Received 21 February 2022
Revised 25 August 2022
14 December 2022
4 July 2023
Accepted 14 May 2024

"Sorry, I Don't Understand …": effects of task type, personality presentation and performance error on user trust of a personal digital assistant

Xinyi Zhang
Department of Communication, University of California Santa Barbara, Santa Barbara, California, USA, and
Sun Kyong Lee
School of Media and Communication, Korea University, Seoul, South Korea

Abstract

Purpose – Based on the theoretical predictions of media equation theory and the computers-are-social-actors (CASA) perspective, this study aims to examine the effects of performance error type (i.e. logical, semantic or syntactic), task type and personality presentation (i.e. dominant/submissive and/or friendly/unfriendly) on users' level of trust in their personal digital assistant (PDA), Siri.

Design/methodology/approach – An experimental study of human–PDA interactions was performed with two types of tasks (social vs functional) randomly assigned to participants (N = 163). While interacting with Siri in 15 task inquiries, the participants recorded Siri's answers for each inquiry and self-rated their trust in the PDA. The answers were coded and rated by the researchers for personality presentation and error type.

Findings – Logical errors were the most detrimental to user trust. Users' trust of Siri was significantly higher after functional tasks compared to social tasks when the effects of general usage (e.g. proficiency, length and frequency of usage) were controlled for. The perception of a friendly personality from Siri had an opposite effect on social and functional tasks in the perceived reliability dimension of trust and increased intensity of the presented personality reduced perceived reliability in functional tasks.

Originality/value – The research findings contradict predictions from media equation theory and the CASA perspective while contributing to a theoretical refinement of machine errors and their impact on user trust.

Keywords Personal digital assistant, Perception, Reliability, Trust, Computers

Paper type Research paper

1. Introduction

The last decade has witnessed a drastic rise in the popularity of personal digital assistants (PDAs; e.g. Siri, Alexa and Cortana) that interact with humans conversationally. From setting an alarm to telling a joke, PDAs perform a wide range of tasks and have gradually transformed from novel to commonly used interfaces. A recent marketing report revealed that approximately 70% of Americans have some experience with voice-based technologies (Microsoft, 2019), however, user trust remains a concern hampering human interactions with PDAs. Indeed, improvements are needed to reach the true potential of this communication technology.

Funding: The study was supported by the Ministry of Education of the Republic of Korea, the National Research Foundation of Korea (NRF-2019S1A3A2099973).

Information Technology & People
Vol. 37 No. 8, 2024
pp. 154-174
Emerald Publishing Limited
0959-3845
DOI 10.1108/ITP-01-2022-0058

User trust is critical for the adoption and appropriate use of technology (Lewis *et al.*, 2018). This study investigates three factors that influence post-interaction user trust in PDAs: (1) performance error type, (2) task type and (3) personality presentation. The research builds on prior empirical findings in the field of human–machine communication (HMC) and theoretical propositions of media equation (Reeves and Nass, 1996). First, this study analyzes three basic error types (i.e. logical, semantic and syntactic) adapted from categorizations in computer science (McCall and Kölling, 2014) and their distinct effects on user trust in Siri. Next, this study examines how the performance of a voice-enabled PDA, *Siri*, affects user trust in two task types: functional and social tasks (Gaudiello *et al.*, 2016). Finally, it also tests hypotheses about preferred personality types (i.e. dominant vs submissive and friendly vs unfriendly) and personality identifiability (i.e. intensity and consistency) from media equation theory (Reeves and Nass, 1996), with the effects of personality presentation explored under both task types.

2. Literature review and hypotheses

2.1 Prevalence of PDAs

Voice-based PDAs have become an indispensable part of people's daily routines, supporting various aspects of life including but not limited to information seeking, online shopping, workload management, social networking, healthcare, education and entertainment. According to Morgan's (2019) estimation, there are over 90 million monthly active users of mobile PDAs and 45.7 million smart home speakers in the USA. As a subset of artificial intelligence (AI), PDAs, sometimes called virtual personal assistants, intelligent virtual assistants or intelligent personal assistants, can be defined as systems capable of autonomously interpreting and utilizing external data (Kaplan and Haenlein, 2019) and that deal with tasks on a personal level. As humanized AIs, PDAs pursue cognitive, emotional and social intelligence (Spencer *et al.*, 2018) to bolster effective interactions with users.

In 2019, Microsoft reported that 72% of US survey respondents had used conversational AI, and tech giants such as Google, Apple, Amazon and Microsoft dominated the marketplace. From their perspective, this innovation is still in its early adoption stage, with infinite business opportunities foretelling a new era driven by voice commands. Among various PDAs, Apple's Siri has emerged as a pioneer in the field. Despite its success, the troubles Siri faces as a major PDA service provider epitomize the difficulties that the whole industry needs to overcome. Although it has active usage on over 500 million devices (Apple, 2018), Siri struggles with serious challenges such as a relatively high chance of error occurrence (Berdasco *et al.*, 2019; Enge, 2019), bottlenecks of product innovation and rapid growth of its opponents' market share. Additionally, strong concerns about user trust persist: 52% of participants reported uneasiness about information security and privacy (Microsoft, 2019). Therefore, finding solutions to promote user trust has been a vital issue for Siri's development team. Focusing on user trust in PDAs under the context of Apple's Siri, this study contributes to the extant knowledge by demonstrating how performance errors, two types of everyday interactions (i.e. social vs functional tasks) and PDA personality presentation in task strings may impact user trust, eventually providing practical implications for improving user experience and refining interface designs.

2.2 User trust

Scholars have examined various dimensions of human trust in machines. Similar to trust fostered in interpersonal relationships, technology-based trust is jointly shaped by factors influencing three components: (1) dispositional trust (e.g. culture, age and personality), (2) situational trust (e.g. system types, task difficulty and mood) and (3) learned trust (performance reliability, predictability, error timing and technological features; Hoff and Bashir, 2015). Because dispositional trust is developed as an enduring trait preceding specific interaction episodes above and beyond a single technology product, it is less central to the inquiry of the present study. We are instead interested in trust formed based on specific user interactions with Siri, encompassing situational trust (i.e. the influences of specific interaction context on trust) and learned trust (i.e. trust specifically related to interactions with one system) of Siri (Hoff and Bashir, 2015). In particular, this study explores the impact of task type on situational trust as well as different interaction characteristics (i.e. errors and personality presentation) on the learned trust of Siri with an in-lab experiment.

Akin to human-to-human trust, human trust in machines also has multiple dimensions. In the field of HMC research, trust is often conceptualized as a bifactor structure, which includes cognitive and affective trust (Lewis *et al.*, 2018). Whereas cognitive trust is based on rational evaluations of PDAs' competence and relevant information, affective trust encompasses trust resulting from the relationships that users have developed with PDAs (Pal *et al.*, 2022). Past research in automation has a heavy emphasis on rational evaluations of machine performance and competence (Lewis *et al.*, 2018). However, as the capabilities of AI grow, building social relationships with more intelligent machines becomes possible, so affect-based trust now plays an equally, if not more, critical role in the process (Kyung and Kwon, 2022). Therefore, it is important to consider both cognitive and affective dimensions in trust measurement. Because we are primarily interested in how Siri's performance impacts user trust, morality-related issues concerning privacy and security are beyond the current research scope.

2.3 Performance errors

System errors, defined as "system states (electrical, logical, or mechanical) that can lead to a failure" caused by "faults" (Honig and Oron-Gilad, 2018), are ubiquitous in everyday usage of technology. These errors result from imperfections in technical designs and operations, which can potentially lead to trust violations if recognized by users. Error occurrence has been a major obstacle to PDA acceptance because the current stage of PDA design is far from perfection, and Siri has left the public an impression of making comparatively high rates of errors. According to Enge (2019), Siri had the second lowest percentage of providing fully correct and complete answers among seven PDA products, and its rank has declined for three consecutive years. Similarly, in Berdasco *et al.* (2019), Siri took the last place in answer correctness and quality.

A machine's task performance is one of the most influential determinants of user trust (Hancock *et al.*, 2011, 2021), with errors being deleterious to task performance. Prior research has supported that system failures have an adverse impact on user assessment under diverse contexts, including user trust of informational or transactional websites (Corritore *et al.*, 2003), automatics (De Vries *et al.*, 2003) and robots (Salem *et al.*, 2015). Although human interactions with computers, automations or robots can differ from ones with PDAs, the literature suggests meaningful negative relationships between system errors and user trust.

H1. The number of Siri's incorrect responses will be negatively related to the level of user trust.

