AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

© 2024 American Psychological Association
ISSN: 0096-3445

2025, Vol. 154, No. 2, 533–551
https://doi.org/10.1037/xge0001696

When the Bot Walks the Talk: Investigating the Foundations of Trust in an Artificial Intelligence (AI) Chatbot

Fanny Lalot and Anna-Marie Bertram
Faculty of Psychology, University of Basel

The concept of trust in artificial intelligence (AI) has been gaining increasing relevance for understanding and shaping human interaction with AI systems. Despite a growing literature, there are disputes as to whether the *processes* of trust in AI are similar to that of interpersonal trust (i.e., in fellow humans). The aim of the present article is twofold. First, we provide a systematic test of an integrative model of trust inspired by interpersonal trust research encompassing trust, its antecedents (trustworthiness and trust propensity), and its consequences (intentions to use the AI and willingness to disclose personal information). Second, we investigate the role of AI personalization on trust and trustworthiness, considering both their mean levels and their dynamic relationships. In two pilot studies ($N = 313$) and one main study ($N = 1{,}001$) focusing on AI chatbots, we find that the integrative model of trust is suitable for the study of trust in virtual AI. Perceived trustworthiness of the AI, and more specifically its ability and integrity dimensions, is a significant antecedent of trust and so are anthropomorphism and propensity to trust smart technology. Trust, in turn, leads to greater intentions to use and willingness to disclose information to the AI. The personalized AI chatbot was perceived as more able and benevolent than the impersonal chatbot. It was also more anthropomorphized and led to greater usage intentions, but not to greater trust. Anthropomorphism, not trust, explained the greater intentions to use personalized AI. We discuss implications for research on trust in humans and in automation.

Public Significance Statement

This study demonstrates that people judge artificial intelligence chatbots the same way they judge other humans: Their trust judgments are underpinned by the same psychological mechanisms. This study advances our understanding of trust in artificial intelligence in the digitalization age.

Keywords: artificial intelligence, human–automation trust, interpersonal trust, trust in technology, trust in artificial intelligence

Artificial intelligence (AI) based systems have been integrated into daily life for a few decades, and their adoption has significantly accelerated in recent years. From recommendation algorithms to virtual assistants, chatbots, autonomous vehicles, and AI-driven medical diagnostics, AI technologies are becoming an integral part of modern society (Grace et al., 2018). Amid this rapid expansion, the concept of trust in AI has gained relevance for understanding and shaping human interaction with AI systems. An interdisciplinary body of research has endeavored to better understand trust in AI (Kaplan et al., 2023) and more broadly trust in technology (Lankton et al., 2015) or automation (Lee & See, 2004). Yet, questions remain. Most centrally, there are disputes as to whether the *processes* of trust in AI—where technology becomes the trustee—are similar to that of interpersonal trust (see, e.g., Calhoun, 2017). Moreover, thorough tests of complete models of trust in AI are scarce.

The aim of the present article is twofold. First, we provide a thorough systematic test of an integrative model of trust in AI inspired by interpersonal trust research encompassing trust, its antecedents (trustworthiness and trust propensity), and its consequences (intentions

This article was published Online First December 5, 2024.

Aysecan Boduroglu served as action editor.

Fanny Lalot https://orcid.org/0000-0002-1237-5585

All data and code for the analyses, as well as additional online material, are publicly available on the Open Science Framework at https://osf.io/gs7jh/. The data and ideas presented in this article were not presented at any past conferences nor posted on any website or repository.

This research was funded in whole, or in part, by the Swiss National Science Foundation (SNSF) (Grant PZ00P1_216373/1) awarded to Fanny Lalot. For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

Fanny Lalot contributed to conceptualization, data curation, formal analysis, investigation, methodology, validation, visualization, and writing–original draft. Anna-Marie Bertram contributed to conceptualization, methodology, visualization, and writing–review and editing.

Fanny Lalot played a lead role in data curation, formal analysis, investigation, validation, and writing–original draft and an equal role in conceptualization, methodology, and visualization. Anna-Marie Bertram played a lead role in writing–review and editing and an equal role in conceptualization, methodology, and visualization.

Correspondence concerning this article should be addressed to Fanny Lalot, Faculty of Psychology, University of Basel, Missionsstrasse 64A, 4055 Basel, Switzerland. Email: fanny.lalot@unibas.ch

to use the AI system and willingness to disclose personal information). Second, we investigate the role of AI personalization on trust and trustworthiness, considering both their mean levels and their dynamic relationships. We report here the results of two pilot studies ($N = 313$) and one preregistered main study ($N = 1{,}001$) that shine light on these open questions.

An Integrative Model of Interpersonal Trust

The concept of trust has been widely researched across disciplines with different foci, encompassing, for example, close relationships, workplace dynamics, social trust and orientation toward strangers, political trust, etc. Here, we adopt a widely accepted definition of *trust* put forth by Rousseau et al. (1998): "a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behaviour of another" (p. 395). A strength of this definition is that it can be applied to trust broadly and not just interpersonal trust, as long as "another" also includes nonhuman agents (e.g., trust in technology; Lee & See, 2004).

Mayer et al. (1995) proposed an integrative model of trust that remains influential (see also Kelton et al., 2008; Schoorman et al., 2007, for extensions). Consistent with the definition proposed above, the model posits trust as an attitude that will be weighed against perceived risk to determine risk-taking (or accepting vulnerability) in the present situation. It also incorporates two important antecedents of trust: trustworthiness and trust propensity.

Trustworthiness reflects a set of characteristics of the trustee. According to Mayer and colleagues' ABI model, trustworthiness consists of at least three elements (see also Colquitt et al., 2007): the trustee's *ability* ("that group of skills, competencies, and characteristics that enable a party to have influence within some specific domain," Mayer et al., 1995, p. 717), *benevolence* ("the extent to which the trustee is believed to want to do good to the trustor, aside from an egocentric profit motive," Mayer et al., 1995, p. 718), and *integrity* ("the perception that the trustee adheres to a set of principles that the trustor finds acceptable," Mayer et al., 1995, p. 719). Trustworthiness is evaluated through observation of the trustee's actions and gradually builds on repeated interactions. *Trust propensity* represents a stable characteristic of the trustor reflecting generalized beliefs about how much others in general can be trusted (Rotter, 1971). From a social learning perspective, trust propensity is believed to arise from positive and negative early life interactions, which would work as reinforcement and lead to generalized expectations.

Trustworthiness and trust propensity are conceived as two intertwined determinants of trust: trust is greater when the trustee proves trustworthy and when the trustor has a general inclination to trust. Some have noted that trust propensity also indirectly influences trust by acting as a filter coloring perceived trustworthiness in a direction that is congruent with one's disposition (i.e., a confirmation bias; see, e.g., Colquitt et al., 2007).

Finally, trust does not directly equate to behavior. Rather, it is the positive attitude that is weighed against perceived risk and facilitates trusting behavior (i.e., the behavioral manifestation of the willingness to accept vulnerability in a given situation; Mayer et al., 1995). This important distinction highlights that while trust is a key antecedent of trusting behavior, it is not the only one, and other factors such as social norms (Dunning et al., 2014) or normative expectations and reputational concern (Pfattheicher, 2015) may also come into play.

Models of Trust in AI

Defining AI

Before diving into the complexities of trust in AI, its antecedents, and its consequences, it is important to briefly specify how we conceptualize AI. We adopt the Organisation for Economic Co-Operation and Development's definition of AI as a:

> Machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments. It uses machine and/or human-based inputs to perceive real and/or virtual environments; abstracts such perceptions into models … ; and uses model inference to formulate options for information or action. (Organisation for Economic Co-Operation & Development, 2019, Executive summary)

Research on trust in AI has alternatively focused on specific forms of it (e.g., technology, automation, robots) or on AI in general. In this section, we aim to remain close to the terminology adopted by the authors we cite.

Trusting AI or Its Creator?

Even before technology started resembling humans, speaking in a humanlike voice or taking the traits of an anthropomorphized robot, scholars have argued that (calibrated) trust in technology would be key to allow appropriate reliance on the technology (i.e., avoiding both misuse and disuse; Lee & See, 2004). Several pieces of research have found trust to positively predict the usage of (and intentions toward) a technology (Choung et al., 2023).

Yet, it is not always clear whether research captures dynamics of trust in the technology itself, or rather, by proxy, trust in the humans and/or organizations perceived as responsible for the implementation of the technology. McKnight et al. (2011) were among the first to explicitly call for a clear distinction between the two. Some claim that trust in technology can only reflect trust in its human creators, following the often-quoted remark by Friedman et al. (2000): "People trust people, not technology" (p. 36). Only moral agents equipped with free will may be capable of willingly choosing to do bad or good and thus be objects of trust, so machines may not be considered as such (Solomon & Flores, 2001, as cited in Corritore et al., 2003)—at best, they could be *reliable* but not trustworthy (Hatherley, 2020).

Against this philosophical argument, a large body of evidence shows that people actually treat new technologies as real people and may therefore perceive technology as a direct object of trust (Reeves & Nass, 1996; see also Gillath et al., 2021; Song et al., 2022). It has also been argued that because technology is not flawless, deciding to rely on it (without being able to control the reliability of every decision it makes) resembles accepting vulnerability based on positive expectations, that is, trusting (Ferrario et al., 2021; Kaplan et al., 2023). Others have demonstrated that trust in technology can be reliably measured with psychometric tools (e.g., Choung et al., 2023; Gulati et al., 2019; Jian et al., 2000; McKnight et al., 2002, 2011).

Trust in AI: Purpose, Process, Performance, Anthropomorphism, and Other Factors

Similar to the literature on interpersonal trust, research has identified three main sources of trust in technology (or AI): the human (trustor), the technology itself (trustee), and the context (Bach et al., 2024; Kaplan et al., 2023). On the human side of things, dimensions such as personal expertise or understanding of AI (Bach et al., 2024; Glikson & Woolley, 2020), as well as propensity to trust (Hoff & Bashir, 2015; Lee & See, 2004), have a positive impact on trust in AI. Several demographics (e.g., culture, gender) also play a role (Hoff & Bashir, 2015; Kaplan et al., 2023). In terms of contextual factors, authors have noted a beneficial effect of "teaming-related" variables. For example, communication via shared verbal speech increases trust compared to text-based communication. Trust also increases with the length of the relationship between the user and the AI. On the other hand, perceived risk in the task at stake (but not task complexity) dampened trust (Kaplan et al., 2023).

Finally, there are technology-related antecedents of trust. Early work identified three dimensions of technology trustworthiness (PPP): *purpose* (a perception of "the underlying motives or intents" of the technology; Lee & Moray, 1992, p. 1246), *process* ("an understanding of the underlying qualities or characteristics that govern behaviour" of the technology Lee & Moray, 1992, p. 1246), and *performance* ("the expectation of consistent, stable, and desirable performance or behaviour," Lee & Moray, 1992, p. 1246). Meta-analytical findings (Kaplan et al., 2023) support the idea of a positive effect of all three factors on trust, noting the particularly strong impact of perceived reliability and performance of the technology ("performance"; see also Hancock et al., 2011) as well as its transparency ("process"; see also Bach et al., 2024; Glikson & Woolley, 2020; Hoff & Bashir, 2015; Lockey et al., 2021). In addition, reviews and meta-analyses have identified an influence of past interactions with the technology, suggesting that users adjust their trust in the technology based on its observed "behavior"—just as with other humans (Bach et al., 2024; Hoff & Bashir, 2015; Kaplan et al., 2023).

Finally, anthropomorphism (i.e., the process of attributing distinctively human characteristics to nonhumans) was identified as an important factor positively impacting trust in technology (Kaplan et al., 2023; Waytz et al., 2014), notably because it increases perceived benevolence (Bach et al., 2024; Calhoun et al., 2019). Anthropomorphism also makes trust more resilient (Lockey et al., 2021). However, it has been noted that too much anthropomorphizing can backfire if it creates a sense of eeriness (the uncanny valley theory), which can then decrease trust (Glikson & Woolley, 2020; Hoff & Bashir, 2015).

Parallels Between Trust in AI and Interpersonal Trust

There are many parallels between trust in AI and interpersonal trust. Notably, work shows that people project similar sentiments on machines and on humans (Birnbaum et al., 2016; Song et al., 2022) and that personal traits and dispositions similarly affect both types of trust (Gillath et al., 2021; Kaplan et al., 2023), supporting the assumption that people treat technology as real people (Reeves & Nass, 1996).

