (e.g., changing steps, phrasing, or sequencing), it introduces procedural unpredictability. This variation may lead users to perceive the CA as unstable or unreliable, even if the outcomes are acceptable. Process variation violates users' expectations of behavioral

consistency, a key aspect of perceived reliability in trustees (Lankton et al., 2015; McKnight et al., 2002).

*H4b: The presence of process variation in a CA decreases user trust in the CA.*[4]

## METHODOLOGY

To test our research model, we conducted a 2×2 between-subject factorial design experiment and independently manipulated process and outcome variation.

### Participants

We recruited 180 participants, of whom 163 (66 females and 97 males) passed the attention check question[5] and were retained for subsequent analysis. Participants averaged 43 years of age (ages ranged from 18 to 77).[6] Half of them had an education of 4 years of college or more and experience interacting with digital assistants once a week or more. We chose to recruit the participants from Amazon's Mechanical Turk because samples from MTurk are demographically diverse (Chandler et al., 2019) and yield generalizable findings (Coppock, 2019).

### Artifact

We developed a car parking simulation environment in which users could instruct a CA to park a car, modeled after Tesla's AutoPark app. This context was selected for three reasons. First, it allowed the CA to serve in a clearly agentic role by carrying out an action directly for the user, closely aligning with our problem formulation. Second, the structure of the parking task provided a rigorous and practical means to independently manipulate outcome variation (whether the car reached the designated spot) and process variation (the path taken to get there). Finally, autonomous driving and parking technologies were highly salient in mainstream media during the study period, which enhanced the realism of the scenario and increased participant engagement.

---

[4] This hypothesis was introduced in the revision process for completeness in response to reviewers' suggestions and was not part of the initial version of the manuscript.

[5] After we measured the dependent variables, participants were instructed to read the following paragraph carefully: "Your experiences on Mturk are important for this survey. In order to demonstrate that you have read this question, please select other and type the word shoe as your answer to the question below. How often do you participate in Mturk surveys?" Participants who did not type the word "shoe" under the "other" option failed our attention check question.

[6] While age variations exist in the sample, random assignment ensures that such factors do not systematically influence the experimental results. This diversity also contributes to the broader generalizability of our findings.

We developed this simulation in JavaScript and integrated it with the rest of our study through the Qualtrics XM platform's APIs. Programmatically, the environment consisted of a terrain, a house object, a parking spot object, a vehicle object, and a CA object. The terrain was internally divided to 16 location nodes to keep the program lightweight for the web interface. The CA object could 'scan' the terrain by querying the terrain object and identifying the empty location nodes. It then calculated all feasible paths from its current location to the parking spot using a depth-first search (DFS) algorithm. To minimize variation in participant inputs, we limited user commands by providing a pre-set, clickable sentence, such as "Drew park my car in the parking space." Upon receiving the user's message, the CA would scan the terrain, select a path, and drive the vehicle to the parking spot. The simulation and its seamless integration with the rest of the study created an engaging task environment with a high degree of psychological realism for participants (Berkowitz & Donnerstein, 1982), thus bolstering the ecological validity of our experiment.

To assess participant engagement, we included a feedback question at the end of the study within our qualitative questions, which confirmed that participants found the task engaging. For instance, one participant noted, "The questions were easy to understand, and the concept of auto parking is definitely interesting, especially if the car can park perfectly within the established parking space." Another remarked, "I thought this study was well-made, and it was rather interesting to interact with artificial intelligence that could park your car." Yet another participant shared, "Self-driving and parking cars are a real boon for poor drivers like me. Humans err, and machines can fail, but it reduces the risk associated with the human element. I love the technology."

## Procedure

Figure 3 provides an overview of the experiment procedure. After providing informed consent, participants were shown a cover story, narrated by a newscaster's voice, to enhance the psychological realism of the experiment (see Figure 4). This approach aligns with previous

research on anthropomorphism (Waytz, Morewedge, et al., 2010). Participants were told that they needed to interact with an AI-based driving assistant to park a car in 10 randomly generated simulation environments. We asked the participants to pay attention to how the CA worked in order to answer follow-up questions after the interaction. The CA's name was randomized between Amanda and Drew to control for gender expectations (Nass et al., 1997). Next, participants used the CA to park the car 10 times (see Figure 5). After each attempt, they clicked on "Next Attempt," which loaded a new randomly generated environment and reset the chatbox. In all environments, the locations of the parking spot and a house, as well as the starting position of the car, were fixed. However, to keep the task engaging, we randomly selected the ground texture (from a pool of 5 desert ground texture images) and the house (from a pool of 5 house images), and randomly placed 15 desert bunchgrasses to create the environment (see Appendix B). After the 10 attempts, the participants answered a series of questions about the CA. Finally, they were debriefed and compensated. For a detailed sample of participants' interaction with the CA, see Appendix C.

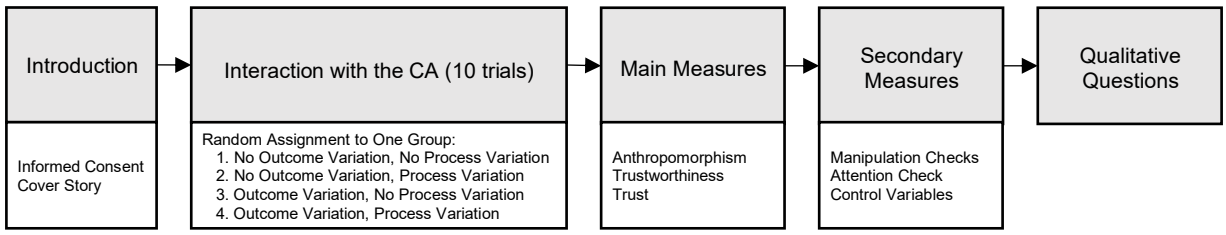



**Figure 3.** Experiment Overview

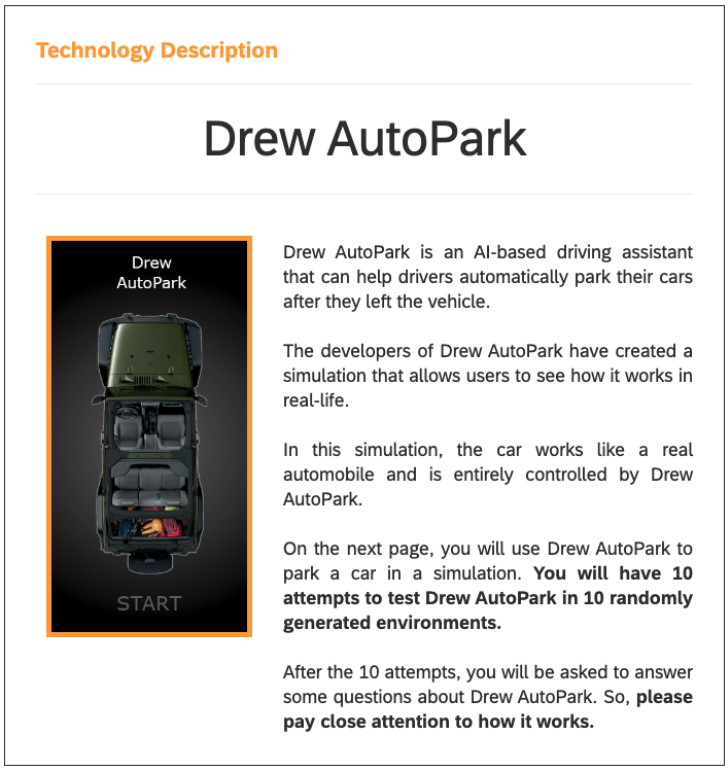



**Figure 4.** The Cover Story that Participants Read Before Interacting with the CA


**Figure 5.** A Sample of Participants' Interaction with the CA Illustrating Parking in the Wrong Spot

## Measures

We measured *anthropomorphism* using measurement items adopted from previous studies (i.e., Waytz, Morewedge, et al., 2010). This operationalization of anthropomorphism includes one omnibus item that directly measures whether people attribute a mind, two items that measure

whether they attribute agency, and two items that measure whether they attribute the ability to experience (e.g., the ability to experience emotions) to the CA.

We measured *trust* using three measurement items adopted from previous literature (Srivastava & Chandra, 2018). We also measured perceived competence, perceived integrity, and perceived benevolence separately based on items adopted from previous studies on trust in recommendation agents (Wang et al., 2016). These factors indicate the perceived *trustworthiness* of the trustee (Mayer et al., 1995). We created perceived trustworthiness as a composite construct based on an average of perceived competence, perceived integrity, and perceived benevolence.