To describe and explore such system flaws, prior research has proposed various error typologies, incorporating human-, machine- and environment-introduced faults (see Honig and Oron-Gilad, 2018 for a review). Technical errors are commonly categorized based on a locus of causes. For example, Parasuraman *et al.* (2000) divided the origins of faults by information acquisition, information analysis, decision/action selection and action implementation. Carlson and Murphy (2005) categorized failures by interaction, algorithms/methods and software design/implementation. Brooks (2017) classified errors according to whether they emerge in communication or processing. Yet despite the plethora of theoretical classifications, human-factor studies have rarely applied them in experimental designs. One reason is that mechanism-based typologies can be highly context-specific, so it is hard to examine them across different types of technology. Additionally, it can be difficult to apply mechanism-based error categories because lay users most likely cannot precisely speculate the causes of errors. That is, when Siri makes a mistake, users have difficulties identifying whether the error is caused by human input, algorithms, hardware or software, and they are more likely to judge response errors merely based on the answers they receive. As a result, the existing categorizations cannot easily satisfy our research needs.

False alarms and mistakes in automation have so far been the error category that has attracted the most scholarly interest (e.g. Davenport and Bustamante, 2010; Guznov *et al.*, 2016; Johnson *et al.*, 2004; Rovira and Parasuraman, 2010), but errors in everyday technology usage certainly exceed alarm errors. Since Siri is fundamentally different from automation alarm systems because of its machinery and task types, these findings do not easily transfer to the context of using Siri. Thus, we introduce an error categorization that can afford a greater variety of error responses than the dichotomous outcomes (i.e. activation or inactivation) led by the division between commission vs omission errors.

Three broad categories of machine errors conceptualized in computer science (McCall and Kölling, 2014) commonly occur in PDAs' performance: logical errors, which engender relevant but wrong results (e.g. Siri responds to "What is a 20% tip on $437?" with "$98.60"); semantic errors, which produce irrelevant and meaningless results (e.g. Siri responds to "Are you a robot?" with "What's the name of the app you want to launch?"); and syntactic errors, which involve a failure to operate the program (e.g. Siri responds to "Where is Starbucks?" with "Seems like I cut you off. Can you please repeat that?"). These three categories have some overlap with the conceptualizations of commission and omission errors but provide a more detailed classification of system errors. Moreover, the categorization assumes an output-centered orientation and highlights the discrepancy between expected and actual output, based on which laypeople can make judgments. To our knowledge, previous research has not directly examined these three types of errors, to our knowledge, in the context of PDA usage. Accordingly, we explore the following research question (RQ):

RQ. What kinds of errors have more negative effects on user trust when Siri makes mistakes?

2.4 Task type

Machines can assist humans with a variety of tasks, and scholars have proposed copious task typologies such as revocable vs irrevocable actions (Salem *et al.*, 2015), achievement vs maintenance tasks (Thórisson *et al.*, 2016), functional vs social tasks (Gaudiello *et al.*, 2016), social vs analytical tasks (Smith *et al.*, 2016), physical vs virtual service, tangible vs intangible actions (Wirtz *et al.*, 2018) and simple vs complex tasks (Guo *et al.*, 2020). Given the simplicity of PDAs, the classification of functional vs social tasks provides the most parsimonious applicability here. According to Gaudiello *et al.* (2016), functional tasks test an agent's ability to "efficiently operate by ensuring useful and accurate performances with relation to the functions it was designed for," such as searching for certain information for the user, while social tasks require an agent's capability of fitting into "the social structures and activities of a given context" (p. 635), such as exchanging jokes with the user. These two types of tasks form the basis for users' everyday interactions with PDAs, yet it remains unclear how they may exert impacts on user trust in similar and/or different manners.

Goetz *et al.* (2003) and Smith *et al.* (2016) confirmed that machines' humanness levels achieve their best effects when matched with appropriate task types. Goetz *et al.* reported that humanlike robotics are preferred when tasks require more social skills, and Smith *et al.* further supported this viewpoint by finding that agents appearing less humanlike elicited more compliance in functional tasks but vice versa for social tasks. Since Siri's appearance is more machine-like than humanlike (i.e. an interface displaying a dynamic sound wave, not a human face), we predict the effects of task type will be consistent with the existing findings about machines with low humanness displays.

H2. Users' level of trust in Siri will be higher after they perform functional tasks compared to social tasks with Siri's assistance.

2.5 Machine personality

2.5.1 Media equation and inequality. Extant research on machines' personality presentation is largely built on media equation theory (Reeves and Nass, 1996) and the computers-as-social-actors (CASA) paradigm (Nass and Moon, 2000), which have been the primary guidelines for PDA development. After conducting a series of in-lab studies in 1980s and 1990s concerning the effects of manners, personality, emotions, social roles and physical forms in human–media interactions, Reeves and Nass (1996) argued that "individuals' interactions with computers, televisions, and new media are fundamentally social and natural" (p. 5), an idea that was labeled as media equation. Media equation theory proposes that when mindlessly delegating to the automatic responses from human brains, humans treat media agents (i.e. "any technological artifact that demonstrates sufficient social cues to indicate the potential to be a source of social interaction," Gambino *et al.*, 2020, p. 73) like other humans or real objects in the physical world regardless of ontological differences. Later, the theory was extended into one of the most influential paradigms in human–computer interaction research, CASA, which proposes that media equations occur when a machine presents enough contextual cues to intrigue social schema (Nass and Moon, 2000).

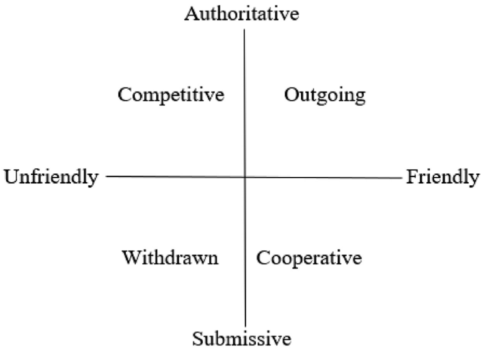
The CASA perspective has gained much support from empirical research across various contexts, such as human interactions with website interfaces (Nass and Lee, 2001), robots (Siegel *et al.*, 2009; Złotowski *et al.*, 2018), embodied conversational agents (Hoffmann *et al.*, 2009), smartphones (Carolus *et al.*, 2018) and autonomous vehicle voice agents (Lee *et al.*, 2019). However, some studies have called CASA's proposition into question. Previous research (e.g. Edwards *et al.*, 2016; Fischer, 2011; Kanda *et al.*, 2008; Melo *et al.*, 2016; Mou and Xu, 2017) disclosed that even though machines could be treated socially, certain cognitive, attitudinal and behavioral gaps exist between interpersonal and HMC. For example, Kanda *et al.* (2008) noticed that participants had slower responses in HMC than in interpersonal communication, and Mou and Xu (2017) showed that individuals displayed different personality traits when interacting with Microsoft's chatbot, Little Ice compared with human partners. The viewpoint of media inequality is reinforced by evidence from neuroscience indicating that HMC activates different brain activity patterns from human–human interactions (Gallagher *et al.*, 2002; Kircher *et al.*, 2009). Therefore, directly applying social rules of human communication to HMC is likely to elude crucial differences, and existing evidence of these inequalities emphasizes the necessity of additional investigation into media equations.

2.5.2 Preference for machine personalities. Media equation theory explains that humans perceive machines to possess human-like personalities based on the messages they convey (Reeves and Nass, 1996), which can be effectively and efficiently embedded with minimal cues (Ball and Breese, 2000; Nass *et al.*, 1995). As "the pattern of collective character, behavioral, temperamental, emotional and mental traits of an individual have consistency over time and situations" (Tapus *et al.*, 2008, p. 171), machine personalities are considered sets of perceived traits derived from observations in HMC. Reeves and Nass (1996) defined two of the Big Five dimensions of human personality, dominance/submissiveness and friendliness/unfriendliness, as two axes of machines' personality typology, as they argued that the other three dimensions, contentiousness, emotional stability and openness, are not equally applicable to computers (Okuno *et al.*, 2003). Their configuration was later refined by Dryer (1999) (see Figure 1). In media equation theory, dominance is described as assertiveness, self-confidence and competitiveness, suggesting a disposition to direct others, whereas friendliness is synonymous with warmth, sympathy and agreeableness, indicating a degree of cordialness. Reeves and Nass (1996) showed that users had a predilection for dominance and friendliness in computer personality presentation, and the most favoured combination of machine personalities was dominance (i.e. extraversion) plus friendliness (i.e. agreeableness).

This proposition from media equation theory gained support from subsequent studies in HMC areas such as human–robot and human–agent interactions. For example, Hwang *et al.* (2013) found that extroverted robots elicited more positive emotions from humans than introverted robots, and Hanna and Richards (2015) observed agreeable intelligent virtual agents positively influenced the development of shared mental models in teams. Yet other findings have challenged the proposition. In the same study by Hanna and Richards, extraversion was not associated with better team outcomes. The design research led by Braun and Alt (2020), on the other hand, revealed that the digital agent with a neutral personality presentation achieved higher likeability and user trust than the one with moderate levels of dominance and friendliness, which challenges the aforementioned proposition by Reeves and Nass (1996). Therefore, user preference for machine personality needs to be further examined across various contexts and settings.