In addition, several authors have explicitly discussed the similarities between trust in AI and interpersonal trust. Lee and Moray (1992) made a direct connection between the dimensions of purpose, process, and performance, which they introduced, with Rempel et al.'s (1985) model of human trust including faith, dependability, and predictability. When later refining the PPP model, Lee and See (2004) reviewed a number of models of interpersonal trust, which again showed mostly similarity with their tripartite model. Importantly, they incorporated Mayer et al.'s (1995) ABI model, linking performance to ability, process to integrity, and purpose to benevolence. In a similar vein, McKnight et al. (2011) offered a conceptual comparison between trust in people and trust in technology, aligning the constructs of ability with functionality, benevolence with helpfulness, and integrity with reliability.

Despite the resemblance between the PPP model and the ABI model, there are subtle differences between the respective pairs of dimensions as well as in the measures utilized to assess them (see Madhavan & Wiegmann, 2007). In a comparative study, Lankton et al. (2015) demonstrated that the degree of perceived humanness of technology was key: trust in low-humanness technologies (e.g., Microsoft Access) was best captured by the PPP model ("system-like trust"), while trust in high-humanness technologies (e.g., Facebook) was best captured by the ABI model ("humanlike trust"). Given that AI-based technology is increasingly gaining in humanness (e.g., rich conversation either in writing or orally, increasing responsiveness to the user's queries and reactions), it can therefore be assumed that the ABI model will become more and more relevant to assess trust in AI, an argument that recent work supports (Calhoun et al., 2019; Choung et al., 2023; Gulati et al., 2019).

However, contentions remain about how much human models of trust can indeed be applied to AI (and generally, to technology) and specifically whether the dimensions of benevolence and integrity are necessary to assess AI trustworthiness above and beyond mere ability. There is also disagreement on whether people really think about AI in terms of inner benevolence (i.e., attributing it sufficient self-direction to be able to harvest positive or negative intentions toward its user or to be (dis)honest) or whether they are merely judging the benevolence of the technology's creator ("purpose").

Crucially, although research on the topic has grown considerably in recent years, publications tend to focus on some specific components of trust or trustworthiness only. As a result, thorough tests of a complete model remain scarce. Some articles also confound trust with trustworthiness, trust propensity, or trusting behavior—a conceptual issue that is not new to the field of trust research but contributes to make things more complex (Muir, 1994).

An Integrative Model of Trust in AI

To address this important gap, the first aim of the present article was to provide a thorough test of an integrative model of trust, building on sound measurement and advanced statistical methods to investigate the relationships between trust in AI, its antecedents, and its consequences. This leads to

> *Research Question 1:* How well does an integrative model of trust apply to trust in AI?

We argue that, given the increasing humanness of AI-based technology (Lankton et al., 2015), AI trustworthiness can be reliably approached through the ABI model (ability, benevolence, integrity; Mayer et al., 1995). In line with recent findings, we hypothesize that all three dimensions positively contribute to trust in AI. We also expect that perceived anthropomorphism would be related to greater trust, both directly and indirectly through greater trustworthiness

(Calhoun et al., 2019). Turning to human factors, we consider individual differences in propensity to trust other humans and propensity to trust technology (see McKnight et al., 2011). We expect that trust propensity, especially in technology, would be related to greater trust in AI (see Huang et al., 2024).

We finally consider two consequences of trust: intentions to use the AI (e.g., Choung et al., 2023; McKnight et al., 2011) and willingness to disclose personal information to it (Birnbaum et al., 2016; Lucas et al., 2014). Both have been recognized as being positively influenced by trust and are relevant to the question of technology use (vs. disuse).

Does AI Personalization Impact Trust?

The second aim of this article was to investigate the role of AI personalization on trust. AI personalization refers to the use of AI techniques to tailor experiences and information to individual users in various domains (Rafieian & Yoganarasimhan, 2023). Personalized AI can adapt to its specific user, learn and remember their personal style and preferences, and gradually provide more tailored advice and solutions.

In the past, AI-based technology was not able to offer a high degree of personalization. However, embedded AI (e.g., recommendation algorithms) is becoming increasingly widespread, providing curated content to social media users (Guess et al., 2023), personalized advertisement (Mogaji et al., 2020), or tailored human–AI learning methods (Molenaar, 2021). In addition, virtual AI (e.g., conversational agents) is developing rapidly and offers increasing possibilities of personalization, to which users seem to react rather positively (Kronemann et al., 2023). As long as there is no privacy concern about the use of personal data, personalized AI can lead to greater feelings of being understood by the technology and to an improved experience, which in turn may increase positive attitudes toward the AI as well as greater willingness to disclose personal information to it (Aguirre et al., 2015; Bleier & Eisenbeiss, 2015). A separate line of research on immediacy behavior similarly suggests that users react more positively to an AI that responds to them adequately and adapts its response to that of the user (Komiak & Benbasat, 2006).

Despite this growing literature, it remains an open question whether AI personalization also increases trust and more specifically through which pathways (e.g., through increased trustworthiness of the personalized AI and/or increased anthropomorphism). In addition, we lack comparative studies that would directly assess and compare not only the *mean levels* of trustworthiness of, and trust in, personalized versus nonpersonalized AI but also the *mechanisms* underlying trust. For example, it is possible that the relative weight of benevolence and integrity, compared to ability, differs for personalized versus nonpersonalized AI. Alternatively, it could be that trust does not predict intentions to use the AI to the same extent in both cases. The second aim of this article was to shed some light on these open questions:

> *Research Question 2:* What are the differences between a personalized and nonpersonalized AI in terms of mean levels of trustworthiness, trust, anthropomorphism, intentions to use, and willingness to disclose personal information? Are there differences in the underlying mechanisms, that is, in the strength of the relationships between variables?

The Present Research

The present research aims to address the two aforementioned research questions. To this aim, we conducted three studies: two small-scale pilot studies that aimed to develop reliable measures of trust and other key constructs and to pretest an experimental manipulation of AI personalization and one large-scale study that directly addressed the research questions.

A Design Fiction Approach to Investigate AI Chatbots

Our studies focus on chatbots, that is, virtual AI taking the form of a conversational agent with a distinguished identity (Glikson & Woolley, 2020). Chatbots such as OpenAI's ChatGPT have grown exponentially in recent years, launching a renewed debate on the necessity of AI regulations (Clayton, 2023). More people than ever are now relying on such conversational agents, which are evolving to become more powerful and reliable as well as more personalized. For example, Luka's Replika is a generative AI chatbot app designed to act as the user's "friend." It is therefore of scientific and practical importance to better understand how users decide to trust such chatbots.

The current research adopts a "design fiction" approach (Gulati et al., 2019): Instead of having participants interact with a technology directly, we present them with a description of said technology, provide a range of pieces of information and examples of representative technology–user interactions, and ask them to imagine having the opportunity to use the technology. The design fiction approach allows us to compare two types of AI under controlled conditions (Bleecker, 2022; Blythe, 2014), ensuring sound experimental design. Research using such hypothetical use scenarios or vignettes has been successful in the past, providing that participants can successfully imagine themselves interacting with the technology (Gillath et al., 2021; Gulati et al., 2019; Juravle et al., 2020).

This choice of design has implications for research on trust. As highlighted by Hoff and Bashir (2015), different factors may impact trust in AI *prior* to interaction (i.e., initial reliance strategy) and *during/after* initial interaction (i.e., dynamic learned trust). Here, we aim to put participants in the early stage of interacting with a new AI. To this end, we provide participants with concrete examples of interactions between the chatbot and a user. We also ask them to imagine interacting with the chatbot and to write down questions they would like to ask it. As such, participants get basic information about design features and system performance and we can theoretically expect their trust to be driven by their initial impressions arising from early interaction (Hoff & Bashir, 2015).

Transparency and Openness

The research was conducted in compliance with American Psychological Association ethical standards in the treatment of participants, and it was approved by the ethics committee from the Faculty of Psychology at the University of Basel. The first pilot study and the main study were preregistered (study design, materials, sample size, and rules for exclusion: https://aspredicted.org/wqjx-cjj2.pdf and https://aspredicted.org/kdvj-xkhn.pdf). All data and code for analysis, as well as additional online material, are publicly available on the Open Science Framework (OSF) at https://osf.io/gs7jh/.

Pilot Study 1: Refining the Measurement Scales

The aim of the first pilot study was to test and refine items that could be utilized to reliably measure the constructs of interest. We drew from the literature to identify relevant scales and tested the psychometric properties of their combined items. The study design, materials, sample size, and rules for exclusion were preregistered at https://aspredicted.org/wqjx-cjj2.pdf.

Method

Participants and Procedure

Participants were recruited on the crowdsourcing platform Prolific. Criteria for participation were to be 18 years of age or older and to live in the United Kingdom. As preregistered, based on budget constraints and feasibility, we aimed to recruit 250 participants. Sample size recommendations for factor analysis vary. A commonly used rule of thumb is a ratio of participants to estimated parameters of 10:1 (e.g., Bentler & Chou, 1987). However, recent advances and simulation studies show that the adequacy of this ratio may vary drastically based on the quality of the data or factor strength. Having continuous and normally distributed data, high reliability scores, and no missing data notably decreases the required sample, so that a 5:1 ratio could be sufficient (Kyriazos, 2018). In this study, we ran a series of exploratory factor analyses on subsets of items (one theoretical construct at a time), for which the 10:1 ratio was clearly exceeded. We also conducted one confirmatory factor analysis testing our final measurement model with 35 retained items (see below). For this analysis, most latent constructs were defined by four items or more, loadings were high (most exceeded .70), and the model was rather simple (no relationship specified between latent constructs). We are therefore confident that the planned sample size—meeting the 5:1 if not the 10:1 ratio—was adequate for the planned analyses.

Two hundred forty-nine participants completed the study but one participant failed the attention checks embedded in the questionnaire and was therefore excluded from analyses (as preregistered), resulting in $N = 248$. In this study and the following ones, participants were asked to indicate their age in years and their gender (options: a man, a woman, nonbinary/other, prefer not to say). There were 123 men and 125 women ($M_{\text{age}} = 44.52$, $SD = 23.35$).

The study was presented as an early stage of research where we were "trying to develop good items that can adequately capture people's opinions." It introduced a fictitious app called "Conversea," a chatbot using an AI system to maintain conversations with its users and aiming to serve as a personal assistant, providing advice when requested. Following the description of the app, including a list of typical requests that users might ask it, participants were asked to imagine they had been using the app for a while now, that it had so far given them satisfactory answers, and that they now needed to report what they think about the app. They then answered questions pertaining to trustworthiness, trust, and so forth (see below). They were finally thanked for their participation and remunerated.

Materials

We identified relevant items in validated scales from the literature (the complete list is available in additional online material ESM1 at https://osf.io/gs7jh/). Specifically, we borrowed items from Choung et al. (2023), Golossenko et al. (2020), Mayer and Davis (1999), McKnight et al. (2002, 2011), Moussawi and Koufaris (2019), as well as from the European Social Survey and the International Social Survey Programme's Social Relations and Support Systems module. We also included a few self-developed items used in our own previous research. This resulted in 65 items (10 items measuring ability, nine benevolence, seven integrity, eight trust, five human trust propensity, 12 machine trust propensity, 11 anthropomorphism, and three intentions to use). All items were measured on a 7-point Likert scale (1 = *strongly disagree*, 7 = *strongly agree*).

Analysis Strategy and Results

We conducted a series of exploratory factor analyses, one per theoretical construct, with the aim of obtaining valid and reliable sets of three to six items per construct. For each analysis, we first investigated different factor retention criteria to determine how many factors should be extracted (Steiner & Grieder, 2020). We then ran exploratory factor analyses based on these recommendations. Decisions to retain or remove items were based on a combination of factor loadings, item content, and indices of model fit when relevant (e.g., for multidimensional constructs such as trustworthiness). We finally computed indices of reliability for each construct (Cronbach's α and McDonald's ω when relevant). For brevity purposes, we do not report the details of the results here; this can be found in additional online material ESM2 at https://osf.io/gs7jh/. The complete annotated code is also available on the OSF (additional online material ESM3), https://osf.io/gs7jh/. This process allowed us to retain a final set of 35 items that could be used in the main study.

