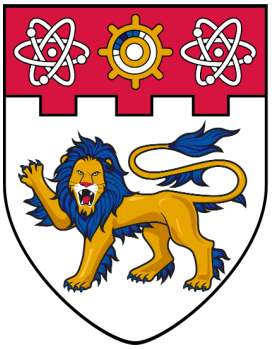


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

THREE ESSAYS ON TRUST IN ARTIFICIAL INTELLIGENCE IN THE ORGANIZATIONAL CONTEXT

SITONG YU

NANYANG BUSINESS SCHOOL

2025

THREE ESSAYS ON TRUST IN ARTIFICIAL INTELLIGENCE IN THE ORGANIZATIONAL CONTEXT

SITONG YU

NANYANG BUSINESS SCHOOL

A thesis submitted to the Nanyang Technological University in partial fulfilment of the requirement for the degree of Doctor of Philosophy

2025

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

21/01/2025

. . . . . .

Date

. . . . . .

Sitong YU

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement, with amendments, changes and improvements as suggested by me as the Supervisor. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

22/01/2025

. . . . . .

Date

. . . . . .

Kang Yang Trevor YU

Authorship Attribution Statement

This thesis contains material from 1 paper (Essay 1) accepted at conferences in which I am listed as an author.

Yu, S., Yu, K. Y. T, & Kawasaki, S. (2024). Trust in Artificial Intelligence (AI): An Integrative and Meta-Analytic Review. Navigating New Horizons: Management Research in the Greater Bay Area. Shenzhen, China.

Yu, S., Kawasaki, S., & Yu, K. Y. T. (2023). Trust in Artificial Intelligence (AI): An Integrative and Meta-Analytic Review. The 83rd Annual Meeting of the Academy of Management (AOM), Boston, MA, United States.

The contributions of the co-authors are as follows:

- Sitong Yu developed the model, collected and analyzed the data, and prepared the manuscript.
- Shota Kawasaki participated in data collection and edited the manuscript drafts.
- Assoc. Prof. Trevor Yu provided suggestions for this project and edited the manuscript drafts.

21/01/2025

. . . . . .

Date

. . . . . .

Sitong YU

Acknowledgements

"Thanks to" Covid-19, I feel that time has flown particularly fast during my PhD journey from 2020 until now. As I wrap up my thesis, the finish line of this journey is almost within reach. I would not have been able to reach this stage without the guidance and support of the following people:

First and foremost, my PhD advisor, Dr. Trevor Yu. I want to express my sincerest gratitude to you for welcoming me into this PhD program and kindly offering immense guidance and support on my research projects and dissertation over these years. I deeply appreciate the valuable opportunities to collaborate on various research projects, during which I gained knowledge as a PhD student and expertise and independence as a researcher.

To my thesis advisory committee members, Dr. Xi Zou and Dr. Filip Lievens. Many thanks to you for your insightful seminars and invaluable feedback on my thesis. Two fun facts: Xi was the first NBS faculty member I connected with at the start of my program (besides my advisor), and Filip kindly invited me to his HR seminar, which turned out to be my last PhD course. Thank you for making both the beginning and the end of this journey meaningful and fruitful. I sincerely appreciate the serendipity of our connection.

To my undergraduate advisor, Dr. Xin Qin, and his research team, *BUT*. From you, I fortunately discovered my interest in OB research, received foundational research training, and cultivated research mindsets (do not study *b*oring and *u*nimportant *t*hings!). Although I have been abroad for these years, I feel fortunate to remain connected to you, maintaining and even strengthening the friendship ties.

To my first class of students, thank you for offering me a wonderful teaching experience. This experience reminds me of the positive impact that all the splendid teachers in my life have had on me and reaffirms my passion for this profession.

To my parents, partner, and friends, thank you for your unconditional love, for your genuine care for my happiness over my accomplishments, and for showing me what it means to be kind, intelligent, and brave. I am grateful for the various perspectives that you brought to my knowledge, as well as the many possibilities in life that you have helped me envision. Many of you are not in academia, and some have even left it. Thank you, my academia friends, for sharing your academic milestones and downs with me. And thank you, my non-academia friends, for respecting, understanding, and supporting my work.

Finally, to myself. Thank you for investing time in knowing, accepting, and loving yourself, for the effort you've put into developing healthy interpersonal relationships, for your resilience during extremely challenging times, and for your persistence in pursuing this academic career.

Table of Contents

Summary

This thesis comprises three essays investigating trust in artificial intelligence (AI) in the organizational context. In Essay 1, I reviewed prior literature on trust in AI to develop a nomological network, outlining its antecedents and consequences. I then conducted a meta-analysis based on 104 empirical articles, 120 independent studies, and 592 effect sizes to quantitatively summarize the empirical findings relevant to trust in AI and theoretical relationships in the nomological network. I also explored potential moderators of those relationships, including trust measurement, AI embodiment, and study artifacts. In Essay 2 and Essay 3, I examine a critical antecedent of trust in AI identified from Essay 1 – AI transparency. I propose a multifaceted typology of AI transparency consisting of four key facets: AI information disclosure (i.e., the extent to which information about how and/or why AI reaches a particular decision/prediction is disclosed), accuracy (i.e., the extent to which the information delivered by AI is accurate), clarity (i.e., the extent to which the information delivered by AI is understandable for recipients, even with limited technical knowledge), and personalization (i.e., the extent to which information delivered by AI is tailored to receiver's unique characteristics or preferences). Two between-subject experiments examined how AI information disclosure and accuracy (Essay 2) and clarity and personalization (Essay 3) influenced the development of trust in AI during an AI-mediated career assessment and recommendation session.