In addition to user preference for machine personalities, previous research has explored the moderators in this process. The most frequently investigated cross-over interaction is the interaction between user personality and machine display (e.g. Isbister and Nass, 2000; Lee *et al.*, 2006; Nass and Lee, 2000; Park *et al.*, 2012; Tapus *et al.*, 2008), although studies on this topic have produced contradictory results. For example, Isbister and Nass (2000), Nass and Lee (2000), Park *et al.* (2012) and Tapus *et al.* (2008) found that similarity attraction affected user evaluations of virtual agents, whereas Lee *et al.* (2006) found that complementarity attraction appealed to human users in robotic designs.

Other factors include the roles assigned to the machine. Tay *et al.* (2014) found that an extroverted healthcare robot was evaluated more positively than an introverted one, but extroversion was not rated positively when the robot was designated to be a security robot. Additionally, the divergence between similarity and complementarity attraction could be partially explained by the machine attributes, task type, context and cultural background, according to Weiss and Evers (2011). For instance, Joosse (2013) showed that users' robot personality preferences could not be solely explained by either similarity or complementary



Source(s): Adapted from Dryer (1999)

Figure 1.
Typology of machine personality

attraction but varied by the assigned task or role. Therefore, it is necessary to explore personality-related issues with more diverse types of machines and tasks, considering that social attraction to PDAs is a determinant of use intentions (Choi and Choi, 2023) and that presented personality plays a central role in the social process. Correspondingly, this study focuses on the effects of PDA personalities on trust in different task types. As illustrated by Neff *et al.* (2010), manipulating linguistic cues is the most effective way to shape human perception of machine personality. Thus, the verbal expressions alone in a PDA's responses are sufficient to influence user impressions of the PDA's personality.

Apple envisioned Siri to be "friendly and humble—but also with an edge" (Kim, 2011). As a PDA with wide applications in real-world settings, Siri differs significantly from agents or robots tested by previous experimental designs, in which their presentations were stringently scripted with predefined personality cues. Instead, Siri's personality presentation fluctuates over time via each response in interaction threads, potentially suffering from inconsistency because of its technical limitations. Supposing the positive effects of dominance and friendliness hold salient in Siri's personality presentation through accumulative interactions, we propose:

> *H3.* The presentation of (a) dominance and (b) friendliness in Siri's responses will be positively associated with the levels of user trust.

Reeves and Nass (1996) also underscored personality identifiability (i.e. the recognizability of machine personalities by human users) as a part of user preference preceding different personality combinations, which resorts to both intensity and consistency of personality presentation in the two dimensions in Figure 1. A few studies have verified the positive correlation between personality intensity or consistency and user preference: for intensity, Dryer (1999) showed that strong personalities were liked better than subtle ones; and for consistency, Weiss and Evers (2011) found that consistency in machine personality presentation increased perceived trustworthiness. Overall, users have a penchant for media with clear personalities regardless of whether they liked the personalities or not (Reeves and Nass, 1996). Machine personalities need to be both salient and consistent to be identifiable. Therefore,

> *H4.* The more intense the personality presentation is in Siri's responses, the higher the level of user trust will be assigned to the agent.

> *H5.* The more consistent the personality presentation is in Siri's responses, the higher the level of user trust will be assigned to the agent.

3. Methods

3.1 Participants

The study sample consisted of 170 undergraduate students enrolled in communication courses at a Southwestern U.S. university, and seven incomplete responses were excluded from the analysis. Among the remaining 163 participants, 109 were female (64.1%) and 54 were male (33.1%). Their ages ranged from 18 to 27 (M = 19.63, SD = 1.54). Over half the participants self-identified as Caucasian or white (n = 114, 69.9%), followed by Asian (n = 21, 12.9%) and Hispanic/Latino (n = 8, 4.9%). The others reported themselves as Native American (n = 5, 3.1%), black or African American (n = 5, 3.1%), Pacific Islander (n = 1, 0.6%) or bi- or multi-racial (n = 8, 4.9%), with one self-reported as "other" (n = 1, 0.6%).

3.2 Procedures

The participants were recruited via a university-sponsored research pool, and they were offered extra credits for their choice of class. As the recruitment criteria, they were required to be iPhone users of at least 18 years old, with system software updated to iOS 12. The experiment was carried out in the college laboratory. The participants were randomly assigned to 15 functional ($n = 82$, 50.3%) or social ($n = 81$, 49.7%) task conditions to perform with Siri. After each task inquiry, participants were asked to enter Siri's responses verbatim into the online questionnaire presented via a laptop screen, which, in sum, generated 15 records per participant. After completing all 15 tasks, they evaluated Siri's performance, reported subjective trust toward Siri and finally provided their demographic and general usage information. The questionnaire had been pre-tested with a small group of graduate students who provided feedback on wording and understandability. Except for minor revisions to wording, no significant changes were made after the pretest as the participants were able to follow the instructions without any issues.

The experimental tasks synthesized some popular questions that users might ask Siri, recommended by several customer websites (e.g. Ben and Rahmanan, 2018). The set of functional tasks focused on common information-seeking attempts with Siri spanning various domains of life, ranging from everyday applications of PDAs (e.g. looking up weather information, geographic locations and time zone differences) to search for specific pieces of information (e.g. asking about the stock market or an uncommon word) and solving math problems. Social tasks involved exchanging jokes (e.g. "Will pigs fly?") and asking personal questions (e.g. "Siri, do you sleep?"). The data collection lasted from mid-October 2018 to late April 2019, and the study was approved by the Institutional Review Board of the university.

3.3 Measures

3.3.1 General usage. Participants were asked to report the length, frequency and proficiency of their Siri usage. Their usage duration was distributed as follows: 42.9% four years or more; 19.5% from two to three years; 19.5% from three to four years; 8.6% from one to two years; and 9.2% less than one year. In addition, 73 participants reported using Siri occasionally (44.8%), while 52 (31.9%) participants used it once a week or more. Usage proficiency was measured on a five-point scale (*1 = Novice; 2 = Below average; 3 = Average; 4 = Above average; 5 = Expert*), with 88 participants identifying their proficiency as average (54%), 46 as above average (28.2%) and one as expert (0.6%). Thus, most were experienced Siri users. Proficiency scores were positively correlated with usage length ($r = 0.16$, $p < 0.05$) and frequency ($r = 0.39$, $p < 0.001$), indicating the participants perceived themselves to be more skilled at using Siri if they had used it longer or more often.

3.3.2 User trust. The measures for user trust were adapted from a human-computer trust scale (Madsen and Gregor, 2000) with a total of 25 items (*1 = Strongly disagree to 5 = Strongly agree*) assessing five subconstructs of trust (i.e. perceived reliability, perceived technical competence, perceived understandability, faith and personal attachment). The scale has demonstrated good internal reliability in previous studies (e.g. Chavaillaz *et al.*, 2016). Madsen and Gregor (2000) further conducted a principal component analysis over their data and found two major components that corresponded with the bifactor model of trust (i.e. cognitive and affective trust). According to the results, perceived understandability (i.e. the perception that users can form a mental model of the system and predict its future performance) formulated cognitive trust, while faith (i.e. the positive beliefs in the system's future performance) and personal attachment (i.e. the positive affect and/or stronger preference for the system resulting from usage) contributed to affective trust (Madsen and Gregor, 2000). However, they found that perceived technical competence (i.e. the perception of performance accuracy and correctness) was related to both components, while perceived reliability (i.e. the perception of consistent functioning) was loaded on neither component, but this subdimension was still retained because of conceptual importance. In sum, the five subdimensions of trust perception are related but distinct, lending partial support to the bifactor structure of user trust while reflecting further complexity of user trust.

Cronbach's alpha scores indicated all subscales were reliable: perceived reliability (M = 3.41, SD = 0.73, α = 0.79), perceived technical competence (M = 3.58, SD = 0.72, α = 0.74), perceived understandability (M = 3.82, SD = 0.80, α = 0.83), faith (M = 2.53, SD = 0.89, α = 0.86) and personal attachment (M = 2.13, SD = 0.80, α = 0.86). To further examine the dimensionality of user trust, we performed a confirmatory factor analysis to inspect the five-factor structure proposed by Madsen and Gregor (2000), and the results showed that all initial goodness of fit indices met Hu and Bentler's (1999) criteria ($\chi^2/df \leq 2$; RMSEA ≤ 0.06; CFI ≥ 0.95; SRMR ≤ 0.08) except for CFI: $\chi^2/df = 1.55$, $p < 0.001$; RMSEA = 0.06; CFI = 0.92; SRMR = 0.06. The model fit was improved after allowing for error covariance between the latent constructs and error covariance of items within each subdimension: $\chi^2/df = 1.22$, $p < 0.001$; RMSEA = 0.04; CFI = 0.95; SRMR = 0.05, so the five-factor structure was retained. The overall trust level was indicated by the aggregated score of all 25 items (M = 77.01, SD = 15.36), and the average score of each subdimension was further analyzed. The data were normally distributed for both the overall trust score and averaged scores of subscales, based on Osborne's (2013) criteria that skewness and kurtosis with absolute values smaller than one should not raise concern.