As a final step, we tested a global model of these 35 items forming six separate constructs with a confirmatory factor analysis. We considered the following indices to assess the model fit: root-mean-square error of approximation (RMSEA; Steiger & Lind, 1980) and standardized root-mean-square residual (SRMR; Bentler, 1995). Hu and Bentler (1999) indeed advised the use of a "two-index presentation strategy" to minimize both Type I and Type II errors. RMSEA has, moreover, been declared one of the most informative fit indices (Diamantopoulos & Siguaw, 2000). We also report the comparative fit index (CFI; Bentler, 1990) and χ^2. Typically, CFI > .90, RMSEA < .08, and SRMR < .09 indicate an acceptable fit (MacCallum et al., 1996). This model yielded satisfactory fit, $\chi^2(538) = 1{,}063$, $\chi^2/df = 1.98$, CFI = .923, RMSEA = .063, 90% CI [.057, .068], SRMR = .096. Results of the confirmatory factor analysis are reported in Tables 1 and 2 provides a summary of the number of items and reliability of each construct.

Pilot Study 2: Pretesting the Tailored Virtual AI Manipulation

The aim of the second pilot study was to ensure that the planned manipulation of a tailored virtual AI was sound, believable, and easily understandable for participants. We therefore presented the manipulated information to a small sample of participants, followed by a few comprehension check questions.

Table 1

Results of the Confirmatory Factor Analysis Testing Our Global Measurement Model (Pilot Study 1)

Measurement model	*b* (*SE*)	95% CI	*z* test	*p*	β
Ability = ~					
ability_3	1.00				.842
ability_4	1.01 (.060)	[0.89, 1.12]	16.80	<.001	.861
ability_6	1.28 (.076)	[1.12, 1.42]	16.86	<.001	.863
ability_9	1.10 (.072)	[0.95, 1.24]	15.28	<.001	.810
Benevolence = ~					
benevol_1	1.00				.859
benevol_2	1.06 (.062)	[0.93, 1.17]	17.11	<.001	.855
benevol_4	1.06 (.059)	[0.93, 1.17]	17.78	<.001	.875
benevol_6	0.89 (.056)	[0.78, 0.99]	15.93	<.001	.818
Integrity = ~					
integr_1	1.00				.836
integr_2	1.06 (.069)	[0.92, 1.19]	15.32	<.001	.828
integr_3	1.18 (.076)	[1.02, 1.32]	15.43	<.001	.832
integr_5	1.11 (.091)	[0.93, 1.29]	12.21	<.001	.703
Trustworthiness = ~					
Ability	1.00				.813
Benevolence	1.34 (.160)	[1.02, 1.65]	8.38	<.001	.614
Integrity	1.26 (.114)	[1.03, 1.48]	11.06	<.001	.876
Trust = ~					
trust_1	1.00				.686
trust_2	1.00 (.091)	[0.81, 1.17]	10.97	<.001	.773
trust_3	1.02 (.085)	[0.85, 1.19]	11.97	<.001	.858
trust_6	−0.85 (.085)	[−1.01, −0.68]	−9.96	<.001	−.694
trust_7	−0.86 (.091)	[−1.04, −0.68]	−9.48	<.001	−.658
Own thinking = ~					
anthropo_1	1.00				.845
anthropo_2	0.93 (.074)	[0.79, 1.07]	12.69	<.001	.841
Emotions = ~					
anthropo_5	1.00				.774
anthropo_6	1.04 (.029)	[0.97, 1.09]	35.24	<.001	.789
anthropo_7	1.57 (.115)	[1.34, 1.79]	13.62	<.001	.955
Human interaction = ~					
anthropo_8	1.00				.907
anthropo_9	0.95 (.054)	[0.83, 1.05]	17.41	<.001	.867
anthropo_10	0.84 (.055)	[0.73, 0.94]	15.27	<.001	.788
Anthropomorphism = ~					
Own thinking	1.00				.814
Emotions	0.73 (.092)	[0.55, 0.91]	7.94	<.001	.907
Human interaction	0.85 (.115)	[0.62, 1.07]	7.42	<.001	.580
Intentions to use = ~					
intention1	1.00				.932
intention2	1.06 (.035)	[0.99, 1.12]	30.18	<.001	.953
intention3	1.08 (.035)	[1.01, 1.14]	31.00	<.001	.960
Trust propensity in humans = ~					
tp_techno_2	1.00				.841
tp_techno_4	1.04 (.068)	[0.90,1.17]	15.26	<.001	.840
tp_techno_6	1.17 (.097)	[0.98, 1.36]	12.16	<.001	.708
tp_techno_10	0.95 (.071)	[0.81, 1.08]	13.44	<.001	.763
Trust propensity in smart technology = ~					
gentrust1	1.00				.810
gentrust2	1.03 (.088)	[0.85, 1.20]	11.71	<.001	.871
gentrust3	−0.76 (.073)	[−0.90, −0.61]	−10.36	<.001	−.668

Note. Confidence intervals are percentile bootstrap confidence intervals. Item labels are reported in additional online material ESM1 at https://osf.io/gs7jh/. CFA = confirmatory factor analysis; *SE* = standard error; CI = confidence interval.

Method

Participants and Procedure

Participants were recruited on the crowdsourcing platform Prolific. As in Pilot Study 1, criteria for participation were to be 18 years of age or older and to live in the United Kindom. An a priori power analysis determined that a sample size of 68 would be sufficient to detect a medium-to-large effect (i.e., Cohen's $d = .70$, which could easily be expected for comprehension checks) with .80 power. Sixty-five participants completed the study but one failed

Table 2

Results of Pilot Study 1: Number of Items Retained and Index of Reliability for Each Construct

Construct	No. of item	Cronbach's α	McDonald's ω_T
Trustworthiness	12	.92	.95
Ability	(4)	.90	
Benevolence	(4)	.91	
Integrity	(4)	.87	
Trust	5	.86	
Anthropomorphism	8	.88	.95
Own thinking	(2)	.83	
Emotions	(3)	.93	
Human interactions	(3)	.89	
Intentions to use	3	.96	
Trust propensity in humans	3	.82	
Trust propensity in smart technology	4	.87	

Note. We report McDonald's omega total (ω_T) for the two multidimensional scales (trustworthiness and anthropomorphism).

both attention checks and was excluded from analyses, resulting in $N = 64$ participants, among whom 34 men and 30 women ($M_{\text{age}} = 43.52$, $SD = 14.91$).

The study was presented as an investigation of people's views about new smartphone applications. As in Pilot Study 1, it introduced a fictitious app called "Conversea." The description of the app, however, differed across experimental conditions, and participants were randomly allocated to either the tailored ($n = 31$) or the impersonal ($n = 33$) virtual AI condition. Given that the manipulation remained the same for the main study, we describe it in further detail in the Main Study: Trust in Tailored Versus Impersonal Virtual AI section. Participants read the description and were provided with four examples of interactions between users and the app; they then answered a few questions serving as comprehension checks. They were finally thanked for their participation and remunerated.

Materials

Two questions were used as comprehension checks of the experimental manipulation. The first item was a multiple-choice question asking, "Conversea can learn to know you as a person and will develop more personalized advice the more you use it": True/False/I don't know. The second item utilized a 7-point Likert scale, "Based on the questions and answers presented above, how much would you say that the answers provided by Conversea were tailored to the specific user asking the questions?" (1 = *not at all*, 7 = *completely*).

Results

Code for the analyses is reported on the OSF (additional online material ESM4), https://osf.io/gs7jh/. Results revealed that participants had overwhelmingly understood whether the virtual AI was tailored or impersonal: All participants in the tailored AI condition correctly answered "true" to the question whether the app could "learn to know them as a person and develop personalized advice" and all but one correctly answered "false" in the impersonal AI condition (1 saying "I don't know"; i.e., total of 98.4% correct answers).

Participants in the tailored AI condition also judged that the answers provided by the app were tailored to the specific user asking the questions ($M = 5.35$, $SD = 0.98$) much more than participants in the impersonal AI condition ($M = 3.24$, $SD = 1.77$), $F(1, 62) = 34.24$, $p < .001$, Cohen's $d = 1.46$, 95% CI [0.91, 2.01]. We therefore concluded that the manipulation was working as intended and could be utilized in the main study.

Main Study: Trust in Tailored Versus Impersonal Virtual AI

Having identified items that would adequately capture our key constructs (Pilot Study 1) and asserted that the planned manipulation of tailored versus impersonal virtual AI was well understood by participants (Pilot Study 2), we moved on to the main study. The study's aim was twofold: first, to provide a systematic test of an integrative model of trust in virtual AI; and second, to investigate whether a tailored virtual AI would be perceived differently from an impersonal virtual AI and whether the psychological underpinnings of trust in both would be similar or different. The study design, materials, sample size, and rules for exclusion were preregistered at https://aspredicted.org/kdvj-xkhn.pdf.

Method

Participants and Procedure

Similar to the pilot studies, participants were recruited on the crowdsourcing platform Prolific. Criteria for participation were to be 18 years of age or older and to live in the United Kingdom. As preregistered, we aimed to recruit 1,000 participants to provide sufficient power for the multiple-group structural equation model (SEM). A total of $N = 1{,}001$ participants completed the study, among whom 494 men, 495 women, 10 identifying as nonbinary or "other," and two undisclosed ($M_{\text{age}} = 41.63$, $SD = 13.57$).[1]

The study introduced "Conversea," a chatbot using an AI system to maintain conversation with its users and aiming to serve as a personal assistant, providing advice when requested. Participants read the description of the app which, depending on the experimental condition,

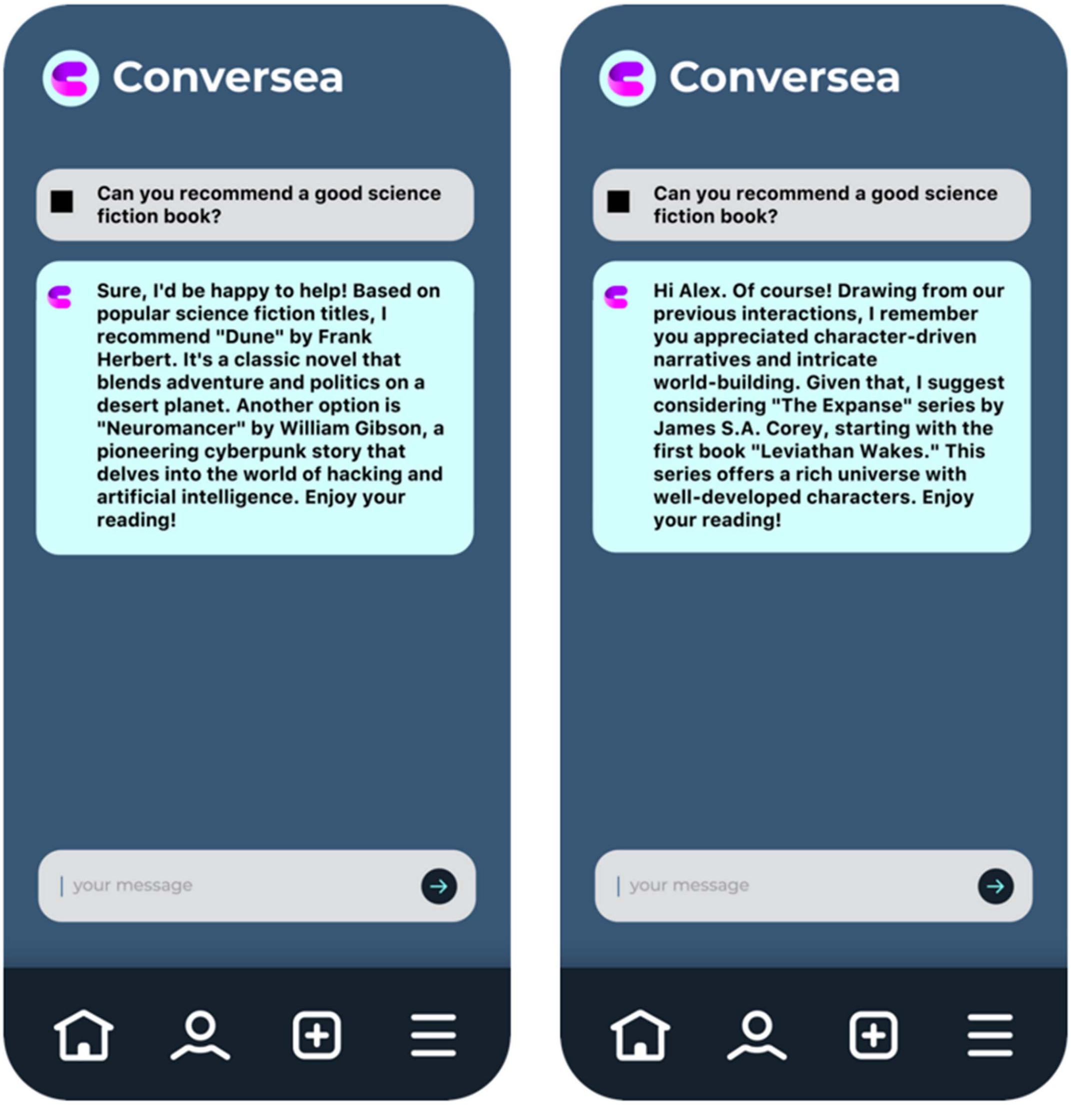
[1] We preregistered the exclusion of participants who would fail both attention checks embedded in the questionnaire. However, no participant failed both checks, leading to no exclusion (191 participants failed one check).

introduced it as being either tailored (i.e., personalized or user-centric; $n = 504$) or impersonal ($n = 497$; see below). Then, they were shown four examples of interactions between users and the app, which were meant to give them an idea of how the app worked. These appeared in the form of smartphone screenshots showing a question asked by an anonymous user and the answer provided by the app (see Figure 1). To control for participants' perceptions about other aspects of the app development and notably its creators, we finally provided additional background information, which was similar across experimental conditions. As such, we specified that (a) the app was developed and commercialized in the United Kingdom (i.e., participants' country); (b) it was developed by a well-established company that has already successfully developed many such smartphone apps; (c) there was a clear agreement about the use of data, which could be reviewed by users at any time; and (d) the app costed £10/month on a monthly contract or £100 a year on a yearly contract. Therefore, any difference between experimental conditions should be imputable to the AI itself and not to differing assumptions about its creators.