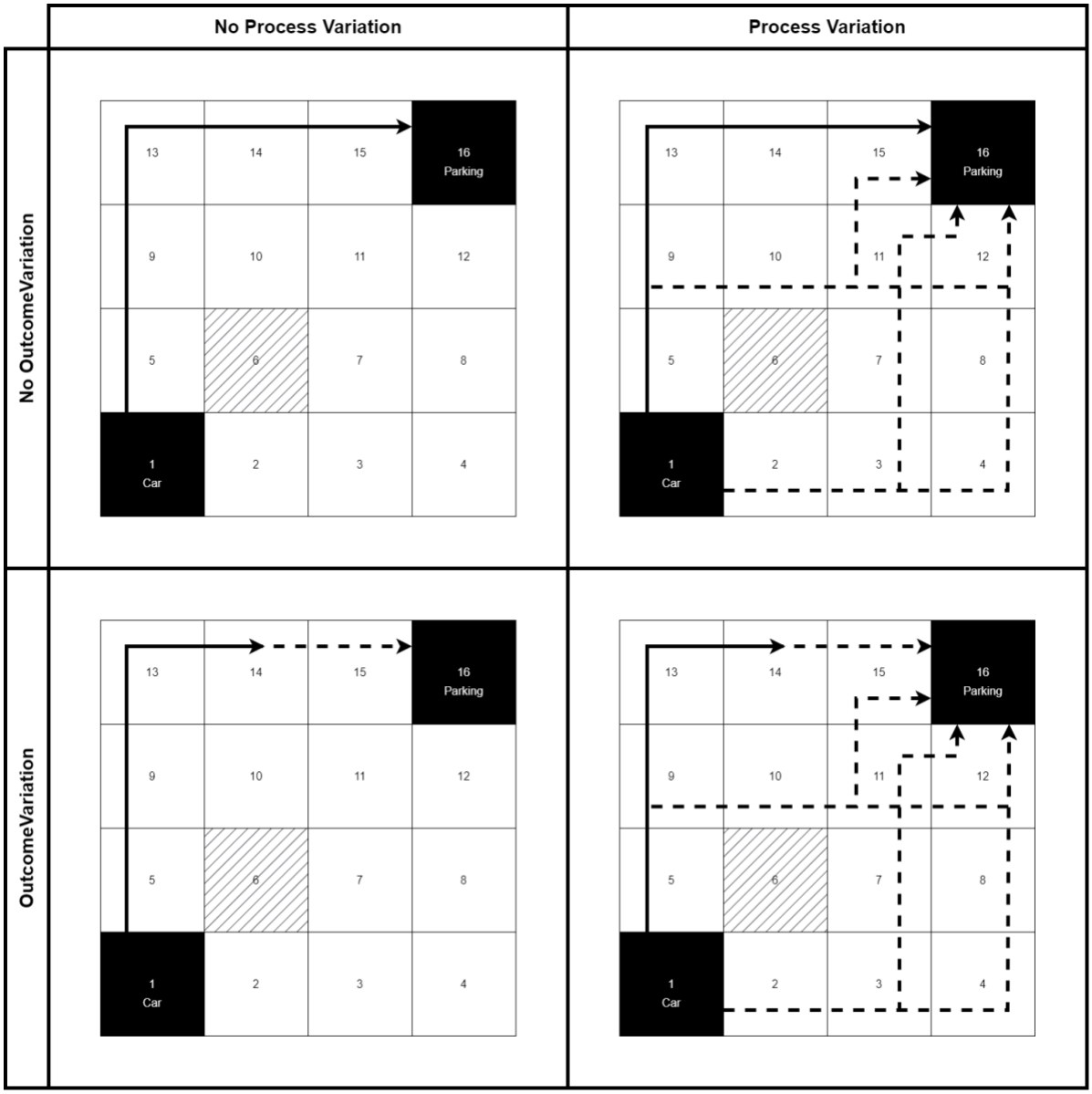
Trustworthiness captures participants' trusting beliefs about the characteristics of the CA, while trust reflects their intentions or willingness to rely on the CA (Mayer et al., 1995; McKnight et al., 2002). Together, these concepts represent participants' trust-related perceptions of the CA (Lankton et al., 2015), and we include both measures for completeness. In line with the previous literature (Riedl et al., 2010; Yuan & Dennis, 2019), we also controlled for sex, age, education, and previous experience using AI assistants. See Appendix D for more details about the measures.

## Manipulations

The car parking scenario provided an ideal context for manipulating process variation and outcome variation independently. To manipulate process and outcome variation, we built on the existing approaches in the literature (Dietvorst & Bharti, 2020; Ebert & Wegner, 2011). Process variation was manipulated by making the CA park the car using either the same path (no-process variation condition) or a randomly selected path from all possible paths (process variation condition) every time the user asked it to park the car. To ensure that no specific path drives the results in the no-process variation condition, for each specific participant in this condition, we randomly picked one path from all possible paths and kept the path constant throughout all interactions the participant had with the agent. To limit the possible number of paths from the car

location to the parking spot to a tractable number, we created a graph with 16 nodes, where the car was on node 1 and the parking spot was on node 16. Assuming that the car does not go through the same location more than once and cannot go over the house, there are 22 directional edges in the graph. The agent used a depth-first search (DFS) algorithm to find all possible paths from nodes 1 to 16.

Outcome variation was manipulated by making the CA either always park the car in the designated parking space (no-outcome variation) or occasionally not park the car in the designated parking space (outcome variation condition). More specifically, in the outcome variation condition, there was a 1-in-3 random chance that the CA would park the car somewhere on the path to but before reaching the designated parking space. Since each of the 10 tries could lead to an either successful or failed parking, there are 1,024 possible sequences one of which randomly occurs for a participant in the outcome variation condition. Figure 6 provides a schematic view of the discussed manipulations.




Note: In the experiment, the car parking task was repeated 10 times for each participant, leading to 10 parking paths. Whether these paths were unique or not was dependent on the experimental condition. However, in the figure above, we show only a few paths in the process variation condition for visual clarity and demonstration purposes.

**Figure 6.** Schematic View of Process Variation and Outcome Variation Manipulation

Figure 7 presents additional details on the distributions of processes and outcomes resulting from the exogenous randomizers in both the process variation and outcome variation conditions.

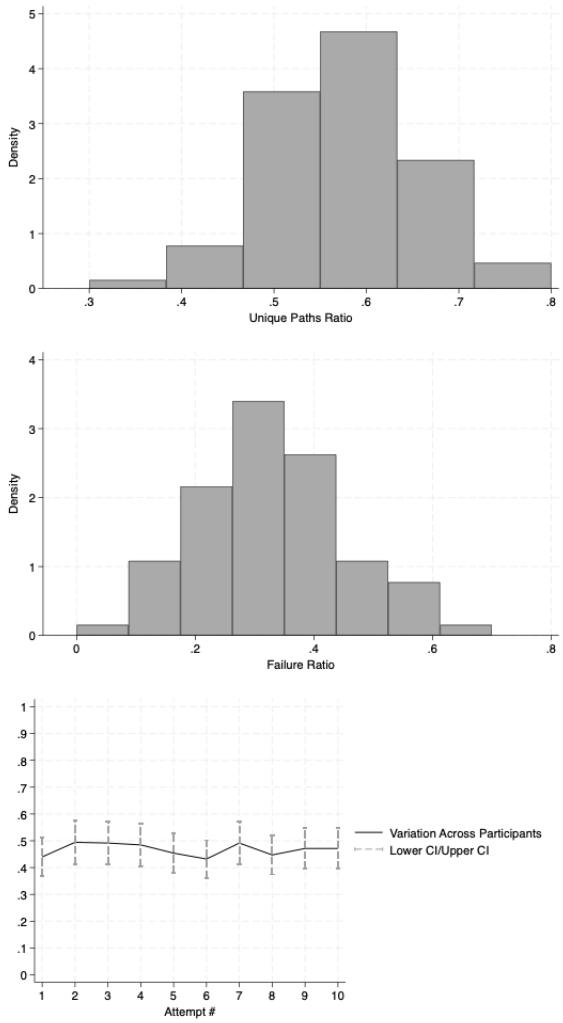



**Figure (a)** shows the distribution of unique paths ratio for individual participants in the process variation condition. Unique paths ratio refers to the number of paths that appears only once in all attempts of each participant divided by the total number of attempts (i.e., 10). The distribution mean is 0.581 (SD= 0.100, N=77).

**Figure (b)** shows the distribution of failure rates in the outcome variation condition. Failure ratio refers to the number of failures divided by the total number of attempts (i.e., 10). The distribution mean is 0.326 (SD= 0.144, N=74), which represents the 1 out of 3 random chance that the CA would park the car somewhere on the path to but before reaching the designated parking space.

**Figure (c)** shows the variation of success and failures in each attempt in the sequence of 10 attempts across all participants in the outcome variation condition. As shown, the distribution of variations in all attempts appear similar, which provides further evidence for the successful implementation of outcome variation manipulation.

**Figure 7.** Distribution of Failure Ratio, Unique Paths Ratio, and Outcome Variation in Each Attempt

## Results

The manipulation check questions for process variation ($\alpha = 0.954; 3\ items$) and outcome variation ($\alpha = 0.875; 3\ items$) revealed that both process variation ($M_{No} = 2.628, SD_{No} = 1.559;\ M_{Yes} = 4.719, SD_{Yes} = 1.612;\ \mathrm{t}(161) = 8.411,\ p < 0.0001, \mathrm{d} = 1.320$) and outcome variation ($M_{No} = 2.400, SD_{No} = 1.278;\ M_{Yes} = 4.365, SD_{Yes} = 1.469;\ t(161) = 9.125, p < 0.0001, \mathrm{d} = 1.435$) were successfully manipulated. Appendix D includes the manipulation check questions.

Table 1 shows the group means and standard deviations for anthropomorphism, trust, and trustworthiness. We employed a system of hierarchical regressions with heteroscedasticity robust

standard errors to test our research model.[7] This approach allows for errors to be correlated for each given participant across the set of regressions used to estimate paths.

**Table 1.** Group Means for Anthropomorphism, Trust, and Trustworthiness Across Experimental Conditions

| Outcome Variation | Process Variation | Anthropomorphism | Trust | Trustworthiness |
|---|---|---|---|---|
| No | No | 2.258 | 4.868 | 5.094 |
| | | (1.034) | (1.202) | (0.937) |
| | Yes | 2.141 | 4.992 | 5.045 |
| | | (0.953) | (1.471) | (1.235) |
| Yes | No | 1.979 | 3.333 | 3.825 |
| | | (0.775) | (1.596) | (1.267) |
| | Yes | 2.389 | 3.917 | 4.352 |
| | | (0.889) | (1.527) | (1.012) |

Standard deviations in parentheses.