3.3.3 Response categories. Siri's responses for functional tasks were categorized according to how Siri framed the information attained. For instance, Siri might respond to "What's the weather like today?" with "It should be nice today . . . up to [temperature]," with real-time temperature varied, so all responses framed exactly this way were classified into one category. Responses for social tasks were coded based on the exact words used (e.g. Siri might respond to "Siri, do you sleep?" with "I never rest. But thanks for asking"), which were predetermined by Siri's algorithms. The whole unitizing procedure was completed by the first author. The functional tasks generated a total of 141 distinct response categories, while the social tasks generated 104 categories across 15 task inquiries for each task type, summing up to 245 response categories in total.

3.3.4 Error occurrence. Errors in any responses from Siri were identified and categorized by the first author as logical, semantic or syntactic, using the definitions by McCall and Kölling (2014). Seventy of 82 participants came across at least one error when performing functional tasks, with a total of 109 mistakes in Siri's responses (M = 1.33, SD = 1.11) out of 1,196 valid responses (9.1%), while only one individual came across errors repeatedly in social tasks (M = 0.09, SD = 1.89), which added up to 8 mistakes out of 1,214 (0.7%) responses. The number of incorrect responses had no significant correlations with the general usage patterns of usage length, frequency or fluency. Since errors in social tasks only occurred with one participant, H1 and the RQ were examined only for functional tasks. For functional tasks, there were 70 logical errors, 16 semantic errors and 23 syntactic errors. No significant correlation was found among the error types, indicating that the different types occurred independently.

3.3.5 Personality presentation. Personality presentation was assessed on two dimensions: dominance and friendliness. Dominance was operationalized as certainty, directness and authoritativeness of the linguistic style, and friendliness was operationalized as social warmth, politeness and tentativeness. Personality presentation in Siri's responses recorded by the participants was rated on two 5-point scales (−2 = submissive, 2 = dominating; −2 = unfriendly, 2 = friendly) by the two authors. Neutrality in responses was rated as zero. First, the researchers independently rated all 245 response categories after an initial discussion about conceptualization and operationalization, and 70.0% of the initial attempts reached an agreement, which was acceptable based on the conventional cutoff value of reliability, 0.70. Next, the two researchers met again to clarify the coding standards and jointly worked on the final codebook. Each author again independently rated 50% of the response categories based on the final codebook. The second round of ratings were combined and analyzed as final ratings for personality presentation.

Dominance levels in Siri's responses were calculated as the sum of the first scale scores (*−2 = Submissive, 2 = Dominating*), and its friendliness levels were calculated by summing the second scale scores (*−2 = Unfriendly, 2 = Friendly*). On average, Siri showed lower levels of dominance in social tasks ($M = 12.63$, $SD = 3.36$) than in functional tasks ($M = 14.31$, $SD = 4.33$) and higher levels of friendliness in social tasks ($M = 5.71$, $SD = 3.20$) than in functional tasks ($M = 5.24$, $SD = 1.82$). Correlation analysis showed that dominance was negatively correlated with friendliness in functional tasks ($r = -0.73$, $p < 0.001$) but uncorrelated in social tasks.

Personality identifiability contained two constructs: intensity and consistency. Intensity was calculated as products of the absolute values of all dominance and friendliness scores for every individual ($M = 204.10$, $SD = 44.29$ for social tasks; $M = 90.02$, $SD = 27.08$ for functional tasks), which reflected the strength of perceived personality in Siri's responses. Intensity of dominance was positively related to that of friendliness ($r = 0.32, p < 0.01$) for social tasks, but the correlation was negative for functional tasks ($r = -0.76$, $p < 0.001$), preliminarily revealing differences in Siri's personality displays in the two task types.

Personality consistency was calculated as products of the standard deviations of every participant's scores on two scales (i.e. dominance and friendliness) ($M = 1.01$, $SD = 0.27$ for social tasks; $M = 0.44$, $SD = 0.16$ for functional tasks), indicating the extent to which the participants were exposed to the presentation of similar personality characteristics across the 15 task inquiries. A lower score would mean high consistency due to smaller variations. The consistency of dominant personalities was positively correlated with the consistency of friendliness scores in functional tasks ($r = 0.36$, $p < 0.01$), but this correlation was non-significant in social tasks. Such correlations reveal that Siri's responses had more consistent personality patterns in functional tasks compared to social tasks.

4. Results

4.1 Results for H1 test and RQ

H1 proposed that error occurrence would be negatively correlated with user trust. With social tasks excluded due to the lack of statistical power, the results of correlation analysis which was performed using IBM SPSS version 28.0 showed the number of errors was negatively related to trust in functional tasks (partial $r = -0.33$, $p < 0.01$), after controlling for proficiency, length and frequency of usage; thus, H1 was supported. Further analysis revealed that such connections between error occurrence and trust were significant for the dimension of perceived reliability (partial $r = -0.37$, $p < 0.01$), perceived technical competence (partial $r = -0.36$, $p < 0.01$) and faith (partial $r = -0.31$, $p < 0.05$), after controlling for general usage variables, but not with perceived understandability (partial $r = -0.15, p = 0.24$) or personal attachment (partial $r = -0.10, p = 0.45$).

To explore which error types had stronger influences over user trust in the PDA, partial correlation analyses were performed using SPSS for the three types of errors (i.e. logical, semantic and syntactic) in relation to functional tasks and user trust, controlling for the effects of proficiency, length and frequency of usage. Social tasks were not examined because logical errors were the only type that appeared in them. The number of logical errors was significantly correlated with overall trust, partial $r = -0.27$, $p = 0.05$. Further analysis showed it was negatively associated with perceived competence (partial $r = -0.29, p < 0.05$) and perceived reliability (partial $r = -0.36$, $p < 0.05$) but not with perceived understandability (partial $r = -0.06$, $p = 0.67$), faith (partial $r = -0.27$, $p = 0.06$) or personal attachment (partial $r = -0.09$, $p = 0.52$). The numbers of semantic errors and syntactic errors, however, were not significantly correlated with either overall trust or any of the five subdimensions of user trust. These results indicated that the more frequently Siri produced logical errors in responses, the less reliable and competent users evaluated it to be.

4.2 Results for H2 test

H2 predicted users would report higher levels of trust in Siri after completing functional rather than social tasks with the PDA's assistance, which was supported by the result from an analysis of covariance (ANCOVA) using IBM SPSS. Functional tasks (M = 80.48, SD = 13.59) overall elicited higher levels of trust than social tasks (M = 73.36, SD = 16.33), $F(1, 4) = 8.06, p < 0.01$, partial eta square = 0.05. The difference was statistically significant after controlling for the effects of usage proficiency, $F(1, 4) = 9.99, p < 0.01$, partial eta^2 = 0.06.

A further examination of each subconstruct of trust via univariate tests revealed that functional tasks yielded higher levels of perceived technical competence, $F(1, 153) = 10.95$, $p < 0.01$, partial $\eta^2 = 0.07$ and faith, $F(1, 153) = 4.79, p < 0.05$, partial $\eta^2 = 0.03$, in Siri than social tasks, but there was no significant difference between the two task types when it came to personal attachment, $F(1, 153) = 0.18$, $p = 0.67$, partial $\eta^2 = 0.01$. The difference in perceived reliability approached statistical significance, $F(1, 153) = 3.70$, $p = 0.06$, partial $\eta^2 = 0.02$ and significance in perceived understandability could not be claimed as Levene's test of equality of error variance was significant for the variable. See Table 1 for a summary of the comparison of means in each dimension of trust by task type.

4.3 Results for H3 test

H3 anticipated that higher levels of overall perceived dominance and friendliness would result in higher user trust. Hierarchical regressions were conducted using SPSS to test H3 with the dominance and friendliness scores as the predictors, overall trust and its five subdimensions as the criteria, and proficiency, length and frequency of usage as the controls. Analysis using Q-Q plots showed that the error terms were normally distributed, corresponding with the values attained through Durbin Watson tests between 1 and 3 (Field, 2009). Perceived dominance and friendliness levels had no significant effects on overall trust in either social tasks, R^2 change = 0.003, $p = 0.87$ or functional tasks, R^2 change = 0.02, $p = 0.66$; therefore, H2 was unsupported. See the first two columns in Table 2 for a summary of the regression results.