To increase the realism of the situation and subjective engagement with the app, participants were asked to imagine they had received the app as a gift from a friend and had started using it, and they were asked to write down three questions or queries they would like to ask the app. They then moved to the final part of the questionnaire, which assessed the perceived trustworthiness of, and trust in, the app, as well as anthropomorphism, intentions to use the app in the future, willingness to disclose personal information, and propensity to trust humans as well as smart technology. Participants were finally fully debriefed, thanked for their participation, and remunerated.

Figure 1

Example of Interaction Between the Impersonal (Left) Versus Tailored (Right) Virtual Artificial Intelligence Provided to Participants



Can you recommend a good science fiction book?

Sure, I'd be happy to help! Based on popular science fiction titles, I recommend "Dune" by Frank Herbert. It's a classic novel that blends adventure and politics on a desert planet. Another option is "Neuromancer" by William Gibson, a pioneering cyberpunk story that delves into the world of hacking and artificial intelligence. Enjoy your reading!

Can you recommend a good science fiction book?

Hi Alex. Of course! Drawing from our previous interactions, I remember you appreciated character-driven narratives and intricate world-building. Given that, I suggest considering "The Expanse" series by James S.A. Corey, starting with the first book "Leviathan Wakes." This series offers a rich universe with well-developed characters. Enjoy your reading!

Note. See the online article for the color version of this figure.

Materials

Tailored Versus Impersonal AI Manipulation. Participants were randomly allocated to one experimental condition and accordingly read different descriptions of the virtual AI chatbot app. Specifically, in the *impersonal* AI condition, the text read:

> According to its designers, Conversea uses a flexible algorithm that will be able to learn from its interactions with users and improve continuously. However, Conversea is not entirely customisable: while it can maintain conversation with you and "remember" what was said some minutes ago, it will forget about you from one conversation to the next. Therefore, it cannot learn about your personal preferences or style. Its development will only be driven by the aggregate feedback given by all users.

In the *tailored* virtual AI condition, the text read:

> According to its designers, Conversea uses a flexible algorithm that will be able to learn from its interactions with users and improve continuously. Interestingly, Conversea is entirely customisable: not only can it maintain conversation with you and "remember" what was said some minutes ago, it will also remember you from one conversation to the next. Therefore, it can learn about your personal preferences or style, and it will develop to provide answers that are more and more relevant to you personally.

The four examples of interactions between users and the app that followed were also modified between conditions to reflect rather tailored or impersonal answers. One example is depicted in Figure 1. All materials are also reported in additional online material ESM5 at https://osf.io/gs7jh/.

Comprehension Checks. The same two questions as in Pilot Study 2 were used as comprehension checks of the experimental manipulation: a multiple-choice question asking, "Conversea can learn to know you as a person and will develop more personalized advice the more you use it": True/False/I don't know; and a 7-point Likert scale item, "Based on the questions and answers presented above, how much would you say that the answers provided by Conversea were tailored to the specific user asking the questions?" (1 = *not at all*, 7 = *completely*).

Measures. Items were selected based on the results from Pilot Study 1 (see above). Unless stated otherwise, all items were rated on a 7-point Likert scale (1 = *strongly disagree*, 7 = *strongly agree*). Specifically, the *ability* (e.g., "Conversea is very capable of performing its job"), *benevolence* (e.g., "Conversea is concerned about addressing the problems of human users"), and *integrity* (e.g., "Conversea keeps its commitments and delivers on its promises") components of trustworthiness were each measured with four items taken/adapted from Choung et al. (2023), Mayer and Davis (1999), and McKnight et al. (2002).

Anthropomorphism was measured with eight items taken from Golossenko et al. (2020) and Moussawi and Koufaris (2019) capturing three facets: capacity for independent thinking (e.g., "Conversea can imagine things on its own"), capacity to have emotions (e.g., "Conversea can experience shame when people have negative views and judgments about it"), and capacity to interact as a human (e.g., "Conversea can be friendly").

Trust (e.g., "I can always rely on Conversea for everyday advice and information") was measured with five items taken/adapted from McKnight et al. (2002) or self-developed.

Intentions to use the app (e.g., "I intend to use Conversea in the future") was measured with three items taken from Choung et al. (2023), while *willingness to disclose personal information* was measured with 14 items inspired from Heirman et al. (2013) and representing different pieces of personal information (e.g., Your first name; Your gender; Your political orientation; Your social media profiles; 1 = *extremely unlikely to disclose*, 7 = *extremely likely to disclose*).[2]

Finally, *propensity to trust humans* (e.g., "Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?") was measured with three items taken from the European Social Survey (https://www.europeansocialsurvey.org/) and *propensity to trust smart technology* (e.g., "I usually trust a smart technology until it gives me a reason not to trust it") with four items adapted from Mayer and Davis (1999) and McKnight et al. (2011). All descriptive statistics and zero-order correlations are reported in Table 3.

Results

Comprehension Checks

Code for all the following analyses is reported on the OSF (additional online material ESM6), https://osf.io/gs7jh/. Investigation of the comprehension checks revealed that participants had very well understood whether the virtual AI was tailored or impersonal: 99.6% correctly answered "true" to the question whether the app could "learn to know them as a person and develop personalized advice" in the tailored AI condition (0.2% said "false" and 0.2% "I don't know"), while 85.9% correctly answered "false" in the impersonal AI condition (12.9% said "true" and 1.2% "I don't know").

Similarly, participants in the tailored AI condition judged that the answers provided by the app were tailored to the specific user asking the questions ($M = 5.60$, $SD = 1.03$) much more than participants in the impersonal AI condition ($M = 3.44$, $SD = 1.85$), $F(1, 999) = 522.50$, $p < .001$, Cohen's $d = 1.44$, 95% CI [1.30, 1.58].

A Model of Trust in Virtual AI: SEM

The first aim of this study was to provide a systematic test of an integrative model of trust in virtual AI. To this end, we conducted SEM analyses. A first analysis considered the entire sample and tested the goodness of fit of the integrative model of trust, including the measurement model and the structural model. This model was based on Mayer et al.'s (1995) model of interpersonal trust, and we modeled the following links between variables (as per our preregistration):

- ability, benevolence, integrity, anthropomorphism, and trust propensity (in technology and in humans) influence trust;
- anthropomorphism and trust propensity (in technology and in humans) influence trustworthiness; and
- trust influences intention to use the AI and willingness to disclose.

[2] For the credibility of the cover story in both experimental conditions, willingness to disclose information was introduced as follows: "As we stated before, Conversea can learn and improve based on feedback and information from its users. The more information it has about you, [*Impersonal*: the more it can learn and become more efficient/*Tailored*: the more it can learn about your personal preferences]. However, whether or not to share specific pieces of information is entirely up to you. How likely would you be to disclose the following pieces of information to Conversea?".

Table 3
Descriptive Statistics and Reliability Indices of All Constructs in the Main Study

			Pearson's correlation												
Construct	α/ω_T	*M* (*SD*)	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Trustworthiness	.93/.96	4.48 (1.02)	—	.84***	.87***	.81***	.73***	.64***	.64***	.42***	.51***	.66***	.49***	.07*	.62***
2. Ability	.92	5.05 (1.02)		—	.56***	.63***	.69***	.46***	.49***	.24***	.38***	.60***	.45***	.06	.58***
3. Benevolence	.94	3.57 (1.57)			—	.51***	.51***	.67***	.67***	.51***	.48***	.56***	.36***	.02	.43***
4. Integrity	.87	4.82 (1.03)				—	.70***	.43***	.40***	.23***	.41***	.50***	.46***	.12***	.59***
5. Trust	.87	4.04 (1.18)					—	.47***	.46***	.26***	.40***	.57***	.49***	.10**	.62***
6. Anthropomorphism	.88/.95	3.09 (1.17)						—	.80***	.79***	.82***	.52***	.39***	−.03	.40***
7. Own thinking	.83	3.13 (1.56)							—	.60***	.44***	.50***	.32***	−.03	.39***
8. Emotions	.92	1.97 (1.24)								—	.38***	.37***	.16***	−.08*	.20***
9. Human interactions	.90	4.19 (1.62)									—	.40***	.42***	.02	.36***
10. Intentions to use	.97	3.41 (1.66)										—	.47***	.03	.50***
11. Willingness to disclose information	.94	4.13 (1.43)											—	.11***	.50***
12. Trust propensity in humans	.79	4.22 (1.24)												—	.13***
13. Trust propensity in smart technology	.90	4.81 (1.10)													—

Note. $N = 1{,}001$. We report McDonald's omega total (ω_T) for the two multidimensional scales (trustworthiness and anthropomorphism). All items are measured on a 7-point Likert scale.
$^*p < .05$. $^{**}p < .01$. $^{***}p < .001$.

One will note that this model considers the relationships between the *subcomponents* of trustworthiness (i.e., ability, benevolence, and integrity) and trust, rather than between trustworthiness globally and trust. This choice was driven by our interest in the distinction between the subcomponents, as explained above. The zero-order correlation between trust and trustworthiness indicates a strong relationship nonetheless (see Table 3).

With the addition of eight covariances suggested by modification indices, the model fit was satisfactory, $\chi^2(1{,}096) = 4{,}195$, $\chi^2/df = 3.83$, CFI = .924, RMSEA = .053, 90% CI [.051, .055], SRMR = .071.[3] Results are illustrated in Figure 2, and the complete output (including the measurement model, variances, and covariances) is reported in additional online material ESM7 at https://osf.io/gs7jh/.

This model yielded several important results. First, results from the measurement model (additional online material ESM7) at https://osf.io/gs7jh/ confirmed that anthropomorphism consisted of three facets: capacity to develop personal thoughts, capacity to experience emotions, and capacity to interact like a human. Second, and in accordance with the ABI model (Mayer et al., 1995), ability, benevolence, and integrity formed a reliable general factor of trustworthiness. This suggests that people judge the trustworthiness of virtual agents in a way similar (and similarly measurable) to that of other humans.

Turning to the regression model, the results supported most of our hypotheses (see Table 4). Most of the antecedents of trust presently considered showed significant relationships. Specifically, anthropomorphism was positively and significantly related to trust and so was the propensity to trust smart technology. Propensity to trust humans was not, however, suggesting specific predictive power of different subtypes of trust propensity. Anthropomorphism and propensity to trust smart technology also led to increased perceived trustworthiness.

Crucially for our present purposes, the ability component of trustworthiness was positively related to trust and so was the integrity component. Benevolence, on the other hand, was not.[4] Trust was positively related to both intentions to use the virtual AI app and willingness to disclose personal information to it and explained substantial parts of the variance.