In the first set of regressions, we estimated the effects of outcome variation (no=0, yes=1), process variation (no=0, yes=1), and their interaction on anthropomorphism ($\alpha = 0.866$; average of 5 items). In the second set of regressions, we estimated the effects of outcome variation (no=0, yes=1) and anthropomorphism on trust ($\alpha$*=0.969*; average of 3 items). To ensure that our second set of regressions was not misspecified due to the elimination of process variation, we also included this factor and its interaction with outcome variation. In addition, we re-estimated the second set of regressions with perceived trustworthiness ($\alpha$*=0.848*; average of 3 items)[8] as the dependent variable to determine if our results were robust regardless of how we measured trust in CA. In all regressions, we controlled for age, sex, ethnicity, education, and past experience using AI assistants.

After examining the residuals of the regressions, we detected non-normality in the distribution of residuals in a few of the regressions. While this typically does not pose an issue in large sample sizes, we accompanied all our analyses with bias-corrected and accelerated bootstrap results based on 1,000 replications.

The results showed no significant main effects of outcome variation on anthropomorphism (*$\beta$=-0.073, se=0.138, p=0.600; 95% BootCI [-0.354,0.209]*) or process variation on anthropomorphism (*$\beta$=0.137, se=0.137, p=0.317; 95% BootCI [-0.162,0.436]*).

[7] To obtain estimates with heteroscedasticity-robust standard errors, we used a maximum likelihood estimator via Stata's 'gsem' command. The results were consistent with those from a Feasible Generalized Least Squares (FGLS) estimator using 'sureg' and Ordinary Least Squares (OLS) estimator with robust heteroscedasticity-robust standard errors using 'reg.'

[8] Perceived trustworthiness was calculated based on three factors: perceived benevolence ($\alpha$*=0.885*; average of 3 items), perceived integrity ($\alpha$*=0.930*; average of 4 items), and perceived competence ($\alpha$*=0.972*; average of 3 items).

Thus, **we did not find support for H1a and H1b**. The interaction of process and outcome variation was significant (*β=0.693, se=0.281, p<0.05; 95% BootCI [0.129,1.257]*), thus **providing support for H2**. We further found that anthropomorphism was positively associated with trust (*β=0.479, se=0.112, p<0.001; 95% BootCI [0.234,0.724]*), thereby **supporting H3**.

Our results confirm that outcome variation has a negative direct impact on trust, both on average (*β=-1.297, se=0.214, p<0.001; 95% BootCI [–1.744,–0.850]*) and when process variation is absent *(β=-1.319, se=0.308, p<0.001; 95% BootCI [-1.933,-0.706]),* thus **supporting H4a**. We did not find evidence that process variation has a negative direct effect on trust. When outcome variation was absent, we did not find a significant effect of process variation (*β=0.346, se=0.264, p=0.189; 95% BootCI [-0.207,0.899]*). On average, the effect was marginally significant but in the opposite direction of our hypothesis (*β=0.367, se=0.216, p=0.089; 95% BootCI [–0.097,0.830])*. Thus, **we did not find support for H4b.** This pattern implies that the effect of process variation on trust may be primarily driven by its interaction with outcome variation (we further test this possibility in our mediation analysis).

We found similar results when using perceived trustworthiness as an alternative way of measuring our dependent variable (see Appendix E). Specifically, anthropomorphism was positively associated with perceived trustworthiness (*β=0.399, se=0.092, p<0.001; 95% BootCI [0.198,0.599]*). Moreover, outcome variation had a significant negative direct effect on perceived trustworthiness, both on average (*β=-0.943, se=0.164, p<0.001; 95% BootCI [-1.278,-0.608]*) and when process variation was absent (*β=-1.094, se=0.238, p<0.001; 95% BootCI [-1.578,-0.61]*). In contrast, process variation did not have a statistically significant negative direct effect on perceived trustworthiness, either on average (*β=0.258, se=0.164, p=0.116; 95% BootCI [-0.082,0.598]*) or when outcome variation was absent (*β=0.120, se=0.203, p=0.554; 95% BootCI [-0.294,0.534]*). Table 2 and Figure 5 show the results.

To test whether anthropomorphism mediates the effect of the interaction between process variability and outcome variability on trust, we estimated the indirect effect using 1,000 bias-corrected and accelerated bootstrap samples (Hayes & Preacher, 2014; Zhao et al., 2010). The

results show that **anthropomorphism significantly mediates the impact of the interaction between process variability and outcome variability on trust** (*β=0.332, se=0.178, 95% BootCI [0.073,0.756]*), and perceived trustworthiness (*β=0.276, se=0.146, 95% BootCI [0.053,0.625]*).

We further explored the role of anthropomorphism by testing the influence of process and outcome variation on subdimensions of anthropomorphism (see Appendix F). *The results suggest that people may be less likely to directly admit that they attribute a mind to a CA, and yet, when asked indirectly, they reveal evidence of such attribution.*

**Table 2.** Hierarchical Regression Results for the Main Analysis

| | (1) Anthropomorphism | (2) Anthropomorphism | (3) Trust | (4) Trust |
|---|---|---|---|---|
| **Independent Variables** | | | | |
| Anthropomorphism | - | - | 0.482*** | 0.479*** |
| | - | - | (0.113) | (0.112) |
| Process Variation | 0.137 | -0.179 | 0.367* | 0.346 |
| | (0.137) | (0.187) | (0.216) | (0.264) |
| Outcome Variation | -0.073 | -0.411** | -1.297*** | -1.319*** |
| | (0.138) | (0.199) | (0.214) | (0.308) |
| Process × Outcome | - | 0.693** | - | 0.045 |
| | - | (0.281) | - | (0.441) |
| **Control Variables** | | | | |
| Constant | 3.127*** | 3.447*** | 5.124*** | 5.152*** |
| | (0.407) | (0.405) | (0.629) | (0.675) |
| Age | -0.006 | -0.005 | -0.007 | -0.007 |
| | (0.006) | (0.006) | (0.009) | (0.009) |
| Sex (Female=0) | | | | |
| Male | 0.100 | 0.096 | 0.088 | 0.088 |
| | (0.138) | (0.134) | (0.225) | (0.225) |
| Ethnicity (White=0)[c] | | | | |
| Black or African American | 0.815** | 0.906** | -0.240 | -0.232 |
| | (0.404) | (0.392) | (0.556) | (0.558) |
| American Indian or Alaska Native | -0.335* | -0.536** | 0.799** | 0.785** |
| | (0.202) | (0.234) | (0.379) | (0.395) |
| Asian | -0.192 | -0.272 | 0.305 | 0.299 |
| | (0.232) | (0.241) | (0.311) | (0.306) |
| Latino or Hispanic | -0.723*** | -0.754*** | -1.053** | -1.057** |
| | (0.280) | (0.252) | (0.462) | (0.459) |
| Education (Less than High School = 0) | | | | |
| High School | -0.480* | -0.654** | -1.607*** | -1.620*** |
| | (0.271) | (0.265) | (0.444) | (0.479) |
| Some College | -0.861*** | -1.059*** | -0.873* | -0.888* |
| | (0.241) | (0.261) | (0.469) | (0.489) |
| 2-year College Degree | -0.174 | -0.411 | -0.937* | -0.953* |
| | (0.298) | (0.324) | (0.522) | (0.560) |
| 4-year College Degree | -0.483** | -0.684*** | -1.191*** | -1.205*** |
| | (0.214) | (0.228) | (0.363) | (0.413) |
| Master's Degree | -0.537** | -0.759*** | -1.294*** | -1.309*** |
| | (0.240) | (0.246) | (0.380) | (0.426) |
| Doctorate Degree | -0.711** | -0.807*** | -2.397*** | -2.405*** |
| | (0.286) | (0.306) | (0.609) | (0.618) |
| Past Experience Frequency (At least once a day=0) | | | | |
| At least once a week | -0.443** | -0.437** | 0.151 | 0.150 |
| | (0.195) | (0.194) | (0.304) | (0.303) |

| | | | | |
|---|---|---|---|---|
| At least once a month | -0.378* | -0.406** | 0.044 | 0.042 |
| | (0.202) | (0.196) | (0.321) | (0.323) |
| Never | -0.375** | -0.355** | 0.038 | 0.038 |
| | (0.169) | (0.168) | (0.279) | 0.150 |
| Observations | 163 | 163 | 163 | 163 |
| Pseudo $R^2$ | 0.129 | 0.160 | 0.311 | 0.311 |

*** p<0.01, ** p<0.05, * p<0.1

a. Heteroscedasticity robust standard errors in parentheses under path coefficients.

b. None of the participants self-identified as "Native Hawaiian or Pacific Islander" or "Other." Therefore, these categories were omitted from the results.