A further assessment of the five subdimensions of trust revealed that dominance ratings did not significantly predict any differences in user trust under either task type, but friendliness was positively associated with perceived reliability in social tasks, $\beta = 0.28$, $p < 0.05$. This means if Siri was perceived as more friendly in response to social inquiries,

	Task type	M	SD	n
Competence	Social	16.83_a	3.74	77
	Functional	18.74_b	3.14	81
Reliability	Social	16.36	3.52	77
	Functional	17.54	3.66	81
Understandability	Social	18.04	4.66	77
	Functional	20.05	2.99	81
Faith	Social	11.82_a	4.58	77
	Functional	13.44_b	4.29	81
Attachment	Social	10.31	4.36	77
	Functional	10.70	3.58	81
Overall trust	Social	73.36_a	16.33	77
	Functional	80.48_b	13.59	81

Note(s): Means with different subscripts within each category of trust indicate statistically meaningful differences

Source(s): Authors' own creation/work

Table 1.
Comparison of means by task types on five dimensions of trust

Table 2. Hierarchical multiple regression predicting trust by dominance and friendliness

		Overall trust				Perceived reliability			
		Social task		Functional task		Social task		Functional task	
Model		β	t	β	t	β	t	β	t
1	Proficiency	0.48^{***}	4.41	0.06	0.39	0.31^{*}	2.56	−0.06	−0.44
	Usage Length	−0.08	−0.77	−0.09	−0.65	0.03	0.23	−0.16	−1.18
	Frequency	0.10	0.98	−0.01	−0.04	−0.00	−0.01	0.04	0.26
	Adjusted R^2	0.23^{***}		−0.05		0.06		−0.02	
2	Proficiency	0.46^{***}	4.10	0.06	0.37	0.22	1.88	−0.06	−0.41
	Usage Length	−0.08	−0.77	−0.09	−0.66	0.02	0.17	−0.18	−1.38
	Frequency	0.11	0.97	0.00	0.01	0.02	0.19	0.05	0.35
	Dominance	0.03	0.25	−0.12	−0.53	0.14	1.35	−0.33	−1.69
	Friendliness	0.05	0.46	−0.19	−0.89	0.28^{*}	2.59	-0.46^{*}	−2.41
	R^2 change	0.003		0.02		0.09^{*}		0.10	
	Adjusted R^2	0.22^{***}		−0.07		0.14^{**}		0.05	

Note(s): $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$
Source(s): Authors' own creation/work

participants evaluated Siri's performance as more reliable. Conversely, friendliness rating was negatively associated with perceived reliability in functional tasks, $\beta = -0.46, p < 0.05$. This means the more Siri was presented to be friendly in response to functional inquiries, the less reliable participants considered Siri to be (see the third and fourth columns of Table 2). Due to the limitation of space, we only present the dimension (i.e. perceived reliability) that demonstrates statistically significant relationships.

4.4 Results for H4 and H5 tests

H4 and H5 posited the effects of personality intensity and consistency. Q-Q plots and Durbin–Watson tests indicated the error terms were normally distributed. Contrary to our expectations, hierarchical regression revealed that personality intensity and consistency together did not yield statistically significant effects on overall user trust in either functional tasks, R^2 change = 0.01, $p = 0.81$ or social tasks, R^2 change = 0.002, $p = 0.90$, after controlling for the effects of general usage (i.e. proficiency, length and frequency of usage). Thus, H3 and H4 were not supported. It was notable that as a control variable, the level of proficiency had a positive association with overall trust in social tasks, $\beta = 0.48, p < 0.001$, but not functional tasks, $\beta = 0.06, p = 0.71$. A further assessment of the five subdimensions of trust revealed that the personality intensity score was a significant predictor of perceived reliability in functional tasks, $\beta = -0.34, p < 0.05$. The more intense the personality of Siri was presented to be, the less reliable participants evaluated its performance as in functional tasks. This result was the opposite of H4's prediction.

5. Discussion

This study examined user interactions with a commonly used PDA, Siri, with a focus on task type, personality presentation and performance errors. As weak (Lu *et al.*, 2018) and narrow AI (Kaplan and Haenlein, 2019) systems, PDAs can only operate predefined functions within a specific range of tasks. Apparently, the present stage of PDA development has not reached some customers' expectations for AI performance (e.g. Goertzel, 2010), so elevated sophistication in technological capacity remains a desideratum for promoting the user experience. But besides the issue of technological advancement, the current study further revealed some nuances that PDA designers could differentiate to facilitate the development of user trust. Pragmatically, the new error categorization (i.e. logical, semantic and syntactic) applied in this study can help designers better identify different types of common system errors, understand their impacts on user trust and seek better remedies in pursuit of effective trust repairs, given their omnipresence in human interactions with PDA. As this study revealed, the differentiation between functional and social tasks pre-staged PDAs' functionality. Also, user evaluations of a PDA's personality presentation were affected by the PDA's roles, which advocate for more task-dependent designs and a human-centered approach. We further discuss the implications of our findings as follows.

The findings concerning H1 aligned with previous studies emphasizing the importance of task performance in user evaluations and confirmed that PDA's errors impaired user trust. This study extended previous research by investigating the impacts of error occurrence on different subdimensions of human-machine trust (Madsen and Gregor, 2000). The negativity of errors was identified in three out of the five dimensions of user trust, with the two exceptions of perceived understandability and personal attachment. This suggested the possibility that because of the wide accessibility of Siri and its habitual usage, users' understanding of and attachment to it might be resilient enough to tolerate its mistakes. The occurrence of errors did not significantly hinder the perceived understandability of the PDA, which indicates Siri's erroneous responses did not add to the incomprehensibility of the system. However, frequent errors lowered participants' trust in Siri regarding trust dimensions of technical competence, reliability and faith, again corroborating the powerful effects of performance errors on various components of human-machine trust.

Analyses for RQ adopted three categories of response errors (i.e. logical, semantic and syntactic) used in computer science and examined the effects of the errors in detail. The results revealed logical errors to be the most detrimental to user trust. It was unexpected that semantic errors were not significantly related to trust because, intuitively, semantic errors should be more negative than logical errors as they achieve less correctness and accuracy.

The other finding, namely, that syntactic errors did not exert a significant impact, was also interesting as Siri did not even successfully decode the input in such errors. It seemed participants would rather see Siri not catch their question at all than catch part or all of it but still generate a wrong answer. One possible explanation is that with semantic and syntactic errors, the participants might have surmised that they did not articulate the question well or some background noise had disturbed Siri's performance and therefore took the blame off from Siri. These results also illuminate the complexity in users' psychological processes when they come across mechanical mistakes, possibly affected by causal attributions (i.e. locus, stability and controllability; Weiner, 1985): participants might have been more inclined to perceive semantic and syntactic errors as abnormalities caused by external forces (e.g. noise, users' fault) in comparison to logical errors. Notably, logical errors mainly impeded trust by decreasing Siri's perceived reliability and technical competence. This preliminary finding demonstrates how the three error types might stimulate differential negative user perceptions and detailed mechanisms through which human-machine trust might be impaired.

This study also contributes to the conceptualization of system errors in HMC. Previous HMC studies that introduced performance errors as experimental treatments rarely designed them following theoretical taxonomies, and their choices of errors were more practically rationalized than theoretically justified. Such designs had limitations for generalizable conclusions. Therefore, further exploration of this topic could benefit both research designs and data analysis by considering meaningful distinctions between error types. Existing categorizations of error occurrence largely rely on identifications of the causes when violation outcomes do not differ significantly, but lay users do not have the ability to recognize the locations of system failures, which is why this study adopted an output-oriented categorization (McCall and Kölling, 2014). With its attempt at developing a classification suitable for PDA research, this study directs attention to some central error mechanisms sensible for users, potentially bridging user perceptions and technical mechanisms. This theoretical development can assist future research across different HMC contexts.

The task type effects on trust examined in H2 confirmed previous empirical findings that automated agents with low levels of humanness received more trust in functional or analytical tasks (Lee *et al.*, 2021; Smith *et al.*, 2016), which was probably also influenced by functionality in the designated role of Siri as a PDA. In other words, the participants were more familiar with engaging in functional tasks inquiries assisted by Siri (e.g. asking for information about the weather, stock market or nearby coffee shops) than with socializing with Siri (e.g. exchanging funny jokes). While it is still possible for human users to enjoy such social interactions with PDAs and pleasant emotions experienced from those interactions can increase their trust (Lee and Sun, 2022), compared to functional task inquiries, socializing with Siri seemed to have less of an effect on user trust.