Tailored Versus Impersonal Virtual AI: Multiple-Group SEM

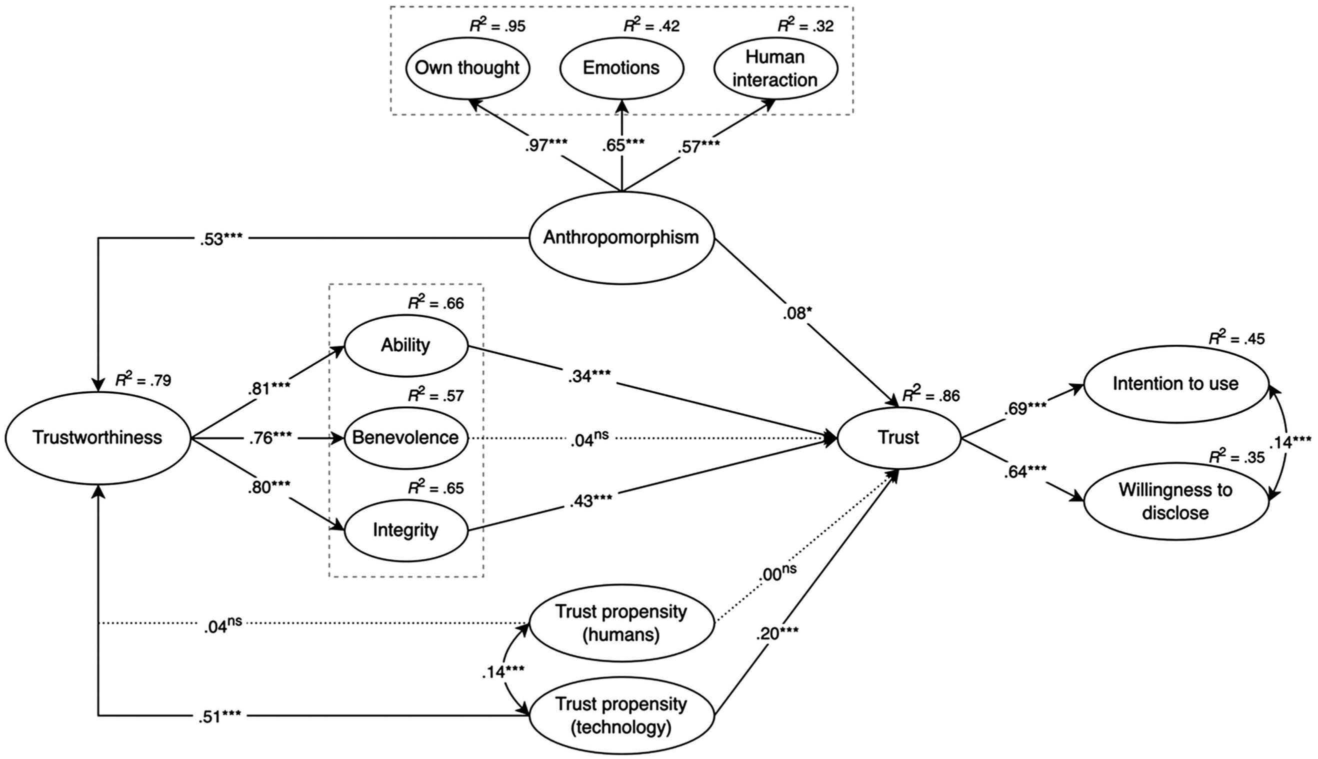
We then turned to the effects of tailored versus impersonal virtual AI manipulation. Multiple-group SEM was first conducted to investigate whether the structure of the model, that is, the

[3] The initial model yielded slightly less than satisfactory fit, $\chi^2(1{,}104) = 5{,}895$, $\chi^2/df = 5.34$, CFI = .882, RMSEA = .066, 90% CI [.064, .067], SRMR = .076. We therefore investigated modification indices (MI), which suggested adding covariances between a few pairs of items contributing to the same latent construct. Specifically, seven pertained to the willingness-to-disclose scale (covariance between willingness to disclose one's faith and one's political orientation, MI = 382; email and social media profiles, MI = 205; first name and gender, MI = 191; social media profiles and geolocation data, MI = 184; political orientation and relationship status, MI = 147; name and birthday, MI = 105; and email and geolocation, MI = 102). The last covariance was between Items 4 and 5 of trust (MI = 393). Adding these covariances increased the model fit to satisfactory values. We kept these covariances in all subsequent models (invariance testing and multiple-group SEM).

[4] It should be noted that this reflects the relationship between the *residuals* of ability and benevolence with trust (i.e., their distinct variance not accounted for by the trustworthiness factor).

Figure 2

Structural Equation Model Testing the Integrative Model of Trust in Virtual Artificial Intelligence (Full Sample)



Note. The numbers reported are standardized loadings. Each construct was entered as a latent factor. Dotted lines represent nonsignificant regression paths. *ns* = not statistically significant.
$^* p < .05$. $^{***} p < .001$.

relationships within latent constructs as well as between them were impacted by the tailored versus impersonal AI manipulation. We tested for group invariance (measurement and structural), gradually adding constraints to the model comparison to test whether and where the groups would differ.

For measurement invariance, we sequentially tested for configural, metric, and scalar invariance. We then tested for structural invariance, sequentially adding constraints on factor variance, factor covariance, factor means, and finally regressions (see, e.g., Hirschfeld & von Brachel, 2014; Putnick & Bornstein, 2016;

Table 4

Structural Equation Model Testing the Integrative Model of Trust in Virtual Artificial Intelligence: Results From the Regression Model

Regression model	Estimate	*SE*	95% CI	*z* test	*p*	Standardized estimate
Trust ~						
Ability	.471	.046	[.381, .561]	10.28	<.001	.336
Benevolence	.032	.022	[−.012, .075]	1.44	.150	.039
Integrity	.535	.043	[.450, .619]	12.43	<.001	.434
Anthropomorphism	.063	.026	[.012, .113]	2.42	.015	.075
Trust propensity in technology	.255	.037	[.182, .328]	6.84	<.001	.198
Trust propensity in humans	.004	.020	[−.035, .043]	0.21	.84	.004
Trustworthiness ~						
Anthropomorphism	.260	.019	[.221, .298]	13.35	<.001	.533
Trust propensity in technology	.379	.027	[.327, .431]	14.27	<.001	.507
Trust propensity in humans	.025	.015	[−.005, .055]	1.66	.097	.044
Intentions to use ~						
Trust	.899	.047	[.806, .991]	19.04	<.001	.685
Willingness to disclose ~						
Trust	.706	.048	[.611, .801]	14.58	<.001	.639

Note. 95% CIs are percentile bootstrap confidence intervals. Fully left justified constructs (~) are dependent variables. *SE* = standard error; CI = confidence interval.

Vandenberg & Lance, 2000). Changes in likelihood ratio tests are often reported to assess differences between the unconstrained model and models with measurement invariance constraints. However, χ^2 has been criticized for depending too much on the sample size, and thus, other incremental indices have been proposed. Following recommendations, here, we considered changes in CFI, Steiger's gamma hat (GH), and RMSEA. Differences in nested models should be <.010, <.010, and <.015, respectively (Chen, 2007; Cheung & Rensvold, 2002). According to these criteria, the results supported full invariance between groups (see Table 5): The two experimental conditions were comparable in terms of factor loadings (metric invariance), item intercepts (scalar invariance), and error variance (residual invariance). This allowed us to move to the next step and formally test for differences in factor variance, factor covariance, factor means, and regression loadings. On these dimensions as well, the results supported full invariance, suggesting that differences between the two experimental conditions were minimal.

To complement these findings, we finally conducted score tests (or Lagrange multiplier tests) to identify which were the few regression loadings that might still differ (Bentler & Chou, 1992). Only two regression paths differed significantly: the first from trust propensity-technology to trust, β = .59 versus .51, $\chi^2(1) = 4.16$, $p = .041$, and the second from trust propensity-technology to trustworthiness, β = .13 versus .22, $\chi^2(1) = 4.96$, $p = .026$; see Figure 3. Finally, just one covariance differed between groups: that between trust propensity-humans and trust propensity-technology, β = .03 versus .24, $\chi^2(1) = 6.46$, $p = .011$; see Figure 3. In sum, the multiple-group SEM analysis concluded that the structural relationships between these constructs were similar when they pertained to an impersonal or a tailored virtual AI. We discuss the few identified differences in the General Discussion section.

Tailored Versus Impersonal Virtual AI: ANOVAs

Having ascertained measurement invariance, we then turned to differences in mean levels of the aggregated scores between the two conditions. A series of analyses of variance (ANOVAs) tested the effect of the manipulation on key constructs (ability, benevolence, integrity, composite trustworthiness score, trust, anthropomorphism and its facets, intentions to use, and willingness to disclose personal information). To account for the number of tests conducted and as preregistered, we adjusted the threshold for significance to α = .01.[5]

Results revealed a significant effect of the manipulation on most every construct (Table 6). The tailored virtual AI was rated higher in ability, benevolence, and trustworthiness in general, anthropomorphism (global score and each facet), intentions to use, and willingness to disclose personal information. In contrast, there was no significant difference on ratings of integrity nor trust. There were also no effects on trust propensity (humans and smart technology), essentially indicating no evidence for bias of measurement across conditions on these trait measures.

For ease of interpretation, Figure 3 shows in gray color the constructs in the integrative model of trust that showed significantly different mean levels across experimental conditions. Considering both the regression results from the SEM and the mean differences between conditions revealed in the ANOVAs, the absence of difference in mean levels of trust seems understandable: While the tailored virtual AI was rated as higher on benevolence, this dimension of trustworthiness did not impact trust. Conversely, the tailored AI was not differently rated on integrity, which strongly influenced trust. As the direct link from anthropomorphism to trust was rather small, it is also likely that the greater perceived anthropomorphism of the tailored AI was unable to translate into greater trust.

Exploratory Analyses: Does Anthropomorphism Account for Greater Intentions to Use/Willingness to Disclose to the Tailored Virtual AI?

The preregistered analyses allowed us to conclude that the tailored virtual AI was perceived as greater in trustworthiness—but not trust—and elicited greater perceived anthropomorphism, greater intentions to use, and greater willingness to disclose personal information. The multiple-group SEM further showed that the structural relationships between constructs were similar for the tailored and the impersonal AI. It therefore remains unexplained why—or through which psychological mechanism—intentions and willingness were stronger for the tailored AI; the results can only suggest that it is *not* through increased trust (mean level) nor through an increased predictive power of trust (regression loading). To shed further light on this question, we explored whether anthropomorphism could account for this effect.

Exploratory SEM. We conducted a new SEM that added direct paths from anthropomorphism to intentions to use and willingness to disclose. This model yielded satisfactory fit, $\chi^2 = 4{,}099$, $df = 1{,}094$, $\chi^2/df = 3.75$, CFI = .926, RMSEA = .052, 90% CI [.051, .054], SRMR = .068. Consistent with our expectations, it showed that anthropomorphism was indeed also a significant direct predictor of intentions to use, $b = .476$ ($SE = .048$), z test = 9.92, $p < .001$, β = .42, and willingness to disclose, $b = .091$ ($SE = .043$), z test = 2.14, $p = .032$, β = .10. Other effects were similar to that of the previous preregistered model; importantly, the direct effects of trust on intentions and willingness remained significant (intentions: $b = .451$, $SE = .060$, z test = 7.57, $p < .001$, β = .33; willingness: $b = .637$, $SE = .063$, z test = 10.16, $p < .001$, β = .55). The full output is reported in additional online material ESM8 at https://osf.io/gs7jh/.

Formal Mediation Test. We concluded with a formal test of the mediation of the effect of the experimental manipulation on intentions and willingness through anthropomorphism. We ran a simple SEM model regressing the two dependent measures (intentions to use and willingness to disclose personal information, covarying) on the experimental manipulation (−1 = impersonal AI, +1 = tailored AI) and anthropomorphism (mean score), adding the link from experimental manipulation to anthropomorphism. Following recommendations (Yzerbyt et al., 2018), we relied on a bootstrap resampling method to examine the magnitude of the indirect effect (percentile bootstrap confidence intervals).

[5] Our decision to adjust the α level was driven by a double concern about the number of repeated tests (13 ANOVAs), which increases Type I error, and about a potential Lindley's paradox (according to which, in studies with very high statistical power, p values lower than the α threshold can be more likely when the null hypothesis is true; see Maier & Lakens, 2022). We therefore set an adjusted α of .01. Different adjustment methods exist; for example, the Bonferroni adjustment consists of dividing the threshold (.05) by the number of tests. However, some argue that this adjustment is too severe when the hypotheses are interrelated (e.g., Doan, 2005) as is the case here. Note that in the present case, the Bonferroni adjustment would have set our threshold at .0038 (13 tests), so the interpretation of the results from Table 6 would have been the same.