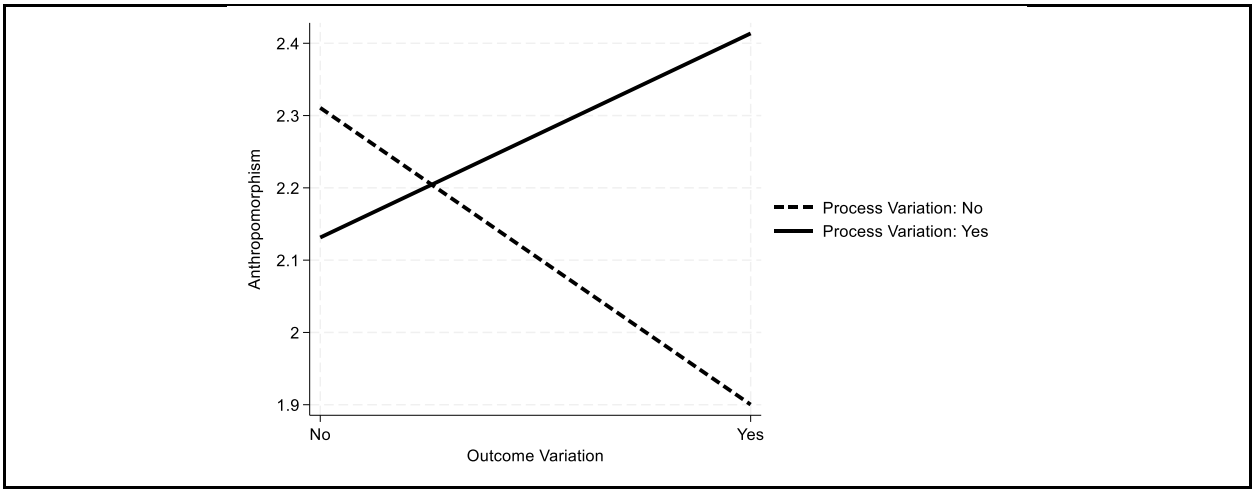



**Figure 5.** Marginal Effects of Experimental Conditions on Anthropomorphism

## Additional Analysis

### Investigating Failure Ratio

To confirm the soundness of our theoretical arguments, we conducted a secondary analysis on the participants who were assigned to the outcome variation experimental condition. Our theoretical development was based on the premise that people tend to make general judgments about the randomness of a sequence based on the observed patterns in the sequence, even when the sequence is short (Tversky & Kahneman, 1974). When the pattern does not "seem" random, they are likely to perceive even a random sequence as intentional. For example, if a person asks a CA to park their car several times and the CA fails most of the time but works sometimes or succeeds most of the time but fails sometimes, the person will likely think that the sudden failure or success cannot be just a coincidence. Therefore, patterns that happen to have "too many" or "too few" failures (or successes) should induce more anthropomorphism than patterns with the same number of successes and failures.

In other words, seemingly sudden anomalies in an otherwise consistent sequence of success or failure likely induce people to anthropomorphize the sequence-generating entity, i.e., the CA. The rationale is that by anthropomorphizing the CA, anomalies can be explained by the user's subjective inference about the CA's agency and ability to experience. A CA with agency can generate behaviors driven by autonomous choice making and a CA with the ability to experience can show performance fluctuations driven by its emotional states (Ebert & Wegner, 2011; Waytz, Gray, et al., 2010). In contrast, when the number of successes and failures are roughly equal, the sequence "looks" random. Therefore, people can easily explain away the variation by correctly assuming that it comes from an actual randomizer, not an intentional agent (Oskarsson et al., 2009). In conclusion, we expect to see a positive curvilinear relationship between failure ratio and anthropomorphism such that anthropomorphism is highest for low and high failure ratios. Failure ratio is the number of failures divided by the total number of attempts.

To test the soundness of this prediction, we estimated the relationship between anthropomorphism and trust and different levels of failure ratio using a separate regression (Adam et al., 2022), as per the following equation.

$$Anthropomorphism = \alpha_0 + \alpha_1 FailureRatio + \alpha_2 FailureRatio^2 + Controls + \varepsilon$$

As shown below, to test whether $\alpha_2$ influences trust in CA mediated through anthropomorphism, we also estimated the effect of anthropomorphism on trust in CA, controlling for failure ratio, age, sex, ethnicity, education, and past experience using AI assistants.

$$Trust = \beta_0 + \beta_1 FailureRatio + \beta_2 Anthropomorphism + Controls + \varepsilon$$

We estimated these two regressions as a system of regressions with heteroscedasticity robust standard errors. We followed Lind & Mehlum's (2010) three-step procedure and confirmed that: (a) the quadratic term was significant ($\alpha_2$*=10.971, se=5.013, p<0.05; 90% BootCI [0.188,21.754]*), (b) the turning point (i.e., $-\alpha_1/2\alpha_2$*=0.342*) fell within the observed range of failure ratio (i.e., 0 to 0.7), and (c) the slope of the relationship (i.e., $d(Anthropomorphism)/d(FailureRatio)=\alpha_1+2\alpha_2$ *FailureRatio*) is significantly negative ($\alpha$*=-7.495, se=3.681, p<0.05; 90% BootCI [-15.357,0.367]*) at the lower bound and significantly

positive at the upper bound ($\alpha$=7.864, se=3.461, p<0.05; 90% BootCI [0.382,15.346]) of the observed range of failure ratio.[9] These results indicate a positive curvilinear relationship such that anthropomorphism is at its highest levels for low and high failure ratios and at its lowest levels for average failure ratios (note that the average failure ratio is one-third by design).

A mediation analysis revealed that this curvilinear relationship influenced trust in CA, marginally mediated through anthropomorphism ($\alpha_2\beta_2$=2.832, se=2.501, 90% BootCI [0.162,10.436]). Table 3 and Figure 6 show the results.

**Table 3.** Hierarchical Regression Results for Failure Ratio

| | (1) Anthropomorphism | (2) Trust |
|---|---|---|
| **Independent Variables** | | |
| Anthropomorphism | - | 0.258** |
| | - | (0.130) |
| Failure Ratio | -7.495** | -6.816*** |
| | (3.681) | (0.823) |
| Failure Ratio$^2$ | 10.971** | - |
| | (5.013) | - |
| **Control Variables** | | |
| Constant | 2.380*** | 8.743*** |
| | (0.864) | (0.918) |
| Age | -0.002 | -0.014 |
| | (0.009) | (0.011) |
| Sex (Female=0) | | |
| Male | 0.037 | 0.209 |
| | (0.207) | (0.272) |
| Ethnicity (White=0)[d] | | |
| Black or African American | 0.272 | -1.219 |
| | (0.599) | (0.764) |
| Asian | 0.114 | -0.819** |
| | (0.509) | (0.344) |
| Latino or Hispanic | -0.435 | -2.467** |
| | (0.437) | (1.027) |
| Education (Less than High School = 0) | | |
| High School | 0.891 | -3.648*** |
| | (0.710) | (0.534) |
| Some College | 1.017 | -2.457*** |
| | (0.739) | (0.643) |
| 2-year College Degree | 1.709** | -2.243** |
| | (0.799) | (0.942) |
| 4-year College Degree | 1.069 | -2.945*** |
| | (0.674) | (0.543) |
| Master's Degree | 1.066 | -3.577*** |
| | (0.722) | (0.593) |
| Doctorate Degree | 1.307* | -4.812*** |
| | (0.746) | (0.596) |
| Past Experience Frequency (At least once a day=0) | | |
| At least once a week | -0.340 | 0.610* |
| | (0.231) | (0.354) |
| At least once a month | 0.004 | 0.777** |
| | (0.374) | (0.372) |

[9] Note that while we failed to find a significant negative slope at the lower bound when using bootstrapping confidence intervals, we estimated the difference between the slopes at the lower bound and upper bound using the same bootstrapped samples and confirmed that they were significantly different (*Slope Difference=15.359, 90% BootCI [0.263, 30.456]*).

| | | |
|---|---|---|
| Never | -0.439* | -0.039 |
| | (0.248) | (0.353) |
| Observations | 74 | 74 |
| Pseudo $R^2$ | 0.164 | 0.583 |

*** p<0.01, ** p<0.05, * p<0.1. The analysis was done on participants in the outcome variation condition.

a. Heteroscedasticity robust standard errors in parentheses under path coefficients

b. None of the participants self-identified as "Native Hawaiian or Pacific Islander," "American Indian or Alaska Native," or "Other." Therefore, these categories were omitted from the results.

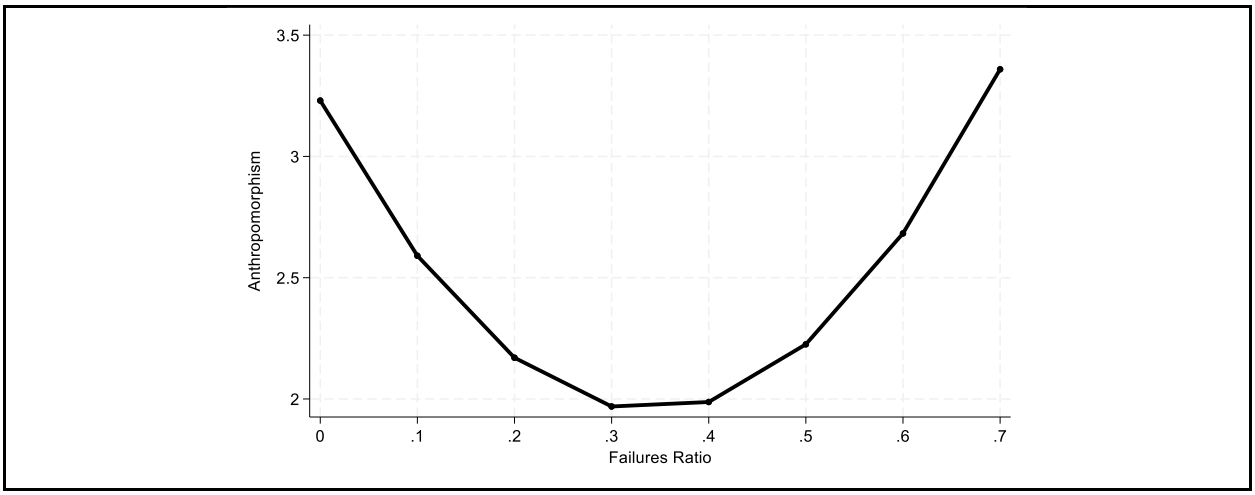



**Figure 6.** Marginal Effects of Failure Ratio on Anthropomorphism

## A Qualitative Measure of Anthropomorphism

To add further robustness to our results, we explored the participants' responses to an open-ended qualitative question that we asked them at the end of the experiment: "briefly explain your thought processes during your interaction with the agent." Exploring the data, we found interesting responses, which showed that people often rationalized the CA's outcome variation through different means such as perceiving it as learning and acting human-like (or lack thereof). For instance, the following is a qualitative response that indicates the person's justification of the CA's outcome variation and positive feelings toward the CA while questioning its competence:

> "After the third failed attempt at parking the car in the correct area, I did not have high hopes for this AI being effective. But then on the fourth try up until the 10th try, when the AI parked the car correctly, I gradually gained confidence in the AI's abilities. My overall feeling is somewhat positive towards it, in that once you give the AI time to learn and improve, it will perform effectively. BUT... if this were the real world and you had to deal with the AI improperly parking your car three times in a row, this would not be acceptable for me and most other people."