The findings over the effects of PDA personality presentation further questioned the previous proposition about users' preference over machine personality by Reeves and Nass (1996). Although H3 received some support from the data, the other hypotheses did not, with H4 even partially contradicting the proposition, indicating that most of the theoretical deductions concerning personality preference based on the media equation theory fell short of their explanatory power in the current context. This finding revealed the interaction between the task type and personality preference. Whereas perceived friendliness increased perceived reliability in social tasks, it conversely reduced perceived reliability in functional tasks. This corresponded with the negative association between personality intensity and perceived reliability in functional tasks, indicating that users might have preferred a less intense and impersonal presentation of machine personality because they regarded Siri simply as a helping tool and did not desire or enjoy social conversations with it under the given circumstances. They might have had distinct preferences for machine personalities in different task types. For instance, they favored cordiality for social interactions but a professional, business-like attitude for functional tasks. Thus, the same dimension of trust can be perceived differently based on the context and nature of the HMC. The division between social and functional tasks could serve as a moderator between the preference for perceived machine personality and the evaluation of performance reliability (Gaudiello *et al.*, 2016).

These findings, distinct from media equation theory and CASA's prediction, can be understood as another evidence demonstrating the unique nature of HMC compared to interpersonal communication. The empirical evidence against CASA abounds. For example, Gambino and Liu (2022) argued that human interlocutors have fewer concerns about impression management or relational maintenance when interacting with a machine agent, which makes the interaction less intimate, more direct and sometimes even more aggressive. Researchers found that 10% of users' commentary toward machine agents involved verbal aggressions (de Angeli and Brahnam, 2008), and almost 40% of students who interacted with a female conversational agent (the case of Siri, too) showed an aggressive attitude by using hypersexualized and dehumanizing expressions (Veletsianos and Miller, 2008). Mou and Xu's (2017) research on online chatting found participants were less open, agreeable, extroverted and conscientious when communicating with a chatbot compared to another human. Following these previous studies, the current research adds to the extant knowledge about the differences between human-human and human-machine communication, which can help better establish the boundary conditions of the media equation and facilitate future theorization of HMC.

Distinguished from prior studies, the current research did not treat personality as a fixed factor and the patterns of Siri's personality presentation to which participants were exposed during the 15 consecutive task responses were far more varied than the personality designs implemented previously (e.g. Dryer, 1999). Our inability to capture the net effects of personality type and identifiability using media equation theory unveils a dilemma in PDA development: though variation in personality presentation might make Siri sound more expressive and intelligent, users could experience difficulties in recognizing its personality due to the variation. Media equation theory proposed that people favor machines with personality changes over ones with a constant presentation if the changes occur in certain directions (e.g. from submissive to dominating; Reeves and Nass, 1996). Due to the weak learning capabilities of most AIs and the algorithmic differentiation between distinct task types, steady transformation in personality across task types is not easy to attain; instead, users are often exposed to diversified personality presentations in ongoing strings of question-and-answer tasks, unable to sense clear trends of personality growth. Therefore, additional exploration over more diversified personality patterns than just dominance and friendliness will certainly benefit technology designs.

Notably, the analysis of these human–machine trust subdimensions also disclosed some relationships worthy of further exploration. The results of this study indicated that, among the five subconstructs of user trust, perceived reliability was potentially the most indicative dimension of overall trust, reflecting that user judgments of PDAs are more based on functional consistency across interaction episodes. Therefore, PDA designs should pay more attention to maintaining stability of performance quality above and beyond the enhancement of PDA capabilities. The affective component of trust appears to be less influenced by Siri's performance. One possible reason for no significant between-group differences in personal attachment in every test could be that participants' general attachment to Siri was influenced more by their long-term and frequent usage outside the lab experiment.

6. Limitations and future directions

Because the responses participants received from Siri were not predefined, some categories had too little statistical power to be tested for our hypotheses. The coding scheme for personality presentation and identifiability could also have been further elaborated by recruiting more coders. Meanwhile, a college student sample prevented us from making any claims generalizable to other population groups despite the random assignment of an experimental design, so replications with diverse samples could verify our findings and enhance overall interpretation.

Several directions for future investigation are proposed. First, the research revealed noticeable distinctions in user trust between the two task types, the mechanisms of which deserve additional exploration within different HMC contexts. Furthermore, future research can investigate more complex patterns in personality display, extending the line of HMC personality research. They might also explore the underlying psychological mechanisms of user responses and the effectiveness of trust repair attempts (e.g. blaming, apology) for different error types. Finally, the error categorization introduced in this study (i.e. logical, semantic and syntactic; McCall and Kölling, 2014) presents possibilities for cross-contextual application.

7. Conclusion

This study shows how different task types influence user evaluations of Siri. As a virtual agent with low humanness levels, Siri elicited more positive evaluations by assisting users with functional tasks rather than social tasks, which accords with previous findings in the field (Goetz *et al.*, 2003; Lee *et al.*, 2021; Smith *et al.*, 2016). Reeves and Nass (1996) suggested that humans favor agents with clearly identifiable personalities that are dominant and friendly, yet these predictions did not gain strong support from the data. In this study, dominance and personality consistency ratings were not significantly associated with human-machine trust, but the effect of friendliness differed by task type. Moreover, the intensity of personality presentation was only associated with one subdimension of trust, perceived reliability. In addition to providing this counterevidence to the media equation's prediction, the findings revealed that Siri mostly failed to establish an influential agent profile in its response threads, which calls for more scholarly investigations in the future.

Finally, this study proposes a new error categorization centered on human-machine users by examining the relationships between intended and actual output. The findings illustrated how user evaluation of a PDA can be influenced by multiple types of performance errors, with logical errors as the most detrimental type. Overall, the research contributes to the ongoing theoretical debate over media equation theory, and the findings can serve as a basis for practical guidelines for human-centered AI agent designs.

References

Apple (2018), "HomePod arrives February 9", available at: https://www.apple.com/newsroom/2018/01/homepod-arrives-february-9-available-to-order-this-friday/ (accessed 23 January 2021).

Ball, G. and Breese, J. (2000), "Emotion and personality in a conversational agent", *Embodied Conversational Agents*, pp. 189-219.

Ben, A. and Rahmanan, Y. (2018), "21 really funny things to ask Siri right now", available at: https://www.timeout.com/usa/things-to-do/funny-things-to-ask-siri (accessed 21 January 2022).

Berdasco, López, G., Diaz, I., Quesada, L. and Guerrero, L.A. (2019), "User experience comparison of intelligent personal assistants: Alexa, google assistant, Siri and cortana", *Proceedings*, Vol. 31 No. 1, p. 51, doi: 10.3390/proceedings2019031051.

Braun, M. and Alt, F. (2020), "Identifying personality dimensions for characters of digital agents", in El Bolock, A., Abdelrahman, Y. and Abdennadher, S. (Eds), *Character Computing*, Springer, Cham, pp. 123-138, doi: 10.1007/978-3-030-15954-2_8.

Brooks, D. (2017), "A human-centric approach to autonomous robot failures", unpublished doctoral dissertation, ProQuest Dissertations and Theses Database. (UMI No. 10643702).

Carlson, J. and Murphy, R.R. (2005), "How UGVs physically fail in the field", *IEEE Transactions on Robotics*, Vol. 21 No. 3, pp. 423-437, doi: 10.1109/tro.2004.838027.

Carolus, A., Schmidt, C., Schneider, F., Mayr, J. and Muench, R. (2018), "Are people polite to smartphones?", in Kurosu, M. (Ed), *Human-Computer Interaction. Interaction in Context. HCI 2018, Lecture Notes in Computer Science*, Springer, Cham, 10902, doi: 10.1007/978-3-319-91244-8_39.

Chavaillaz, A., Wastell, D. and Sauer, J. (2016), "System reliability, performance and trust in adaptable automation", *Applied Ergonomics*, Vol. 52, pp. 333-342, doi: 10.1016/j.apergo.2015.07.012.

Choi, T.R. and Choi, J.H. (2023), "You are not alone: a serial mediation of social attraction, privacy concerns, and satisfaction in voice AI use", *Behavioral Sciences*, Vol. 13 No. 5, p. 431, doi: 10.3390/bs13050431.

Corritore, C.L., Kracher, B. and Wiedenbeck, S. (2003), "On-line trust: concepts, evolving themes, a model", *International Journal of Human-Computer Studies*, Vol. 58 No. 6, pp. 737-758, doi: 10.1016/s1071-5819(03)00041-7.

Davenport, R.B. and Bustamante, E.A. (2010), "Effects of false-alarm vs Miss-prone automation and likelihood alarm technology on trust, reliance, and compliance in a miss-prone task", *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 54 No. 19, pp. 1513-1517, doi: 10.1037/e578802012-034.

de Angeli, A. and Brahnam, S. (2008), "I hate you! Disinhibition with virtual partners", *Interacting with Computers*, Vol. 20 No. 3, pp. 302-310, doi: 10.1016/j.intcom.2008.02.004.

de Vries, P., Midden, C. and Bouwhuis, D. (2003), "The effects of errors on system trust, self-confidence, and the allocation of control in route planning", *International Journal of Human-Computer Studies*, Vol. 58 No. 6, pp. 719-735, doi: 10.1016/s1071-5819(03)00039-9.