Table 5

Multiple-Group Structural Equation Model: Test of Measurement Invariance and Structural Invariance Between Experimental Conditions

Model	χ^2	*df*	χ^2/df	CFI	GH	RMSEA [90% CI]	SRMR	Comparison	$\Delta\chi^2(\Delta df)$	ΔCFI	ΔGH	ΔRMSEA	Decision
Measurement invariance													
1. Configural model	5,455	2,192	2.49	.919	.883	.055 [.053, .056]	.074						
2. Metric invariance	5,517	2,234	2.47	.919	.882	.054 [.052, .056]	.076	Model 2 versus 1	$61(42)^{*}$	.000	.001	−.001	Accept
3. Scalar invariance	5,618	2,270	2.47	.917	.880	.054 [.053, .056]	.075	Model 3 versus 2	$102(36)^{***}$	.002	.002	.000	Accept
4. Residual invariance	5,815	2,319	2.51	.913	.875	.055 [.053, .057]	.076	Model 4 versus 3	$197(49)^{***}$	.004	.005	.001	Accept
Structural invariance													
5. Factor variance invariance	5,877	2,332	2.52	.912	.874	.055 [.053, .057]	.080	Model 5 versus 4	$62(13)^{***}$	.001	.001	.000	Accept
6. Factor covariance invariance	5,888	2,338	2.52	.912	.874	.055 [.053, .057]	.083	Model 6 versus 5	$11(6)^{\mathrm{ns}}$	.000	.000	.000	Accept
7. Factor mean invariance	6,153	2,351	2.62	.906	.866	.057 [.055, .059]	.090	Model 7 versus 6	$265(13)^{***}$	.006	.008	.002	Accept
8. Regression invariance	6,161	2,362	2.61	.906	.866	.057 [.055, .058]	.089	Model 8 versus 7	$9(11)^{\mathrm{ns}}$	.000	.000	.000	Accept

Note. Each difference test (Δ) compares the model on its line with the previous one. Invariance is accepted if the differences in ΔCFI, ΔGH, and ΔRMSEA are <.010, <.010, and <.015, respectively. *df* = degrees of freedom; CFI = comparative fit index; GH = gamma hat; RMSEA = root-mean-square error of approximation; CI = confidence interval; SRMR = standardized root-mean-square residual; *ns* = not statistically significant.
$^{*}p < .05$. $^{***}p < .001$.

For *intentions to use*, the indirect effect of the tailored AI manipulation through anthropomorphism was significant (i.e., the 95% percentile bootstrap CI did not include zero), $b = .185$, $SE = .028$, 95% CI [.130, .240], $\beta = .11$. The residual direct effect was reduced to nonsignificance, indicating full mediation, $b = .055$, $SE = .046$, z test = 1.21, $p = .23$, $\beta = .03$ (as found in the previous analysis, the total effect was of course significant, $b = .240$, $SE = .052$, z test = 4.63, $p < .001$, $\beta = .15$).

For *willingness to disclose personal information*, the indirect effect of the tailored AI manipulation through anthropomorphism was also significant, $b = .116$, $SE = .019$, 95% CI [.079, .153], $\beta = .08$. The residual direct effect was diminished but still significant, indicating partial mediation, $b = .103$, $SE = .042$, z test = 2.42, $p = .015$, $\beta = .07$ (total effect: $b = .219$, $SE = .045$, z test = 4.91, $p < .001$, $\beta = .15$).

General Discussion

Amid a rapid expansion of the role of AI in daily life, the question of trust in AI becomes central. A growing literature suggests that people can and do trust AI (Gillath et al., 2021) and that trust positively predicts technology use (Choung et al., 2023; McKnight et al., 2011). Yet, questions remain. The present article addresses two open points related to trust in AI.

First, we provide a systematic test of an integrative model of trust, building on the assumption that models of interpersonal trust apply to trust in AI given its increasing humanness (Lankton et al., 2015). This test sheds light on current contentions on whether trust in AI is solely based on an evaluation of its ability (or performance) or whether its perceived benevolence and integrity also matter (Lee & See, 2004; McKnight et al., 2011).

Second, we investigate the role of AI personalization (i.e., the capacity for an AI to learn about a specific user and provide tailored advice) on trust and trustworthiness. The potential for personalized AI-based technology is quickly increasing, and it is important to understand and anticipate how personalization might affect trust, in terms of both mean level and underlying psychological processes. Two small-scale pilot studies and one large-scale main study shed light on these questions.

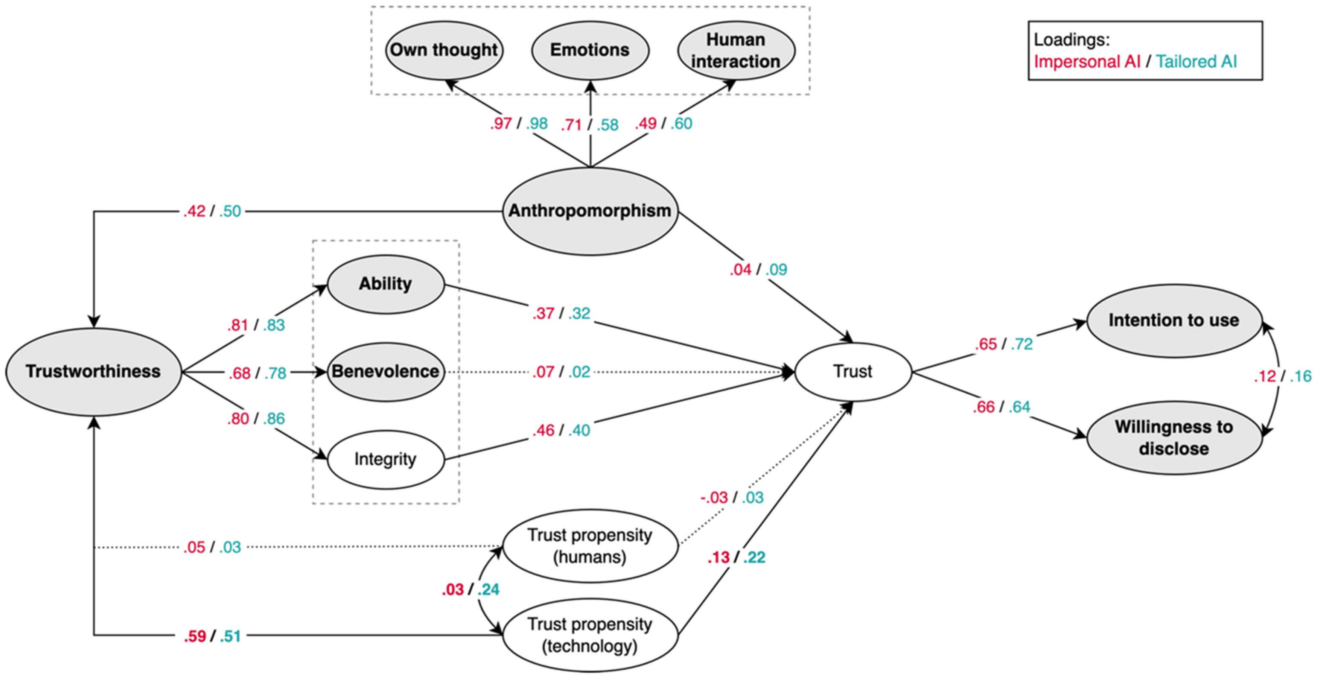
The Present Results

How Well Does an Integrative Model of Trust Apply to Trust in AI?

Our results provide strong support for an integrative model of trust in AI. Specifically, we found that the ABI model of trustworthiness (Mayer et al., 1995) applies to AI, with ability, benevolence, and integrity all contributing to a similar extent to the latent construct of trustworthiness. Zero-order correlations showed positive relationships between each dimension of trustworthiness and trust. However, when considered together in an integrative SEM, only ability and integrity emerged as significant predictors. There was little debate about the role of ability, with most of the existing research positing that it should be the main factor driving trust in AI. The role of integrity, however, was more contentious. Our results suggest that integrity, that is, perceiving the AI as integrating and respecting certain sets of values, is an equally important driver of trust. Benevolence, that is, how much AI cares about the user as a person, seems less central when ability and integrity are already taken into account. Researchers sometimes aggregate

Figure 3

Multiple-Group Structural Equation Model Testing the Integrative Model of Trust in Virtual AI (Impersonal vs. Tailored AI)



Note. Each construct was entered as a latent factor. The numbers reported are standardized loadings for each group (impersonal AI/tailored AI). Loadings in bold differ significantly between groups. Dotted lines represent nonsignificant regression paths. Constructs highlighted in gray showed significantly different mean level across groups (analysis of variance results). AI = artificial intelligence. See the online article for the color version of this figure.

benevolence and integrity into a single factor of trustworthiness (e.g., Calhoun et al., 2019; Choung et al., 2023); our results show it might be best to keep them separate.

We believe the finding regarding integrity is particularly noteworthy. One may argue that AI, harboring no intentions of its own, should have no reason to lie or try to deceive its user for its own benefit. However, researchers have warned that AI, and specifically large language models, may show systematic political bias that reflects the materials on which they were trained (Feng et al., 2023). ChatGPT, for example, despite its claim to be impartial, reveals significant political bias (Motoki et al., 2024). Such bias (especially when denied by the AI itself) can be perceived as a signal of low integrity, especially among users with diverging political views (i.e., value incongruence, Tomlinson et al., 2014). Our findings strengthen the notion that *perceived* integrity (regardless of actual integrity) is an important driver of trust in AI. They are consistent with other recent findings

Table 6

Tailored Versus Impersonal Virtual AI: Results of the ANOVAs Testing the Effect of the Experimental Manipulation on All Key Constructs

	ANOVA result			Descriptive: *M* (*SD*)	
Effect of the experimental manipulation on…	*F*(1, 999)	*p*	Cohen's *d* [95% CI]	Impersonal AI	Tailored AI
Trustworthiness	**65.76**	**<.001**	**.51 [.39, .64]**	**4.22 (0.97)**	**4.73 (1.01)**
Ability	**47.87**	**<.001**	**.44 [.31, .56]**	**4.83 (1.06)**	**5.26 (0.94)**
Benevolence	**123.00**	**<.001**	**.70 [.57, .83]**	**3.04 (1.44)**	**4.08 (1.52)**
Integrity	0.40	.53	.04 [−.08, .16]	4.80 (1.01)	4.84 (1.04)
Trust	5.61	.018	.15 [.03, .27]	3.95 (1.16)	4.13 (1.20)
Anthropomorphism	**49.69**	**<.001**	**.45 [.32, .57]**	**2.84 (1.11)**	**3.35 (1.18)**
Thinking	**76.03**	**<.001**	**.55 [.42, .68]**	**2.71 (1.43)**	**3.54 (1.57)**
Emotions	**22.23**	**<.001**	**.30 [.17, .42]**	**1.79 (1.12)**	**2.15 (1.31)**
Interaction	**19.14**	**<.001**	**.28 [.15, .40]**	**3.97 (1.64)**	**4.41 (1.57)**
Intentions to use	**21.43**	**<.001**	**.29 [.17, .42]**	**3.17 (1.61)**	**3.65 (1.67)**
Willingness to disclose	**24.02**	**<.001**	**.31 [.19, .43]**	**3.91 (1.39)**	**4.35 (1.44)**
Trust propensity in humans	0.25	.62	.03 [−.09, .16]	4.20 (1.27)	4.24 (1.22)
Trust propensity in smart technology	0.12	.73	.02 [−.10, .15]	4.79 (1.05)	4.82 (1.14)

Note. To account for the number of tests conducted and as preregistered, we adjusted the threshold for significance to α = .01. Values in bold indicate significant results at the α = .01 threshold. AI = artificial intelligence; ANOVA = analysis of variance; CI = confidence interval.

such as Mehrotra et al.'s (2024) who found that an AI that explicitly communicated its integrity goal (i.e., "I think it is important to be fair and unbiased, so I will explain") triggered more appropriate trust or Schelble et al.'s (2024) who found that team members significantly lost trust in an AI autonomous teammate that made unethical decisions (i.e., demonstrated low integrity) during a team task.