To be able to explore our qualitative data, we conducted a Latent Dirichlet allocation (LDA) analysis to cluster the responses into several coherent topics. We used the UMass coherence score method to find an optimum number of topics for our LDA model (Mimno et al.,

2011). As shown in Figure 7, using nine topics in our LDA model results in the highest absolute coherence score (please note that UMass coherence scores are logarithmic and negative).

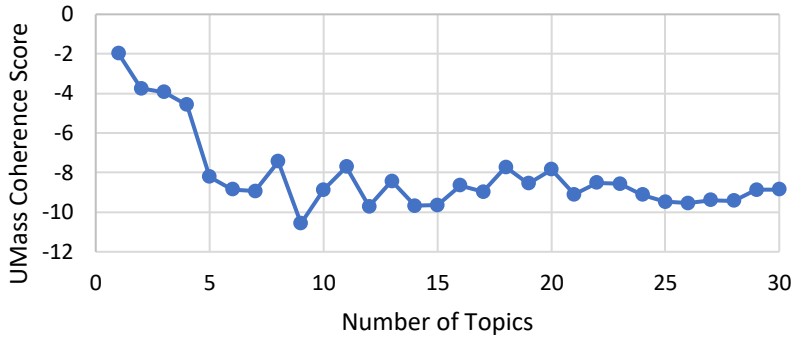



**Figure 7.** UMass Coherence Scores

We ran an LDA model using a Markov chain Monte Carlo with a collapsed Gibbs sampler (total samples: 1,000, burn-in samples: 600, topics: 9). For each response, our model yielded 9 values between 0 and 1, which indicated how likely it was that each topic was discussed in the response. After reviewing the words and responses that were highly scored on each of the topics, we labeled the topic. Table 4 shows the topic labels, the top 10 words associated with each topic, and a sample of a relatively short response that was highly scored on a given topic.

**Table 4.** LDA Results

| **Labels** | Anthropomorphism | Understanding the Interface | Failure | Physical Obstacles | Pattern | Wondering | Trustworthiness | Agent | Route |
|---|---|---|---|---|---|---|---|---|---|
| **Topic** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| **Top 10 Words for Each Topic** | like | didn't | park | like | time | would | try | Autopark | take |
| | think | go | car | drive | seem | parking | less | Drew | route |
| | AI | make | time | right | route | wonder | attempt | Amanda | turn |
| | something | way | would | seem | every | route | keep | draw | zig-zag |
| | human | easy | think | avoid | like | take | pattern | use | sometimes |
| | machine | park | AI | keep | pattern | space | determine | task | make |
| | game | know | try | felt | environment | spot | successful | complete | though |
| | feel | want | fail | straight | notice | also | certain | felt | didn't |
| | video | app | work | good | exact | get | many | efficient | think |
| | program | understand | correctly | line | terrain | see | course | seem | direct |
| **Sample Response** | I was thinking about how on the surface level it acted like a human being, but that in reality it was just an AI program. It's responses weren't dynamic enough for me to think otherwise. | That there wasn't enough interaction. It never seemed to want to know where or how I wanted it to park. (i.e. it never asked) It never let me know what its plans were before moving forward. | I found the experience. Through my trials, the car was correctly parked 7 out of 10 times. I'm still not sure why it failed to cooperate during the three failed trials. | I hope that it doesn't hit the building. It keeps running over the same brush. It's going to get a flat tire because of the stickers and the ground it keeps running over. | I noticed that he kept going in the exact same pattern each time that he drove. | I wondered what route she would take. I wondered if she'd hit the house. I wondered if she'd make it to the parking spot. I wondered if she'd take into consideration the terrain and what was the best route. | I was keeping track of successful vs unsuccessful park numbers to determine the reliability. I also begin looking for patterns, but didn't do that until midway into the test. | Amanda AutoPark would kill me if I were to use it. | It's not very good at its job. It's not taking a very efficient route (unnecessary zig-zags). It needs a lot more work before it's ready for public release. |

To augment our topic modeling, we calculated the sentiment of the responses. In doing so, we used Google Cloud Natural Language Sentiment Analyzer[10], which yielded a number between -100 and 100, corresponding to negative and positive sentiments respectively. To represent each participant's view on each topic, we multiplied their topic scores and sentiments. Based on our theorization and quantitative findings, we expected to see a significant interaction effect of process and outcome variation on the "anthropomorphism" topic. We also expected this interaction effect to indirectly influence the "trustworthiness" topic, mediated through the "anthropomorphism" topic.

We observed a significant interaction effect of process and outcome variation on the "anthropomorphism" topic (*β=3.602, se=1.656, p<0.05; 95% BootCI [0.166,7.038]*). We did not observe a similar effect on any of the other topics. We found that the observed interaction had a positive indirect impact on "trustworthiness" topic, mediated through the "anthropomorphism" topic (*β=0.333, se=0.197, 95% BootCI [0.045,0.878]*). Table 5 shows the regression results of our analysis of the "anthropomorphism" topic and Appendix G demonstrates the results for the other extracted topics.[11] These results indicate that our findings are robust when using alternative measures of anthropomorphism and trustworthiness.

**Table 5.** Hierarchical Regression Results for the LDA Topics

| | (1) Anthropomorphism Topic | (2) Anthropomorphism Topic | (3) Trustworthiness Topic | (4) Trustworthiness Topic |
|---|---|---|---|---|
| **Independent Variables** | | | | |
| Anthropomorphism | - | - | 0.090** | **0.093**** |
| | - | - | (0.039) | (0.038) |
| Process Variation | -0.314 | -1.960* | 0.381 | 0.507* |
| | (0.715) | (1.164) | (0.240) | (0.272) |
| Outcome Variation | 0.415 | -1.342 | -0.116 | 0.017 |
| | (0.803) | (0.981) | (0.291) | (0.451) |
| Process × Outcome | - | **3.602**** | - | -0.274 |
| | - | (1.656) | - | (0.478) |
| **Control Variables** | | | | |
| Constant | 0.573 | 2.235 | -0.394 | -0.522 |
| | (1.761) | (2.051) | (0.633) | (0.742) |
| Age | -0.015 | -0.010 | 0.016 | 0.016 |
| | (0.024) | (0.024) | (0.011) | (0.011) |
| Sex (Female=0) | | | | |
| Male | -0.247 | -0.269 | 0.035 | 0.038 |
| | (0.599) | (0.600) | (0.224) | (0.224) |
| Ethnicity (White=0)[c] | | | | |

[10] https://cloud.google.com/natural-language/docs/basics

[11] It is noteworthy that we repeated our analysis with 5 and 10 topics and the regression results remained the same in terms of significance and direction.

| | | | | |
|---|---|---|---|---|
| Black or African American | 0.338 | 0.811 | 0.239 | 0.202 |
| | (1.105) | (1.056) | (0.619) | (0.615) |
| American Indian or Alaska Native | 4.675** | 3.628* | 1.305** | 1.374** |
| | (2.187) | (1.983) | (0.541) | (0.611) |
| Asian | -0.626 | -1.038 | -0.101 | -0.068 |
| | (0.912) | (0.954) | (0.371) | (0.363) |
| Latino or Hispanic | -8.997 | -9.154 | 0.614 | 0.647 |
| | (8.014) | (7.868) | (0.682) | (0.681) |
| Education (Less than High School = 0) | | | | |
| High School | -0.895 | -1.799 | -1.083* | -1.012 |
| | (1.432) | (1.584) | (0.628) | (0.634) |
| Some College | 1.294 | 0.265 | -0.964 | -0.889 |
| | (1.364) | (1.309) | (0.642) | (0.712) |
| 2-year College Degree | -0.060 | -1.290 | -1.087* | -0.994 |
| | (2.512) | (2.320) | (0.647) | (0.719) |
| 4-year College Degree | 0.779 | -0.264 | -0.595 | -0.518 |
| | (1.419) | (1.362) | (0.629) | (0.710) |
| Master's Degree | -0.723 | -1.877 | -0.774 | -0.685 |
| | (1.412) | (1.568) | (0.527) | (0.608) |
| Doctorate Degree | 0.406 | -0.089 | -0.532 | -0.496 |
| | (1.710) | (1.726) | (0.643) | (0.683) |
| Past Experience Frequency (At least once a day=0) | | | | |
| At least once a week | -1.059 | -1.026 | -0.421 | -0.421 |
| | (0.883) | (0.867) | (0.292) | (0.288) |
| At least once a month | -1.171 | -1.315 | -0.442 | -0.428 |
| | (0.881) | (0.864) | (0.470) | (0.478) |
| Never | -2.629* | -2.524* | -0.179 | -0.181 |
| | (1.485) | (1.439) | (0.242) | (0.244) |
| Observations | 163 | 163 | 163 | 163 |
| Pseudo $R^2$ | 0.156 | 0.182 | 0.135 | 0.137 |

*** p<0.01, ** p<0.05, * p<0.1

a. Heteroscedasticity robust standard errors in parentheses under path coefficients.

b. None of the participants self-identified as "Native Hawaiian or Pacific Islander" or "Other." Therefore, these categories were omitted from the results.