Dryer, D.C. (1999), "Getting personal with computers: how to design personalities for agents", *Applied Artificial Intelligence*, Vol. 13 No. 3, pp. 273-295, doi: 10.1080/088395199117423.

Edwards, A., Edwards, C., Spence, P.R., Harris, C. and Gambino, A. (2016), "Robots in the classroom: differences in students' perceptions of credibility and learning between 'teacher as robot' and 'robot as teacher'", *Computers in Human Behavior*, Vol. 65, pp. 627-634, doi: 10.1016/j.chb.2016.06.005.

Enge, E. (2019), "Rating the smarts of the digital personal assistants in 2019 Perficient", available at: https://www.perficientdigital.com/insights/our-research/digital-personal-assistants-study (assessed 24 October 2021).

Field, A. (2009), *Discovering Statistics Using IBM SPSS Statistics*, 2nd ed., Sage, London.

Fischer, K., Foth, K., Rohlfing, K.J. and Wrede, B. (2011), "Mindful tutors: linguistic choice and action demonstration in speech to infants and a simulated robot", *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, Vol. 12 No. 1, pp. 134-161, doi: 10.1075/is.12.1.06fis.

Gallagher, H.L., Jack, A.I., Roepstorff, A. and Frith, C.D. (2002), "Imaging the intentional stance in a competitive game", *Neuroimage*, Vol. 16 No. 3, pp. 814-821, doi: 10.1006/nimg.2002.1117.

Gambino, A., Fox, J. and Ratan, R.A. (2020), "Building a stronger CASA: extending the computers are social actors paradigm", *Human-Machine Communication*, Vol. 1, pp. 71-86, doi: 10.30658/hmc.1.5.

Gambino, A. and Liu, B. (2022 In this issue), "Considering the context to build theory in HCI, HRI and HMC: Explicating differences in processes of communication and socialization with social technologies", *Human-Machine Communication*, Vol. 4, pp. 111-130, doi: 10.30658/hmc.4.6.

Gaudiello, I., Zibetti, E., Lefort, S., Chetouani, M. and Ivaldi, S. (2016), "Trust as indicator of robot functional and social acceptance. An experimental study on user conformation to iCub answers", *Computers in Human Behavior*, Vol. 61, pp. 633-655, doi: 10.1016/j.chb.2016.03.057.

Goertzel, B. (2010), "Siri, the new iPhone 'AI personal assistant': some useful niche applications, not so much AI [Web log post]", available at: http://multiverseaccordingtoben.blogspot.com/2010/02/siri-new-iphone-personal-assistant-some.html (accessed 6 February 2021).

Goetz, J., Kiesler, S. and Powers, A. (2003), "Matching robot appearance and behavior to tasks to improve human-robot cooperation", *the 12th IEEE International Workshop on Robot and Human Interactive Communication, Proceedings of the. roman, Proceedings 2003. RO-Man. 2003. The 12th IEEE International Workshop on Robot and Human Interactive Communication*. Millbrae, California, USA: IEEE, pp. 55-60, doi: 10.1109/ROMAN.2003.1251796.

Guo, J., Tao, D. and Yang, C. (2020), "The effects of continuous conversation and task complexity on usability of an AI-based conversational agent in smart home environments", in Long, S. and Dhillon, B.S. (Eds), *Man–machine–environment System Engineering*, Springer Singapore (Lecture Notes in Electrical Engineering), Singapore, pp. 695-703, doi: 10.1007/978-981-13-8779-1_79.

Guznov, S., Lyons, J., Nelson, A. and Woolley, M. (2016), "The effects of automation error types on operators' trust and Reliance", in Lackey, S. and Shumaker, R. (Ed.), *Lecture Notes in Computer Science*, Springer International, Cham, pp. 116-124, doi: 10.1007/978-3-319-39907-2_11.

Hancock, P.A., Billings, D.R., Schaefer, K.E., Chen, J.Y.C., de Visser, E.J. and Parasuraman, R. (2011), "A meta-analysis of factors affecting trust in human-robot interaction, human factors", *The Journal of the Human Factors and Ergonomics Society*, Vol. 53 No. 5, pp. 517-527, doi: 10.1177/0018720811417254.

Hancock, P.A., Kessler, T.T., Kaplan, A.D., Brill, J.C. and Szalma, J.L. (2021), "Evolving trust in robots: specification through sequential and comparative meta-analyses", *Human Factors*, Vol. 63 No. 7, pp. 1196-1229, doi: 10.1177/0018720820922080.

Hanna, N. and Richards, D. (2015), "The impact of virtual agent personality on a shared mental model with humans during collaboration", in *AAMAS*, pp. 1777-1778.

Hoff, K.A. and Bashir, M. (2015), “Trust in automation: integrating empirical evidence on factors that influence trust”, *The Journal of the Human Factors and Ergonomics Society*, Vol. 57 No. 3, pp. 407-434, doi: 10.1177/0018720814547570.

Hoffmann, L., Krämer, N.C., Lam-Chi, A. and Kopp, S. (2009), “Media equation revisited: do users show polite reactions towards an embodied agent?”, in *Lecture Notes in Computer Science International Workshop on Intelligent Virtual Agents*, Springer, Berlin, Heidelberg, pp. 159-165.

Honig, S. and Oron-Gilad, T. (2018), “Understanding and resolving failures in human-robot interaction: literature review and model development”, *Frontiers in Psychology*, Vol. 9, p. 861, doi: 10.3389/fpsyg.2018.00861.

Hu, L. and Bentler, P.M. (1999), “Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives”, *Structural Equation Modeling: A Multidisciplinary Journal*, Vol. 6 No. 1, pp. 1-55, doi: 10.1080/10705519909540118.

Hwang, J., Park, T. and Hwang, W. (2013), “The effects of overall robot shape on the emotions invoked in users and the perceived personalities of robot”, *Applied Ergonomics*, Vol. 44 No. 3, pp. 459-471, doi: 10.1016/j.apergo.2012.10.010.

Isbister, K. and Nass, C. (2000), “Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics”, *International Journal of Human-Computer Studies*, Vol. 53 No. 2, pp. 251-267, doi: 10.1006/ijhc.2000.0368.

Johnson, J.D., Sanchez, J., Fisk, A.D. and Rogers, W.A. (2004), “Type of automation failure: the effects on trust and Reliance in automation”, *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 48 No. 18, pp. 2163-2167, doi: 10.1037/e577212012-007.

Joosse, M., Lohse, M., Perez, J.G. and Evers, V. (2013), “What you do is who you are: the role of task context in perceived social robot personality”, *IEEE International Conference on Robotics and Automation 2013. IEEE*, pp. 2134-2139, doi: 10.1109/ICRA.2013.6630863.

Kanda, T., Miyashita, T., Osada, T., Haikawa, Y. and Ishiguro, H. (2008), “Analysis of humanoid appearances in human-robot interaction”, *IEEE Transactions on Robotics*, Vol. 24 No. 3, pp. 725-735, doi: 10.1109/tro.2008.921566.

Kaplan, A. and Haenlein, M. (2019), “Siri, Siri, in my hand: who’s the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence”, *Business Horizons*, Vol. 62 No. 1, pp. 15-25, doi: 10.1016/j.bushor.2018.08.004.

Kim, A. (2011), “How Apple approached developing Siri’s personality”, *MacRumors*, available at: https://www.macrumors.com/2011/10/15/how-apple-approached-developing-siris-personality/ (accessed 15 October 2021).

Kircher, T., Blümel, I., Marjoram, D., Lataster, T., Krabbendam, L., Weber, J., van Os, J. and Krach, S. (2009), “Online mentalizing investigated with functional MRI”, *Neuroscience Letters*, Vol. 454 No. 3, pp. 176-181, doi: 10.1016/j.neulet.2009.03.026.

Kyung, N. and Kwon, H.E. (2022), “Rationally trust, but emotionally? The roles of cognitive and affective trust in laypeople’s acceptance of AI for preventive care operations”, *Production and Operations Management*, pp. 1-20, doi: 10.1111/poms.13785.

Lee, S. and Sun, J. (2022), “Testing a theoretical model of trust in human-machine communication: emotional experiences and social presence”, *Behaviour and Information Technology*, Vol. 42 No. 16, pp. 2754-2767, doi: 10.1080/0144929X.2022.2145998.

Lee, K.M., Peng, W., Jin, S. and Yan, C. (2006), “Can robots manifest personality?: an empirical test of personality recognition, social responses, and social presence in human-robot interaction”, *Journal of Communication*, Vol. 56 No. 4, pp. 754-772, doi: 10.1111/j.1460-2466.2006.00318.x.