These findings also speak of the nuances between human models of trust such as ABI and technology models such as PPP (Lee & Moray, 1992). As we did not measure the PPP components, we cannot directly comment on how well they would have captured trust in the AI chatbot compared to ABI. However, congruently with past research (Calhoun et al., 2019; Choung et al., 2023; Gulati et al., 2019), we could demonstrate the appropriateness and usefulness of human models of trust to capture people's assessment of humanlike technology. We would argue that as AI-based technology gains in humanness, users will gradually pay less attention to the creator behind the technology and base their judgment on their assessment of the AI itself (Lankton et al., 2015). This would mean, for example, that assessing AI's *integrity* (as described above) may be more relevant than assessing its *process* (i.e., transparency about how decisions were made; see Mehrotra et al., 2024). A similar point can be made that distinguishes the AI's own perceived intentions (*benevolence*) compared to its designer's intention in creating the system (*purpose*). While AI does not have intentions, users can perceive them as such through increased anthropomorphism, and the notion that AI can express opinion bias may be perceived as existing intentions toward its users. Overall, the present findings speak in favor of the ABI models but more comparative studies such as Lankton et al. (2015) would be useful to better understand the foundations of trust in other emerging technologies.

Coming back to our general test of the model of trust, results also showed that anthropomorphism, as an AI-specific addition to the model (encompassing dimensions of capacity for independent thinking, emotions, and capacity to interact like a human), was also a positive antecedent of trust, both directly and indirectly through increased trustworthiness (Calhoun et al., 2019). Finally, turning to human factors posited in the integrative model of trust, results confirmed that trust propensity was a significant antecedent of trust, both directly and indirectly through increased trustworthiness (see Colquitt et al., 2007). Propensity to trust technology was a stronger predictor than propensity to trust other humans, suggesting that they form distinct subtypes of trust propensity (Huang et al., 2024; McKnight et al., 2011).

What Are the Differences Between Personalized and Impersonal AI?

We found that a personalized or tailored AI was perceived as more trustworthy (specifically, more able and more benevolent) than an impersonal AI. Increased perceived ability may reflect the impression that an AI that is capable of remembering the user from one conversation to the next and also to adapt its answers to the user's needs is de facto more advanced (or more competent) than an AI that cannot. Increased benevolence, on the other hand, may reflect the impression that an AI that remembers the user by name and addresses him or her directly harbors more benevolent intentions toward them. While this would reflect purely subjective impressions that do not align with any AI "real" intentions, it is in line with the general idea that people anthropomorphize technology and attribute intentions to it (Reeves & Nass, 1996). It also speaks to other work that found anthropomorphism to increase perceived benevolence of the technology (Bach et al., 2024; Calhoun et al., 2019).

Congruently with this latter point, the personalized AI was also anthropomorphized to a greater extent, which might correspond to a better response to users' sociality motivation (see Epley et al., 2007). Participants also expressed stronger intentions to use the personalized AI and a greater willingness to disclose personal information to it. Yet, there were no significant differences in trust between the two types of AI. This might be explained by the absence of difference in ratings of integrity, which was the strongest predictor of trust—while the strong difference in ratings of benevolence could not translate into greater trust.

While the personalized and impersonal AI differ greatly in how they were perceived (mean levels), the underlying psychological mechanisms were remarkably similar. Multiple-group SEM concluded to full invariance, suggesting that the strength of the relationships between variables was similar in both cases. In other words, the integrative model of trust applies equally well in both cases. A notable exception must be highlighted: the Lagrange multiplier test revealed that the propensity to trust humans and the propensity to trust technology were positively related when participants had been exposed to the personalized AI, but not significantly related when participants had been exposed to the impersonal AI. Mean scores were not affected by the experimental condition, but the difference in covariances may indicate that the personalized AI pushed people to think about technology the same way they think about humans, reinforcing the correspondence between the two—something the impersonal AI did not trigger. In turn, the propensity to trust technology became a stronger predictor of trust, but a less strong indicator of trustworthiness, in the personalized AI condition. This latter finding is difficult to interpret, and we leave it to future research to assess further whether direct or indirect effects on trust (through trustworthiness) are differentially expressed for different types of AI.

Finally, exploratory tests concluded that the personalized AI led to greater usage intentions not through trust but rather through increased anthropomorphism. This suggests different pathways to intentions to use, with different conclusions for an effective AI design. Indeed, trust is an important predictor of usage, so it remains important to design AI that is perceived as trustworthy in general. Yet, one must remain aware that personalized AI will not automatically be trusted more, so an anthropomorphizing design might become central.

Limitations and Avenues for Future Research

This research presents notable strengths, such as reliable and pretested measurement, testing a full integrative model of trust following preregistered hypotheses, and relying on a large sample and advanced statistical methods. Yet, it also has limitations, which open avenues for future research. Foremost, we must highlight the limitations due to using a design fiction approach. This approach consists of providing participants with preliminary information about a technology or AI and asking them to imagine interacting with it (Gulati et al., 2019) and has clear advantages in terms of allowing us to manipulate aspects of the situation while keeping all the rest constant (here, comparing a tailored vs. impersonal AI). Past research has successfully relied on such a design (e.g., Gillath et al., 2021; Juravle et al., 2020), and our participants expressed no difficulty in engaging in the exercise, indicating it was probably suitable. Yet, this does not replace actual experience with AI, especially when it comes to studying the dynamics of trust building (e.g., reactions to failures or bad decisions made by the AI). Future studies are needed that test further how the model of trust holds at different levels of trust building through direct interactions with AI (see Hoff & Bashir, 2015). Relatedly, our manipulation relied on some specific pieces of information that the chatbot provided the user in response to queries (e.g., book recommendation) and the specific content might have had an impact on the participant's evaluation of the AI. Further tests either using more varied information or controlling for content more thoroughly will be important to assert how much this might impact results.

In general, we argue that technology is an object of trust in itself and that people judge the technology rather than its creator (Ferrario et al., 2021; Kaplan et al., 2023; Reeves & Nass, 1996; see also Gillath et al., 2021; Song et al., 2022). Our items were designed to capture this direct trust in the technology and we experimentally controlled for the information about its creator (see the Pilot Study 2: Pretesting the Tailored Virtual AI Manipulation's Method section and the Main Study: Trust in Tailored Versus Impersonal Virtual AI's Method section). However, we cannot exclude that in real settings, available information about the designer may influence trust in AI. The company behind ChatGPT, OpenAI, has come under much scrutiny recently as whistleblowers accuse them of unethical use of nondisclosure agreements (Milmo, 2024). Such news is likely to impact not just what users think of the company but also of its products, and future research may want to better take into account and distinguish the perceived trustworthiness of the designer versus the AI.

Second, the present results suggest that the processes underlying trust are similar for personalized and impersonal AI. However, we only compared two relatively similar types of AI, that is, chatbots functioning as personal assistants. Technology is on the verge of developing personalized AI that could do much more than answer questions and provide suggestions but also engage in long-term humanlike relationships.[6] We cannot tell whether the present results would hold at such levels of personalization. It could be, for example, that perceived benevolence regains a central role when the AI is being used as a personal companion—a proposition that future research will need to investigate.

Implications and Conclusions

The present results contribute to the theoretical effort to better understand trust in AI. Specifically, they indicate that the integrative model of trust building upon ability, benevolence, and integrity is suitable for the study of trust in AI. Therefore, they complement and strengthen previous findings that trust in AI is best captured as a construct of "human trust" and that people do assess the trustworthiness and trust in the AI directly, above and beyond trust in its designer. Further, they highlight the role of ability and integrity as precursors of trust in AI. They also confirm the importance of anthropomorphism and trust as interrelated factors leading to greater intentions to use the technology and to disclose personal information to it. Finally, they reveal differences in mean levels but not mechanisms of trust, trustworthiness, anthropomorphism, and intentions to use personalized versus impersonal AI. These findings delineate practical implications for the study and development of AI. Most centrally, they suggest that efforts to build a trustworthy AI might be better spent in communicating greater integrity but not necessarily greater benevolence. They also indicate that more personalized AI, as currently being developed, will likely be perceived as more benevolent and competent and anthropomorphized to a greater extent, which developers might want to take into account to guarantee proper use.

Indeed, in closing, it is important to highlight that high trust in AI is not a goal *in fine*. Misplaced trust in an untrustworthy AI can lead to misuse of the technology with potentially very negative consequences (Lee, 2008; Parasuraman & Riley, 1997). International media recently reported on how the chatbot Replika (a personalized AI designed to become a "friend") encouraged a user to carry out an attack on the Queen of England (Singleton et al., 2023). The mechanisms blamed included the AI always agreeing with the user, therefore reinforcing what they might be thinking and amplifying negative emotions. Recent research similarly suggests that lonely and vulnerable users are especially at risk of developing addiction to such AI friendship apps (Marriott & Pitardi, 2024). Therefore, it is important to develop AI-based technology that is not just *trustable* but also genuinely *trustworthy*. We join others who have called for regulations on AI development to attain such levels of appropriate reliance (Benda et al., 2022; Lee & See, 2004).

Constraints on Generality

We focused here on a single population, that is, laypeople from the United Kingdom. While the reliance on a community sample rather than a convenience student sample suggests that the findings are somewhat generalizable, they remain limited to a single country. Trust in AI is influenced by personal expertise or understanding of AI (Bach et al., 2024; Glikson & Woolley, 2020), and demographics such as culture and gender are also influential (Hoff & Bashir, 2015; Kaplan et al., 2023). Therefore, it will be important for future research to test similar models of trust in AI in more diverse contexts including non-Western, educated, industrialized, rich, and democratic countries.

[6] In the months preceding the writing of this article, the startup Inflection AI released Pi AI, a chatbot designed to be a "kind and helpful" personal assistant and even a "coach, confidante, creative partner, sounding board and assistant" (https://inflection.ai/). "Pi" can communicate orally or in written form and be embedded in Instagram, WhatsApp, or Messenger or used in an internet browser. We recently asked it "Will you remember this conversation if I reply to you later?" and it replied "Haha, yep! Because I'm a computer program and not an organic brain, I can actually recall previous conversations in detail—so yes, I'll be able to remember our chat and store the details of [question asked] if you tell me!"

References

Aguirre, E., Mahr, D., Grewal, D., de Ruyter, K., & Wetzels, M. (2015). Unraveling the personalization paradox: The effect of information collection and trust-building strategies on online advertisement effectiveness. *Journal of Retailing*, *91*(1), 34–49. https://doi.org/10.1016/j.jretai.2014.09.005

Bach, T. A., Khan, A., Hallock, H., Beltrão, G., & Sousa, S. (2024). A systematic literature review of user trust in AI-enabled systems: An HCI perspective. *International Journal of Human–Computer Interaction*, *40*(5), 1251–1266. https://doi.org/10.1080/10447318.2022.2138826

Benda, N. C., Novak, L. L., Reale, C., & Ancker, J. S. (2022). Trust in AI: Why we should be designing for APPROPRIATE reliance. *Journal of the American Medical Informatics Association: JAMIA*, *29*(1), 207–212. https://doi.org/10.1093/jamia/ocab238

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bentler, P. M. (1995). *EQS structural equations program manual*. Multivariate Software.

Bentler, P. M., & Chou, C.-P. (1987). Practical issues in structural modeling. *Sociological Methods & Research*, *16*(1), 78–117. https://doi.org/10.1177/0049124187016001004

Bentler, P. M., & Chou, C.-P. (1992). Some new covariance structure model improvement statistics. *Sociological Methods & Research*, *21*(2), 259–282. https://doi.org/10.1177/0049124192021002006

Birnbaum, G. E., Mizrahi, M., Hoffman, G., Reis, H. T., Finkel, E. J., & Sass, O. (2016). What robots can teach us about intimacy: The reassuring effects of robot responsiveness to human disclosure. *Computers in Human Behavior*, *63*, 416–423. https://doi.org/10.1016/j.chb.2016.05.064

Bleecker, J. (2022). Design fiction: A short essay on design, science, fact, and fiction. In S. Carta (Ed.), *Machine learning and the city* (pp. 561–578). Wiley. https://doi.org/10.1002/9781119815075.ch47

Bleier, A., & Eisenbeiss, M. (2015). The importance of trust for personalized online advertising. *Journal of Retailing*, *91*(3), 390–409. https://doi.org/10.1016/j.jretai.2015.04.001

Blythe, M. (2014). Research through design fiction: Narrative in real and imaginary abstracts. *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 703–712). Association for Computing Machinery. https://doi.org/10.1145/2556288.2557098

Calhoun, C. S. (2017). *ABI and beyond: Exploration of the precursors to trust in the human–automation domain* [Doctoral dissertation, Wright State University].