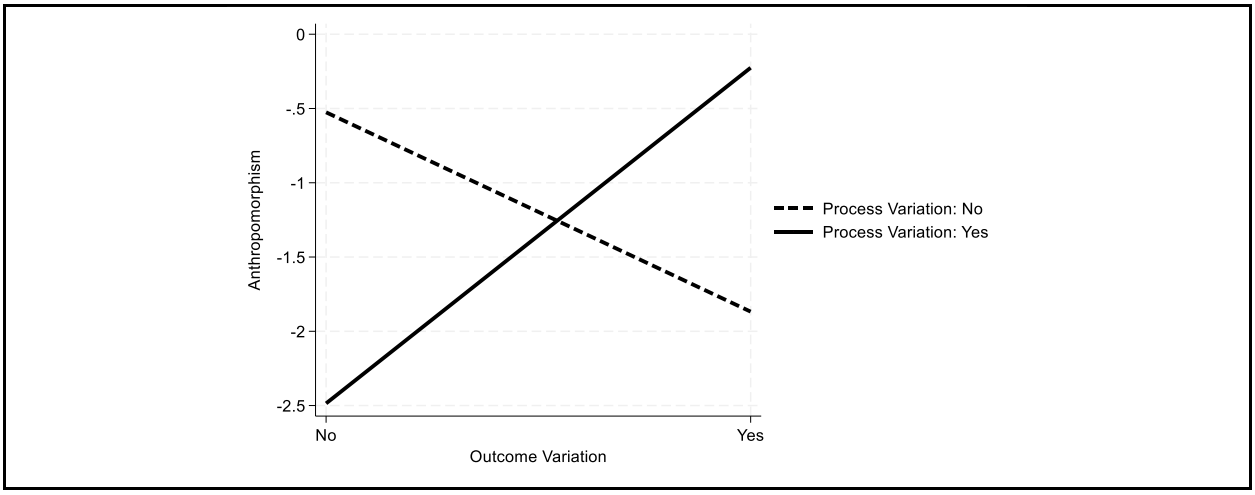



**Figure 3.** Marginal Effects of Experimental Conditions on Anthropomorphism

## DISCUSSION

In this research, we explored the effects of process and outcome variation on people's trust in a CA. More specifically, we investigated how the interplay of process and outcome variation affected trust via anthropomorphism.

Prior research yielded contradictory findings regarding the effect of outcome variation on anthropomorphism, with some studies finding a positive effect (Salem et al., 2013; Waytz, Morewedge, et al., 2010), but others failing to find an effect (Mirnig et al., 2017). Our results did not provide support for the positive effects of outcome and process variation on anthropomorphism (i.e., H1a and H1b were not supported). However, we found that process variation changed the effect of outcome variation on anthropomorphism from directionally negative to directionally positive, and this change was significant (i.e., H2 was supported). In other words, our results indicated that the effect of outcome variation on anthropomorphism depends on process variation. This finding suggests that people's perception of variation in CAs' behavior is more nuanced than was previously proposed in the literature (Mirnig et al., 2017; Salem et al., 2015).

In addition, we found that anthropomorphism was *positively* associated with trust in CA (H3) while outcome variation had a *negative* direct effect on trust in CA (H4a). A mediation analysis revealed the interaction of process and outcome variation had a *positive* indirect effect on trust in CA, mediated through anthropomorphism. These results were robust and replicated using alternative measures of anthropomorphism and trust.

In our secondary analyses, we found a curvilinear relationship between a CA's failure ratio and anthropomorphism such that anthropomorphism was highest for short sequences of user-CA interactions in which the CA had "too many" or "too few" failures. This is consistent with the notion that people are more likely to think that the source of a random sequence is a human agent than a nonhuman when the sequence does not "look" random (Caruso et al., 2010). We further found that this curvilinear relationship influenced trust, mediated via anthropomorphism. These findings have several implications for research and practice.

## Implications for Research

### Implications for Anthropomorphism Research

Our research contributes to the anthropomorphism literature by highlighting the falsification-prevention mechanism through which process variation moderates the effect of outcome variation on anthropomorphic attribution. When a CA varies its approaches, failures appear as exploratory behavior rather than systematic errors. This mechanism, likely rooted in the misconception of chance, explains why variation's effect on anthropomorphism is contingent rather than monolithic. Users perceive varied processes paired with variable outcomes as intentionally adaptive behavior because these patterns violate expectations of randomness, appearing too "streaky" or purposeful rather than properly alternating (Caruso et al., 2010; Gilovich et al., 1985). This violation likely triggers effectance motivation (the drive to explain and predict agent behavior) leading to anthropomorphic attribution as users apply their most accessible mental model for understanding complex, adaptive behavior (Epley et al., 2007; Epley, Waytz, et al., 2008; Waytz, Morewedge, et al., 2010).

By making a crucial distinction between process and outcome variation and focusing on how their interaction influences anthropomorphism in CAs, we resolve contradictions in prior literature. Some studies found positive effects of variation on anthropomorphism (Salem et al., 2013; Waytz, Morewedge, et al., 2010) while others found no effect (Mirnig et al., 2017). Our findings suggest these contradictions arise because researchers did not account for the interaction between different types of variation. We posit that process variation can give context to outcome variation, making it seem more attributable to human-like decision-making processes rather than random errors. This contextualization can shift the perception of outcome variation from being seen as flaws to being characteristic of a human-like entity.

This understanding opens new avenues for research into how AI developers can intentionally manipulate process variation to achieve a desired level of anthropomorphism. Future research should explore the boundary conditions of this effect: At what level does process variation become excessive and break the anthropomorphic illusion? How do individual

differences in tolerance for ambiguity, experience with technology, or need for control moderate these effects? The findings also raise questions about the ethical considerations of such manipulations, especially in scenarios where over-anthropomorphism could lead to misplaced trust or unrealistic expectations of the CA's capabilities.

Our research also reveals that while users might not explicitly acknowledge anthropomorphizing CAs, their responses regarding the CA's agency and ability to experience suggest an underlying attribution of human-like qualities (see Appendix F) (K. Gray et al., 2012; Waytz, Gray, et al., 2010). This finding aligns with recent human-robot interaction studies indicating a reluctance to admit anthropomorphism (Złotowski et al., 2018). This aspect suggests a complex psychological relationship between users and CAs that may not be fully conscious (Kim & Sundar, 2012). Future research could delve deeper into understanding this implicit anthropomorphism, exploring its implications for user behavior and the effectiveness of interactions with CAs.

**Implications for Trust Literature**

Our research contributes to trust literature by demonstrating that outcome variation can directly decrease trust in CAs, yet this effect can be moderated by process variation. When process variation is present, it can moderate the relationship between outcome variation and anthropomorphism and subsequently trust, potentially leading to a more positive perception. This dual nature of trust in response to behavioral variation in CAs presents a complex challenge for AI design. It suggests that since eliminating outcome variation in modern AI is all but impossible, how such variation is contextualized via process variation becomes vital. This insight could lead to more nuanced approaches in AI development, focusing on creating more 'human-like' error patterns or process variation that could foster trust even in the presence of errors.

This finding challenges traditional competence-based trust models (McKnight et al., 2011) by suggesting that anthropomorphism fundamentally alters how users evaluate CAs. When users anthropomorphize a CA through the interaction of process and outcome variation, they—at

least partially—shift from outcome-based to effort-based evaluation. Process variation signals that the CA is "trying" different approaches, transforming outcome failures from evidence of incompetence into evidence of human-like imperfection. This challenges the assumption that behavioral consistency universally signals trustworthiness, suggesting instead that in probabilistic AI contexts, variation paired with anthropomorphism may actually preserve trust through effort-based rather than outcome-based evaluation.

### Implications for Misconception of Chance Literature

Our study adds to the misconception of chance literature by showing why the misconception of chance influences the attribution of human characteristics to CAs. Our empirical results, particularly the curvilinear relationship between failure ratio and anthropomorphism, validate our theoretical proposition that violations of randomness expectations trigger anthropomorphic attribution. People tend to anthropomorphize CAs to rationalize the randomness in their actions, attributing agency and ability to experience to these nonhuman agents. This suggests that people's inherent need to find patterns and intentionality in randomness (Ebert & Wegner, 2011) plays a significant role in how they interact with and perceive CAs. Our study opens potential research areas in exploring how this misconception influences user acceptance, dependency, and the long-term relationship with CAs. Understanding such a misconception could be crucial in designing CAs that are both effective in their function and ethically aligned with human psychological tendencies.

## Implications for Practice

Since outcome variation is often not controlled by CA developers and is an inevitable side-effect of how modern CAs are designed, the role of more controllable aspects of AI agents such as process variation in managing the detrimental impacts of outcome variation becomes crucial. We discuss when developers should leverage process variation to improve users' experience and when they need to be cautious about the unwanted consequences of such an approach.

**Strategic Implementation of Process Variation.** Our research indicates that process variation can positively moderate the effect of outcome variation on anthropomorphism. In practice, this means developers can strategically implement process variation in scenarios where anthropomorphism might enhance user experience and trust. For instance, in customer service CAs (Schanke et al., 2021) or emotional support CAs (Broadbent, 2017), introducing variability in responses or interaction styles could make the CA seem more relatable and less mechanical. While we do not model usage intention directly, trust in CAs is a well-established predictor of intention to adopt, continue using, or rely on such systems (Glikson & Woolley, 2020; Saffarizadeh, Keil, & Maruping, 2024). Thus, our findings offer actionable guidance for how process variation may indirectly influence usage behavior through its effect on trust and anthropomorphism. In settings where user uptake and long-term engagement are critical, developers can leverage process variation not only to mitigate the negative effects of outcome variation but also to encourage sustained usage.