Lee, S., Ratan, R. and Park, T. (2019), “The voice makes the car: enhancing autonomous vehicle perceptions and adoption intention through voice agent gender and style”, *Multimodal Technologies and Interaction*, Vol. 3 No. 1, p. 20, doi: 10.3390/mti3010020.

Lee, S.K., Kavya, P. and Lasser, S.C. (2021), "Social interactions and relationships with an intelligent virtual agent", *International Journal of Human-Computer Studies*, Vol. 150, 102608, doi: 10.1016/j.ijhcs.2021.102608.

Lewis, M., Sycara, K. and Walker, P. (2018), "The role of trust in human-robot interaction", in Abbass, H.A., Scholz, J. and Reid, D.J. (Eds), *Foundations of Trusted Autonomy*, Springer, pp. 135-159.

Lu, H., Li, Y., Chen, M., Kim, H. and Serikawa, S. (2018), "Brain intelligence: go beyond artificial intelligence", *Mobile Networks and Applications*, Vol. 23 No. 2, pp. 368-375, doi: 10.1007/s11036-017-0932-8.

Madsen, M. and Gregor, S. (2000), "Measuring human-computer trust", *Paper Presented at the 11th Australasian Conference on Information Systems*, 30 November December 2. Queensland, available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.93.3874&rep=rep1&type=pdf (accessed 20 February 2021).

McCall, D. and Kolling, M. (2014), "Meaningful categorisation of novice programmer errors", *Proceedings. Madrid, Spain: IEEE, 2014 IEEE Frontiers in Education Conference (FIE)*, pp. 1-8, doi: 10.1109/FIE.2014.7044420.

Melo, C.D., Marsella, S. and Gratch, J. (2016), "People do not feel guilty about exploiting machines", *ACM Transactions on Computer-Human Interaction*, Vol. 23 No. 2, pp. 1-17, doi: 10.1145/2890495.

Microsoft (2019), "Voice report: consumer adoption of voice technology and digital assistants", Microsoft, available at: https://about.ads.microsoft.com/en-us/insights/2019-voice-report (accessed 15 April 2019).

Morgan, B. (2019), "The customer of the future: 10 guiding principles for winning tomorrow's business", *HarperCollins Leadership*.

Mou, Y. and Xu, K. (2017), "The media inequality: comparing the initial human-human and human-AI social interactions", *Computers in Human Behavior*, Vol. 72, pp. 432-440, doi: 10.1016/j.chb.2017.02.067.

Nass, C. and Lee, K.M. (2000), "Does computer-generated speech manifest personality? An experimental test of similarity-attraction", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 329-336, doi: 10.1145/332040.332452.

Nass, C.I. and Lee, K.M. (2001), "Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction", *Journal of Experimental Psychology: Applied*, Vol. 7 No. 3, pp. 171-181, doi: 10.1037//1076-898x.7.3.171.

Nass, C. and Moon, Y. (2000), "Machines and mindlessness: social responses to computers", *Journal of Social Issues*, Vol. 56 No. 1, pp. 81-103, doi: 10.1111/0022-4537.00153.

Nass, C., Moon, Y., Fogg, B.J., Reeves, B. and Dryer, D.C. (1995), "Can computer personalities be human personalities?", *International Journal of Human-Computer Studies*, Vol. 43 No. 2, pp. 223-239, doi: 10.1145/223355.223538.

Neff, M., Wang, Y., Abbott, R. and Walker, M. (2010), "Evaluating the effect of gesture and language on personality perception in conversational agents", *Lecture Notes in Computer Science International Conference on Intelligent Virtual Agents*, Berlin, Heidelberg, Springer, pp. 222-235.

Okuno, H.G., Nakadai, K. and Kitano, H. (2003), "Design and implementation of personality of humanoids in human humanoid nonverbal interaction", *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, Berlin, Heidelberg, Springer, pp. 662-673.

Osborne, J.W. (2013), *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do before and after Collecting Your Data*, Sage, London.

Pal, D., Babakerkhell, M.D. and Roy, P. (2022), "How perceptions of trust and intrusiveness affect the adoption of voice activated personal assistants", *IEEE Access*, Vol. 10, pp. 123094-123113, doi: 10.1109/ACCESS.2022.3224236.

Parasuraman, R., Sheridan, T.B. and Wickens, C.D. (2000), “A model for types and levels of human interaction with automation”, *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 30 No. 3, pp. 286-297, doi: 10.1109/3468.844354.

Park, E., Jin, D. and del Pobil, A.P. (2012), “The law of attraction in human-robot interaction”, *International Journal of Advanced Robotic Systems*, Vol. 9 No. 2, p. 35, doi: 10.5772/50228.

Reeves, B. and Nass, C.I. (1996), *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places*, Cambridge University Press, Oxford.

Rovira, E. and Parasuraman, R. (2010), “Transitioning to future air traffic management: effects of imperfect automation on controller attention and performance, human factors”, *The Journal of the Human Factors and Ergonomics Society*, Vol. 52 No. 3, pp. 411-425, doi: 10.1177/0018720810375692.

Salem, M., Lakatos, G., Amirabdollahian, F. and Dautenhahn, K. (2015), “Would you trust a (faulty) robot?: effects of error, task type and personality on human-robot cooperation and trust”, *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, Academic Medicine/IEEE International Conference on Human-Robot Interaction ’15. HRI*, Portland, ACM, pp. 141-148, doi: 10.1145/2696454.2696497.

Siegel, M., Breazeal, C. and Norton, M.I. (2009), “Persuasive robotics: the influence of robot gender on human behavior” in International, R.S.J. (Ed.), *Conference on Intelligent Robots and Systems 2009*, IEEE, pp. 2563-2568.

Smith, M.A., Allaham, M.M. and Wiese, E. (2016), “Trust in automated agents is modulated by the combined influence of agent and task type”, *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 60 No. 1, pp. 206-210, doi: 10.1177/1541931213601046.

Spencer, J., Poggi, J. and Gheerawo, R. (2018), “Designing out stereotypes in artificial intelligence: involving users in the personality design of a digital assistant”, *Proceedings of the 4th EAI international conference on smart objects and technologies for social good – Goodtechs ’18. the 4th EAI International Conference*, Bologna, ACM Press, pp. 130-135, doi: 10.1145/3284869.3284897.

Tapus, A., Ţăpuş, C. and Matarić, M.J. (2008), “User—robot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy”, *Intelligent Service Robotics*, Vol. 1 No. 2, pp. 169-183, doi: 10.1007/s11370-008-0017-4.

Tay, B., Jung, Y. and Park, T. (2014), “When stereotypes meet robots: the double-edge sword of robot gender and personality in human-robot interaction”, *Computers in Human Behavior*, Vol. 38, pp. 75-84, doi: 10.1016/j.chb.2014.05.014.

Thórisson, K.R., Bieger, J., Thorarensen, T., Sigurðardóttir, J.S. and Steunebrink, B.R. (2016), “Why artificial intelligence needs a task theory: and what it might look like”, in Steunebrink, B., Wang, P. and Goertzel, B. (Eds), *Lecture Notes in Computer Science*, Springer International, Cham, pp. 118-128, doi: 10.1007/978-3-319-41649-6_12.

Veletsianos, G. and Miller, C. (2008), “Conversing with pedagogical agents: a phenomenological exploration of interacting with digital entities”, *British Journal of Educational Technology*, Vol. 39 No. 6, pp. 969-986, doi: 10.1111/j.1467-8535.2007.00797.x.

Weiner, B. (1985), “An attributional theory of achievement motivation and emotion”, *Psychological Review*, Vol. 92 No. 4, pp. 548-573, doi: 10.1037//0033-295x.92.4.548.

Weiss, A. and Evers, V. (2011), “Exploring cultural factors in human-robot interaction: a matter of personality?”, *paper presented at the 2nd International Workshop on Comparative Informatics*, Copenhagen, 9-10 December, available at: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.820.7082&rep=rep1&type=pdf (accessed 20 March 2021).

Wirtz, J., Patterson, P.G., Kunz, W.H., Gruber, T., Lu, V.N., Paluch, S. and Martins, A. (2018), “Brave new world: service robots in the frontline”, *Journal of Service Management*, Vol. 29 No. 5, pp. 907-931, doi: 10.1108/josm-04-2018-0119.

Złotowski, J., Sumioka, H., Eyssel, F., Nishio, S., Bartneck, C. and Ishiguro, H. (2018), "Model of dual anthropomorphism: the relationship between the media equation effect and implicit anthropomorphism", *International Journal of Social Robotics*, Vol. 10 No. 5, pp. 701-714, doi: 10.1007/s12369-018-0476-5.

Corresponding author
Sun Kyong Lee can be contacted at: sunnylee@korea.ac.kr