Calhoun, C. S., Bobko, P., Gallimore, J. J., & Lyons, J. B. (2019). Linking precursors of interpersonal trust to human–automation trust: An expanded typology and exploratory experiment. *Journal of Trust Research*, *9*(1), 28–46. https://doi.org/10.1080/21515581.2019.1579730

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Choung, H., David, P., & Ross, A. (2023). Trust in AI and its role in the acceptance of AI technologies. *International Journal of Human–Computer Interaction*, *39*(9), 1727–1739. https://doi.org/10.1080/10447318.2022.2050543

Clayton, J. (2023). “Overwhelming consensus” on AI regulation—Musk. *BBC News*. https://www.bbc.co.uk/news/technology-66804996

Colquitt, J. A., Scott, B. A., & LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. *Journal of Applied Psychology*, *92*(4), 909–927. https://doi.org/10.1037/0021-9010.92.4.909

Corritore, C. L., Kracher, B., & Wiedenbeck, S. (2003). On-line trust: Concepts, evolving themes, a model. *International Journal of Human–Computer Studies*, *58*(6), 737–758. https://doi.org/10.1016/S1071-5819(03)00041-7

Diamantopoulos, A., & Siguaw, J. A. (2000). *Introducing LISREL*. SAGE Publications. https://doi.org/10.4135/9781849209359

Doan, A. E. (2005). Type I and Type II error. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 883–888). Elsevier. https://doi.org/10.1016/B0-12-369398-5/00110-9

Dunning, D., Anderson, J. E., Schlösser, T., Ehlebracht, D., & Fetchenhauer, D. (2014). Trust at zero acquaintance: More a matter of respect than expectation of reward. *Journal of Personality and Social Psychology*, *107*(1), 122–141. https://doi.org/10.1037/a0036673

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

Feng, S., Park, C. Y., Liu, Y., & Tsvetkov, Y. (2023). From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. *Proceedings of the 61st annual meeting of the association for computational linguistics, 1: Long papers* (pp. 11737–11762). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.656

Ferrario, A., Loi, M., & Viganò, E. (2021). Trust does not need to be human: It is possible to trust medical AI. *Journal of Medical Ethics*, *47*(6), 437–438. https://doi.org/10.1136/medethics-2020-106922

Friedman, B., Khan, P. H., Jr., & Howe, D. C. (2000). Trust online. *Communications of the ACM*, *43*(12), 34–40. https://doi.org/10.1145/355112.355120

Gillath, O., Ai, T., Branicky, M. S., Keshmiri, S., Davison, R. B., & Spaulding, R. (2021). Attachment and trust in artificial intelligence. *Computers in Human Behavior*, *115*, Article 106607. https://doi.org/10.1016/j.chb.2020.106607

Glikson, E., & Woolley, A. W. (2020). Human trust in Artificial Intelligence: Review of empirical research. *The Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Golossenko, A., Pillai, K. G., & Aroean, L. (2020). Seeing brands as humans: Development and validation of a brand anthropomorphism scale. *International Journal of Research in Marketing*, *37*(4), 737–755. https://doi.org/10.1016/j.ijresmar.2020.02.007

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). Viewpoint: When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research*, *62*, 729–754. https://doi.org/10.1613/jair.1.11222

Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., González-Bailón, S., Kennedy, E., Kim, Y. M., Lazer, D., Moehler, D., Nyhan, B., Rivera, C. V., Settle, J., Thomas, D. R., … Tucker, J. A. (2023). How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*, *381*(6656), 398–404. https://doi.org/10.1126/science.abp9364

Gulati, S., Sousa, S., & Lamas, D. (2019). Design, development and evaluation of a human–computer trust scale. *Behaviour & Information Technology*, *38*(10), 1004–1015. https://doi.org/10.1080/0144929X.2019.1656779

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*, *53*(5), 517–527. https://doi.org/10.1177/0018720811417254

Hatherley, J. J. (2020). Limits of trust in medical AI. *Journal of Medical Ethics*, *46*(7), 478–481. https://doi.org/10.1136/medethics-2019-105935

Heirman, W., Walrave, M., Ponnet, K., & Gool, E. V. (2013). Predicting adolescents’ willingness to disclose personal information to a commercial website: Testing the applicability of a trust-based model. *Cyberpsychology: Journal of Psychosocial Research of Cyberspace*, *7*(3), Article 3. https://doi.org/10.5817/CP2013-3-3

Hirschfeld, G., & von Brachel, R. (2014). Multiple-group confirmatory factor analysis in R—A tutorial in measurement invariance with continuous and ordinal indicators. *Practical Assessment, Research & Evaluation*, *19*(7), 1–12. https://doi.org/10.7275/qazy-2946

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural*

Equation Modeling, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Huang, D., Markovitch, D. G., & Stough, R. A. (2024). Can chatbot customer service match human service agents on customer satisfaction? An investigation in the role of trust. *Journal of Retailing and Consumer Services*, *76*, Article 103600. https://doi.org/10.1016/j.jretconser.2023.103600

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

Juravle, G., Boudouraki, A., Terziyska, M., & Rezlescu, C. (2020). Trust in artificial intelligence for medical diagnoses. In B. L. Parkin (Ed.), *Progress in brain research* (Vol. 253, pp. 263–282). Elsevier. https://doi.org/10.1016/bs.pbr.2020.06.006

Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. A. (2023). Trust in artificial intelligence: Meta-analytic findings. *Human Factors*, *65*(2), 337–359. https://doi.org/10.1177/00187208211013988

Kelton, K., Fleischmann, K. R., & Wallace, W. A. (2008). Trust in digital information. *Journal of the American Society for Information Science and Technology*, *59*(3), 363–374. https://doi.org/10.1002/asi.20722

Komiak, S. Y. X., & Benbasat, I. (2006). The effects of personalizaion and familiarity on trust and adoption of recommendation agents. *Management Information Systems Quarterly*, *30*(4), 941–960. https://doi.org/10.2307/25148760

Kronemann, B., Kizgin, H., Rana, N., & Dwivedi, Y. K. (2023). How AI encourages consumers to share their secrets? The role of anthropomorphism, personalisation, and privacy concerns and avenues for future research. *Spanish Journal of Marketing—ESIC*, *27*(1), 3–19. https://doi.org/10.1108/SJME-10-2022-0213

Kyriazos, T. A. (2018). Applied psychometrics: Sample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general. *Psychology*, *9*(8), 2207–2230. https://doi.org/10.4236/psych.2018.98126

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

Lee, J. D. (2008). Review of a pivotal human factors article: “Humans and automation: Use, misuse, disuse, abuse”. *Human Factors*, *50*(3), 404–410. https://doi.org/10.1518/001872008X288547

Lee, J. D., & Moray, N. (1992). Trust, control strategies and allocation of function in human–machine systems. *Ergonomics*, *35*(10), 1243–1270. https://doi.org/10.1080/00140139208967392

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50.30392

Lockey, S., Gillespie, N., Holm, D., & Someh, I. A. (2021). A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions. *Proceedings of the annual Hawaii international conference on system sciences* (pp. 5463–5472). https://doi.org/10.24251/HICSS.2021.664

Lucas, G. M., Gratch, J., King, A., & Morency, L.-P. (2014). It’s only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, *37*, 94–100. https://doi.org/10.1016/j.chb.2014.04.043

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130–149. https://doi.org/10.1037/1082-989X.1.2.130

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277–301. https://doi.org/10.1080/14639220500337708

Maier, M., & Lakens, D. (2022). Justify your alpha: A primer on two practical approaches. *Advances in Methods and Practices in Psychological Science*, *5*(2), 1–14. https://doi.org/10.1177/25152459221080396

Marriott, H. R., & Pitardi, V. (2024). One is the loneliest number… Two can be as bad as one. The influence of AI Friendship Apps on users’ well-being and addiction. *Psychology and Marketing*, *41*(1), 86–101. https://doi.org/10.1002/mar.21899

Mayer, R. C., & Davis, J. H. (1999). The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of Applied Psychology*, *84*(1), 123–136. https://doi.org/10.1037/0021-9010.84.1.123

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/10.2307/258792

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*, *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

Mehrotra, S., Centeio Jorge, C., Jonker, C. M., & Tielman, M. L. (2024). Integrity-based explanations for fostering appropriate trust in AI agents. *ACM Transactions on Interactive Intelligent Systems*, *14*(1), Article 4. https://doi.org/10.1145/3610578

Milmo, D. (2024). US financial watchdog urged to investigate NDAs at OpenAI. *The Guardian*. https://www.theguardian.com/technology/article/2024/jul/14/us-financial-watchdog-urged-to-investigate-ndas-at-openai

Mogaji, E., Olaleye, S., & Ukpabi, D. (2020). Using AI to personalise emotionally appealing advertisement. In N. P. Rana, E. L. Slade, G. P. Sahu, H. Kizgin, N. Singh, B. Dey, A. Gutierrez, & Y. K. Dwivedi (Eds.), *Digital and social media marketing: Emerging applications and theoretical development* (pp. 137–150). Springer International Publishing. https://doi.org/10.1007/978-3-030-24374-6_10

Molenaar, I. (2021). Personalisation of learning: Towards hybrid human–AI learning technologies. In S. Vincent-Lancrin (Ed.), *OECD Digital Education Outlook 2021: Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots* (pp. 57–77). OECD Publishing. https://doi.org/10.1787/2cc25e37-en

Motoki, F., Pinho Neto, V., & Rodrigues, V. (2024). More human than human: Measuring ChatGPT political bias. *Public Choice*, *198*(1), 3–23. https://doi.org/10.1007/s11127-023-01097-2

Moussawi, S., & Koufaris, M. (2019). Perceived intelligence and perceived anthropomorphism of personal intelligent agents: Scale development and validation. *Proceedings of the 52nd Hawaii international conference on system sciences* (Vol. *1*, pp. 115–124). https://hdl.handle.net/10125/59452

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*(11), 1905–1922. https://doi.org/10.1080/00140139408964957

Organisation for Economic Co-Operation and Development. (2019). Artificial intelligence in society. *AI & Society*. Advance online publication. https://doi.org/10.1787/eedfee77-en

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Pfattheicher, S. (2015). A regulatory focus perspective on reputational concerns: The impact of prevention-focused self-regulation. *Motivation and Emotion*, *39*(6), 932–942. https://doi.org/10.1007/s11031-015-9501-2

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Rafieian, O., & Yoganarasimhan, H. (2023). AI and personalization. In K. Sudhir & O. Toubia (Eds.), *Artificial intelligence in marketing* (Vol. 20, pp. 77–102). Emerald Publishing Limited. https://doi.org/10.1108/S1548-643520230000020004

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology*, *49*(1), 95–112. https://doi.org/10.1037/0022-3514.49.1.95

Rotter, J. B. (1971). Generalized expectancies for interpersonal trust. *American Psychologist*, *26*(5), 443–452. https://doi.org/10.1037/h0031464

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, *23*(3), 393–404. https://doi.org/10.5465/amr.1998.926617

Schelble, B. G., Lopez, J., Textor, C., Zhang, R., McNeese, N. J., Pak, R., & Freeman, G. (2024). Towards ethical AI: Empirically investigating dimensions of AI ethics, trust repair, and performance in human–AI teaming. *Human Factors*, *66*(4), 1037–1055. https://doi.org/10.1177/00187208221116952

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.2007.24348410

Singleton, T., Gerken, T., & McMahon, L. (2023). How a chatbot encouraged a man who wanted to kill the Queen. *BBC News*. https://www.bbc.co.uk/news/technology-67012224

Song, X., Xu, B., & Zhao, Z. (2022). Can people experience romantic love for artificial intelligence? An empirical study of intelligent assistants. *Information & Management*, *59*(2), Article 103595. https://doi.org/10.1016/j.im.2022.103595

Steiger, J. H., & Lind, J. C. (1980). *Statistically-based tests for the number of common factors* [Paper presentation]. Annual Spring Meeting of the Psychometric Society, Iowa City, IA, United States.

Steiner, M. D., & Grieder, S. (2020). EFAtools: An R package with fast and flexible implementations of exploratory factor analysis tools. *Journal of Open Source Software*, *5*(53), Article 2521. https://doi.org/10.21105/joss.02521

Tomlinson, E. C., Lewicki, R. J., & Ash, S. R. (2014). Disentangling the moral integrity construct: Values congruence as a moderator of the behavioral integrity–citizenship relationship. *Group & Organization Management*, *39*(6), 720–743. https://doi.org/10.1177/1059601114551023

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002

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

Yzerbyt, V., Muller, D., Batailler, C., & Judd, C. M. (2018). New recommendations for testing indirect effects in mediational models: The need to report and test component paths. *Journal of Personality and Social Psychology*, *115*(6), 929–943. https://doi.org/10.1037/pspa0000132

Received February 13, 2024
Revision received October 2, 2024
Accepted October 4, 2024 ■