**Managing User Expectations in High-Stakes Environments.** In high-stakes environments like military or medical applications, it is crucial to manage user expectations regarding outcome variation (Hancock et al., 2011). While our findings show that process variation combined with outcome variation increases anthropomorphism, which can enhance trust, this may be problematic in high-stakes contexts. Research suggests that anthropomorphism might lead to biased assessments of blame (Waytz et al., 2014), potentially causing users to excuse system failures as "exploratory behavior" rather than recognizing genuine malfunctions. Hence, in these contexts, it is vital to minimize process variation to avoid inappropriate anthropomorphism and ensure users correctly attribute responsibility for outcomes and maintain appropriate vigilance for system errors.

**Redefining Algorithm Aversion Strategies.** Given that process variation could attenuate the negative effect of outcome variation on trust, developers might consider incorporating controlled process variation to combat algorithm aversion. This could complement other methods

such as enhancing transparency (Mikalef et al., 2022; Rai, 2020) and allowing user control over AI outputs (Dietvorst et al., 2018).

**Ethical Implications of Intentional Anthropomorphism.** The findings raise ethical questions about intentionally designing CAs to induce anthropomorphism, particularly in contexts where over-reliance on AI could have serious consequences (Banker & Khetani, 2019). Developers and policymakers should think about these implications and consider establishing ethical guidelines for anthropomorphism in AI design.

## Limitations and Future Directions

As is the case with all experiments, we should be cautious when generalizing the results of this study for a few reasons. First, we used a controlled experiment to test our hypotheses. While experiments are considered the best method to establish causality, they require a parallel design across experimental groups to rule out alternative explanations (Shadish et al., 2002). To do so, we chose not to rely on a machine learning model in our CA because such models would introduce uncontrollable inconsistencies across experimental groups. Instead, we designed and used a CA that behaved identically in all experimental groups except in their manipulated aspects. We believe this design was suitable for our purposes because most participants do not understand how a CA works, even in real settings. However, future research can add to the external validity of our findings by conducting correlational studies of people's daily interactions with CAs that demonstrate process and outcome variation.

Second, we assumed people have repeated interactions with CAs. This assumption may be typically true and essential for a person to observe a level of process and outcome variation in a CA. However, there may be other scenarios in which people have a one-off interaction with the CA. Future research is needed to investigate people's judgment of CAs in one-off interactions.

Third, we found a curvilinear relationship between failure ratio and anthropomorphism in our study. Both low and high failure ratios led to higher anthropomorphism, possibly because they seem less random, suggesting intentional actions by the CA. This finding challenges the

notion of "to err is human" (Salem et al., 2013), as no errors should theoretically result in less anthropomorphism. A possible reason for this could be participants' expectations of human performance in similar tasks. For instance, if they expect humans to make no errors in a task, a CA also making no errors might be seen as more human-like. Future research could investigate how anthropomorphism is influenced by expectations of human failure rates in different tasks and contexts.

While our study focuses on trust, we acknowledge that distrust and ambivalence represent important and distinct psychological states that may also shape user reactions to CAs, particularly in sensitive domains involving privacy, fairness, or safety concerns. Prior work has shown that trust and distrust are not simply opposites on a single continuum but may coexist or emerge independently depending on context (Dimoka, 2010; Lewicki et al., 1998; Moody et al., 2017). Our emphasis on trust aligns with our theoretical objective of examining the role of anthropomorphism and behavioral variation in shaping users' willingness to rely on CAs in agentic roles. In task-specific domains such as autonomous parking, where the agent acts directly on behalf of the user to fulfill a specific task, the trust construct is deemed to be well-suited for capturing users' reliance judgments (Schoorman et al., 2007). Nevertheless, future research would benefit from examining how outcome and process variation might also activate distrust (i.e., negative expectations about the CA's motives or reliability) or ambivalence (i.e., holding simultaneous, conflicting attitudes of similar magnitude), especially in multifaceted, high-stakes or ethically charged applications such as healthcare, financial advising, or surveillance.

Finally, we chose the context of CAs in autonomous driving for our experiment. The rationale behind this selection was threefold: (a) it enabled the CA to serve in a clearly agentic role—executing an action (parking a car) directly on behalf of the user, (b) it provided a clear and effective means to operationally distinguish between outcome and process variations, and (c) autonomous and semi-autonomous driving was a highly visible technological topic in mainstream media at the time of the experiment, which enhanced the plausibility of the scenario and supported participant engagement. However, considering recent advancements in LLM-

based CAs, exploring additional contexts such as automatic trading or code assistance presents valuable opportunities. For instance, an agentic automatic trading AI could achieve outcomes with varying levels of alignment to users' desired results and might employ diverse strategies to accomplish identical outcomes. Similarly, a code-assistance CA might solve problems exhibiting different degrees of alignment with users' intended outcomes and might utilize various approaches or algorithms to achieve comparable outcomes.

## REFERENCES


Adam, M., Croitor, E., Werner, D., Benlian, A., & Wiener, M. (2022). Input control and its signalling effects for complementors' intention to join digital platforms. *Information Systems Journal*.

Adam, M., Roethke, K., & Benlian, A. (2023). Human vs. Automated Sales Agents: How and Why Customer Responses Shift Across Sales Stages. *Information Systems Research*, *34*(3), 1148–1168. https://doi.org/10.1287/isre.2022.1171

Baird, A., & Maruping, L. M. (2021). The Next Generation of Research on IS Use: A Theoretical Framework of Delegation to and from Agentic IS Artifacts. *MIS Quarterly*, *45*(1b), 315–341.

Banker, S., & Khetani, S. (2019). Algorithm overdependence: How the use of algorithmic recommendation systems can increase risks to consumer well-being. *Journal of Public Policy & Marketing*, *38*(4), 500–515.

Barrett, J., & Johnson, A. H. (2003). The Role of Control in Attributing Intentional Agency to Inanimate Objects. *Journal of Cognition and Culture*, *3*(3), 208–217. https://doi.org/10.1163/156853703322336634

Bauer, K., & Gill, A. (2024). Mirror, Mirror on the Wall: Algorithmic Assessments, Transparency, and Self-Fulfilling Prophecies. *Information Systems Research*, *35*(1), 226–248. https://doi.org/10.1287/isre.2023.1217

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

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), 1010–1042.

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, *45*(3), 1433–1450.

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

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34.

Bigman, Y. E., Wilson, D., Arnestad, M. N., Waytz, A., & Gray, K. (2022). Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Experimental Psychology: General*, *152*(1), 4–27.

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

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.

Burgoon, J. K., Bonito, J. A., Bengtsson, B., Ramirez, A., Dunbar, N. E., & Miczo, N. (1999). Testing the Interactivity Model: Communication Processes, Partner Assessments, and the Quality of Collaborative Work. *Journal of Management Information Systems*, *16*(3), 33–56. https://doi.org/10.1080/07421222.1999.11518255

Burton, J. W., Stein, M.-K., & Jensen, T. B. (2019). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*(2), 220–239.

Caruso, E. M., Waytz, A., & Epley, N. (2010). The intentional mind and the hot hand: Perceiving intentions makes streaks seem likely to continue. *Cognition*, *116*(1), 149–153.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809–825.

Chan, E. Y. (2020). Political conservatism and anthropomorphism: An investigation. *Journal of Consumer Psychology*, *30*(3), 515–524.

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

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

Clotfelter, C. T., & Cook, P. J. (1993). The "gambler's fallacy" in lottery play. *Management Science*, *39*(12), 1521–1525.

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

Dang, J., & Liu, L. (2023). Do lonely people seek robot companionship? A comparative examination of the Loneliness–Robot anthropomorphism link in the United States and China. *Computers in Human Behavior*, *141*, 107637.

Diederich, S., Brendel, A. B., Morana, S., & Kolbe, L. (2022). On the design of and interaction with conversational agents: An organizing and assessing review of human-computer interaction research. *Journal of the Association for Information Systems*, *23*(1), 96–138.

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

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.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170.