Essay 1: Trust in Artificial Intelligence: A Narrative and Meta-Analytic Review

Introduction

The past few decades have witnessed the growing prevalence of artificial intelligence (AI) technologies (IBM, 2022) in daily life as well as business activities, such as customer service (Huang & Rust, 2018), medical diagnosis (Yokoi et al., 2021a), automated driving (Ajenaghughrure et al., 2020), and talent attraction, selection, performance management, and learning and development (Black & van Esch, 2020; Lacroux & Martin-Lacroux, 2022; Landers et al., 2023; Tambe et al., 2019). During these AI implementations, trust-building has emerged as a critical challenge. Recent reports revealed that AI was adopted in at least one business unit or function in half of the surveyed organizations (Maslej et al., 2023), yet many organizations have not taken steps to ensure AI is perceived as trustworthy (IBM, 2022), resulting in underutilization of AI-assisted data science by decision-makers.

Beyond its practical relevance, trust in AI also raises important theoretical questions. Trust has long been an important construct in understanding human-human interactions, and it has since evolved into a core concept in facilitating human interactions with technologies, such as computers, automation, and robots. Nowadays, the extension of trust theories into the AI context introduces novel considerations – how is trust in AI similar to or different from interpersonal trust and trust in traditional technologies? Are classic trust frameworks, such as Mayer et al.'s (1995) ability-benevolence-integrity framework, applicable in the AI context (e.g., Lalot & Bertram, 2025; Li & Bitterly, 2024)? This prompts a need to map the current landscape of trust in AI research.

Moreover, despite research findings that trust in AI directly influences whether individuals accept AI-generated information, follow AI-provided

suggestions, and benefit from the AI implementation (Hancock et al., 2011; Glikson & Woolley, 2020), there is a lack of holistic and systematic understanding of how trust in AI connects with other theoretical constructs (e.g., antecedents, consequences) in its nomological network. Researchers have also pinpointed ongoing controversies over the conceptualization of trust (J. D. Lee & See, 2004; McKnight et al., 2002; Rousseau et al., 1998). Due to its multidisciplinary nature, the literature on trust in AI has adopted diverse approaches, potentially creating ambiguity in interpreting prior findings and limiting theoretical progress.

To address these gaps, I comprehensively review current research on trust in AI in Essay 1 by conducting both a narrative review and a meta-analysis. The narrative review fulfills the *redirect* purpose (Cronin & George, 2023), organizing current domain knowledge about trust in AI (e.g., how it has been theoretically and empirically studied), developing a nomological network, and identifying avenues for future research.

The meta-analysis complements this effort by serving the *adjudication* purpose, quantitatively aggregating empirical findings for the theoretical relationships identified in the narrative review. As primary studies often lack sufficient power to detect or accurately estimate effect sizes (Lipsey & Wilson, 2001) or are susceptible to distortion by sampling error and other artifacts (Hunter & Schmidt, 2004), a meta-analytic approach provides more accurate estimates of effect sizes and enables the identification of moderators that help inform future research directions (Geyskens et al., 2008).

For the rest of Essay 1, I first present the theoretical background of this study, then present the results of the narrative review and meta-analysis, respectively. Essay 1 concludes with a discussion of contributions, limitations, and future directions.

Theoretical Background

Artificial Intelligence

Artificial intelligence (AI) has been studied across various disciplines (e.g., computer science, psychology, management) and applied in diverse domains, such as medical diagnosis systems (Alam & Mueller, 2021; Juravle et al., 2020), recommendation agents (Bigras et al., 2018; Shi et al., 2021), and virtual assistants (Hasan et al., 2021; Pitardi & Marriott, 2021). Researchers have proposed different definitions of AI based on the features they believe to be most critical to the concept. In general, AI refers to software-based technologies that, for a given set of human-defined objectives and based on available information, can mimic human decision-making process by learning, reasoning, and making predictions, recommendations, or decisions (Glikson and Woolley 2020; Kaplan et al., 2023; Qin et al., 2025)[1]. Ferràs-Hernández (2018) outlined four core capabilities of AI: (a) interacting with the environment by gathering information from external sources (e.g., natural language, other computer systems), (b) interpreting this information to recognize patterns, induce rules, or predict events; (c) generating outputs such as results, answers, or instructions for other systems; and (d) evaluating the outcomes of their actions and improving their decision processes over time.

AI thus differs from non-AI-driven forms of automations (e.g., traditional conveyor belts), robots (e.g., conventional warehouse robots), and algorithms (e.g., rule-based algorithms) that may execute predefined tasks but lack the capacity to adapt to new environments and improve autonomously through learning.