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

DuBrow, S., & Davachi, L. (2013). The influence of context boundaries on memory for the sequential order of events. *Journal of Experimental Psychology: General*, *142*(4), 1277.
Ebert, J. P., & Wegner, D. M. (2011). Mistaking randomness for free will. *Consciousness and Cognition*, *20*(3), 965–971. https://doi.org/10.1016/j.concog.2010.12.012
Epley, N., Akalis, S., Waytz, A., & Cacioppo, J. T. (2008). Creating social connection through inferential reproduction: Loneliness and perceived agency in gadgets, gods, and greyhounds. *Psychological Science*, *19*(2), 114–120. https://doi.org/10.1111/j.1467-9280.2008.02056.x
Epley, N., Waytz, A., Akalis, S., & Cacioppo, J. T. (2008). When we need a human: Motivational determinants of anthropomorphism. *Social Cognition*, *26*(2), 143–155. https://doi.org/10.1521/soco.2008.26.2.143
Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864
Eyssel, F., & Reich, N. (2013). Loneliness makes the heart grow fonder (of robots)—On the effects of loneliness on psychological anthropomorphism. *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 121–122.
Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127. https://doi.org/10.1038/nrn2787
Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, *30*(3), 411–437.
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.
Gartner. (2023). *Reviews for Enterprise Conversational AI Platforms Reviews 2023 | Gartner Peer Insights*. Gartner. https://www.gartner.com/market/enterprise-conversational-ai-platforms
Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, *17*(3), 295–314.
Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660.
Gong, L. (2008). How social is social responses to computers? The function of the degree of anthropomorphism in computer representations. *Computers in Human Behavior*, *24*(4), 1494–1509.
GoogleCloud. (2023). *Responses | Dialogflow ES*. Google Cloud. https://cloud.google.com/dialogflow/es/docs/intents-responses
Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619. https://doi.org/10.1126/science.1134475
Gray, K., Knobe, J., Sheskin, M., Bloom, P., & Barrett, L. F. (2011). More than a body: Mind perception and the nature of objectification. *Journal of Personality and Social Psychology*, *101*(6), 1207. https://doi.org/10.1037/a0025883
Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*(2), 101–124.
Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *53*(5), 517–527. https://doi.org/10.1177/0018720811417254

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

Hayes, A. F., & Preacher, K. J. (2014). Statistical mediation analysis with a multicategorical independent variable. *British Journal of Mathematical and Statistical Psychology*, *67*(3), 451–470. https://doi.org/10.1111/bmsp.12028

Hodgson, H. (2024). *Agentic AI Is the Perfect Fit for Automotive*. https://www.abiresearch.com/market-research/insight/7785118-agentic-ai-is-the-perfect-fit-for-automoti

Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards Algorithms? A comprehensive literature Review on Algorithm aversion. *ECIS*.

Jussupow, E., Benbasat, I., & Heinzl, A. (2024). An integrative perspective on algorithm aversion and appreciation in decision-making. *MIS Quarterly*, *48*(4), 1575–1590.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454.

Kelley, H. H. (1967). Attribution theory in social psychology. *Nebraska Symposium on Motivation*.

Kim, Y., & Sundar, S. S. (2012). Anthropomorphism of computers: Is it mindful or mindless? *Computers in Human Behavior*, *28*(1), 241–250. https://doi.org/10.1016/j.chb.2011.09.006

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

Kordzadeh, N., & Ghasemaghaei, M. (2021). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems*, 1–22.

Kreye, M. E., Goh, Y. M., & Newnes, L. B. (2011). Manifestation of uncertainty-A classification. *DS 68-6: Proceedings of the 18th International Conference on Engineering Design (ICED 11), Impacting Society through Engineering Design, Vol. 6: Design Information and Knowledge, Lyngby/Copenhagen, Denmark, 15.-19.08. 2011*.

Lankton, N. K., McKnight, D. H., & Tripp, J. (2015). Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems*, *16*(10), 880–918. https://doi.org/10.17705/1jais.00411

Lewicki, R. J., McAllister, D. J., & Bies, R. J. (1998). Trust and distrust: New relationships and realities. *Academy of Management Review*, *23*(3), 438–458.

Lind, J. T., & Mehlum, H. (2010). With or Without U? The Appropriate Test for a U-Shaped Relationship*: Practitioners' Corner. *Oxford Bulletin of Economics and Statistics*, *72*(1), 109–118. https://doi.org/10.1111/j.1468-0084.2009.00569.x

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*(10), 464–470. https://doi.org/j.tics.2006.08.004

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

McCracken, H. (2023, January 11). *If ChatGPT doesn't get a better grasp of facts, nothing else matters*. Fast Company. https://www.fastcompany.com/90833017/openai-chatgpt-accuracy-gpt-4

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

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research*, *13*(3), 334–359. https://doi.org/10.1287/isre.13.3.334.81

Mikalef, P., Conboy, K., Lundström, J. E., & Popovič, A. (2022). Thinking responsibly about responsible AI and 'the dark side'of AI. In *European Journal of Information Systems* (Vol. 31, Issue 3, pp. 257–268). Taylor & Francis.

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 262–272.

Mirnig, N., Stollnberger, G., Miksch, M., Stadler, S., Giuliani, M., & Tscheligi, M. (2017). To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers in Robotics and AI*, *4*, 21.

Moody, G. D., Galletta, D. F., & Lowry, P. B. (2014). When trust and distrust collide online: The engenderment and role of consumer ambivalence in online consumer behavior. *Electronic Commerce Research and Applications*, *13*(4), 266–282.

Moody, G. D., Lowry, P. B., & Galletta, D. F. (2017). It's complicated: Explaining the relationship between trust, distrust, and ambivalence in online transaction relationships using polynomial regression analysis and response surface analysis. *European Journal of Information Systems*, *26*(4), 379–413. https://doi.org/10.1057/s41303-016-0027-9

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

Nass, C., Moon, Y., & Green, N. (1997). Are machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of Applied Social Psychology*, *27*(10), 864–876. https://doi.org/10.1111/j.1559-1816.1997.tb00275.x

Nearsure. (2024). *What's an AI Hallucination and its Impact on the Tech Industry | Nearsure*. https://www.nearsure.com/blog/ai-hallucinations-tech-impact

Nunamaker, J. F., Derrick, D. C., Elkins, A. C., Burgoon, J. K., & Patton, M. W. (2011). Embodied conversational agent-based kiosk for automated interviewing. *Journal of Management Information Systems*, *28*(1), 17–48.

OpenAI. (2023). *GPT-4 Technical Report* (No. arXiv:2303.08774). arXiv. https://doi.org/10.48550/arXiv.2303.08774

Oskarsson, A. T., Van Boven, L., McClelland, G. H., & Hastie, R. (2009). What's next? Judging sequences of binary events. *Psychological Bulletin*, *135*(2), 262.

Prillaman, M. (2024). Is ChatGPT making scientists hyper-productive? The highs and lows of using AI. *Nature*, *627*(8002), 16–17.

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

Rai, A. (2020). Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141.

Riedl, R., Hubert, M., & Kenning, P. (2010). Are there neural gender differences in online trust? An fMRI study on the perceived trustworthiness of eBay offers. *MIS Quarterly*, *34*(2), 397–428. https://doi.org/10.2307/20721434

Riedl, R., Mohr, P. N. C., Kenning, P. H., Davis, F. D., & Heekeren, H. R. (2014). Trusting Humans and Avatars: A Brain Imaging Study Based on Evolution Theory. *Journal of Management Information Systems*, *30*(4), 83–114. https://doi.org/10.2753/MIS0742-1222300404

Saffarizadeh, K., Keil, M., Boodraj, M., & Alashoor, T. (2024). "My Name is Alexa. What's Your Name?" The Impact of Reciprocal Self-Disclosure on Post-Interaction Trust in Conversational Agents. *Journal of the Association for Information Systems*, *25*(3), 528–568.

Saffarizadeh, K., Keil, M., & Maruping, L. (2024). Relationship Between Trust in the AI Creator and Trust in AI Systems: The Crucial Role of AI Alignment and Steerability. *Journal of Management Information Systems*, *41*(3), 645–681. https://doi.org/10.1080/07421222.2024.2376382

Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., & Joublin, F. (2013). To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics*, *5*(3), 313–323.

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015). Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust. *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 1–8.

Salesforce. (2025). *What Are Autonomous Agents?* Salesforce. https://www.salesforce.com/agentforce/autonomous-agents/

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

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

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

Schuetz, S., & Venkatesh, V. (2020). Research Perspectives: The Rise of Human Machines: How Cognitive Computing Systems Challenge Assumptions of User-System Interaction. *Journal of the Association for Information Systems*, *21*(2), 460–482. https://doi.org/10.17705/1jais.00608

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

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

Sohail, M., Wang, F., Archer, N., Wang, W., & Yuan, Y. (2024). Lower than expected but still willing to use: User acceptance toward current intelligent conversational agents. *Information & Management*, *61*(8), 104033. https://doi.org/10.1016/j.im.2024.104033

Spherical Insights. (2024). *Chatbot Market Size, Share, Growth Forecast 2033*. https://www.sphericalinsights.com/reports/chatbot-market

Srivastava, S. C., & Chandra, S. (2018). Social Presence in Virtual World Collaboration: An Uncertainty Reduction Perspective Using a Mixed Methods Approach. *MIS Quarterly*, *42*(3), 779–803. https://doi.org/10.25300/MISQ/2018/11914

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

Tarafdar, M., Page, X., & Marabelli, M. (2022). Algorithms as co-workers: Human algorithm role interactions in algorithmic work. *Information Systems Journal*.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, *185*(4157), 1124–1131.

Valdesolo, P., & Graham, J. (2014). Awe, uncertainty, and agency detection. *Psychological Science*, *25*(1), 170–178.

Wang, W., & Benbasat, I. (2007). Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Management Information Systems*, *23*(4), 217–246.

Wang, W., & Benbasat, I. (2013). Research note—A contingency approach to investigating the effects of user-system interaction modes of online decision aids. *Information Systems Research*, *24*(3), 861–876.

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

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

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

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

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

White, R. W. (1959). Motivation reconsidered: The concept of competence. *Psychological Review*, *66*(5), 297. https://doi.org/10.1037/h0040934

Xu, J., Benbasat, I., & Cenfetelli, R. T. (2014). The nature and consequences of trade-off transparency in the context of recommendation agents. *MIS Quarterly*, *38*(2), 379–406.

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, *32*(4), 403–414.

You, S., Yang, C. L., & Li, X. (2022). Algorithmic versus Human Advice: Does Presenting Prediction Performance Matter for Algorithm Appreciation? *Journal of Management Information Systems*, *39*(2), 336–365.

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

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

Zheng, J. F., & 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.

Złotowski, J., Sumioka, H., Eyssel, F., Nishio, S., Bartneck, C., & Ishiguro, H. (2018). Model of dual anthropomorphism: The relationship between the media equation effect and implicit anthropomorphism. *International Journal of Social Robotics*, *10*, 701–714.