Trust Across Humans, Technologies, and Artificial Intelligence

[1] Generative AI technologies (GenAI) were not discussed in Essay 1, as the data collection for Essay 1 was completed prior to the widespread discourse on GenAI in academic research. For further discussions and comparisons among AI-related technologies, please refer to Appendix 1.

Trust in Humans

Trust is originally studied in the human-human interaction context, such as interpersonal trust (Rotter, 1967), social exchange (Blau, 1964; Luhmann, 1968), service relationship (Johnson & Grayson, 2005), and trust in organization relationships (Mayer et al., 1995; Mayer & Davis, 1999; McKnight et al., 1998). It is conceptualized in various forms, such as expectancy, beliefs, or behavioral intentions. For example, Rotter (1967) theorized trust as "an expectancy held by an individual or a group that the word, promise, verbal or written statement of another individual or group can be relied upon" (p. 651). Trust also represents individual willingness to be vulnerable to the actions of another party (Mayer et al., 1995), or to depend on the party based on its perceived characteristics (McKnight et al., 1998; Rousseau et al., 1998). These conceptualizations underscore the relevance of trust in situations characterized by risk, uncertainty, or vulnerability.

One of the most common conceptualizations of trust is proposed by Mayer et al. (1995, p. 4), defined as "the *willingness* of a party to be *vulnerable* to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party". This definition views trust as a unidimensional construct, but proposes a three-dimensional framework of trustworthiness (i.e., ABI) to predict trust. Specifically, they argue that a major portion of AI trustworthiness is explained by three factors – ability (i.e., group of skills, competencies, and characteristics that enable a party to influence within some specific domain), benevolence (i.e., the extent to which a trustee is believed to want to do good to the trustor, aside from an egocentric profit motive), and integrity (i.e., perception that the trustee adheres to a set of principles that the trustor finds acceptable; Mayer et al., 1995). Individuals form

such beliefs based on observations and evaluations of the abovementioned antecedents, which subsequently contribute to the perception of AI trustworthiness. Trust, in turn, informs risk-taking behaviors and subsequent outcomes.

Another well-known conceptualization is the distinction between cognition-based and affect-based trust proposed by McAllister (1995). Cognition-based trust refers to trust grounded in rational evaluation, where individuals rely on observable cues or direct information to assess another's trustworthiness based on their reliability, character, or competence. In contrast, affect-based trust is rooted in emotional bonds and interpersonal care, arising from emotional investments in the relationships, genuine concern for another's well-being and a belief in reciprocal sentiments.

In addition, Lewicki and Bunker (1996) provide a dynamic view of trust, in which trust develops and emerges over time. They argue that all trust relationships begin with calculus-based trust (grounded in potential gains and costs from transactions in the relationship), then develop to knowledge-based (knowing the other sufficiently to predict the other's behavior) and subsequently identification-based (mutual understanding with identification with the other).

In this dissertation, I employ the conceptualization of Mayer et al. (1995) given its theoretical and empirical relevance to trust in the AI context. This will be further elaborated in subsequent sections.

Trust in Technology

As technologies evolved, researchers have investigated trust between human and non-human entities and technologies, such as automation (Jian et al., 2000; J. D. Lee & See, 2004; Schaefer et al., 2016), e-Commerce (Gefen et al., 2003; McKnight et al., 2002), robot (Hancock et al., 2011), and algorithms (Logg et al., 2019).

McKnight et al. (2011) defined trust in technology as a *belief* that a specific technology will work in a functional, helpful and reliable way in a situation where negative consequences are possible. In such conceptualization, functionality, helpfulness, and reliability correspond to the ability, benevolence, and integrity elements in the ABI model. Lee and See (2004) proposed a relatively new approach to conceptualizing trust as an *attitude*. Their definition of trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (p. 51) has been widely adopted in empirical studies of trust in human-machine interactions (e.g., Ashoori & Weisz, 2019; Liu, 2021; Nasirian et al., 2017).

These conceptualizations of trust in AI put a heavier emphasis on the capability or functionality of technology as the basis of trust, with an implicit assumption that technology lacks agency or intrinsic motives. Nevertheless, the Computers Are Social Actors (CASA) paradigm proposes that individuals often treat computers and automated agents as social beings by applying social norms and stereotypes to them (Nass et al., 1997). There is also empirical and neurological evidence for similarities shared between trust in humans and technologies such as automation (Li et al., 2024; Lewandowsky et al., 2000; Hoff & Bashir, 2015).

Trust in AI

As with traditional technologies, there has been considerable debate about whether AI is capable of being trusted. Consistent with recent research, I identify AI as a capable referent of trust (Dirks & de Jong, 2022; Glikson & Woolley, 2020; Lalot & Bertram, 2025). AI is distinct from other traditional technologies (e.g., automation) in its capability of adaptation and self-learning, making it much closer than previous generations of technologies in interacting with people and serving social functions.

The context of human-AI interactions inherently embeds certain levels of risk and vulnerability for users due to the complex mechanisms and non-determined behaviors of AI, which creates a context for trust formation.

Empirical findings support such arguments – the integrative model of ABI, though originating from interpersonal trust research, is found to be applicable in the AI context and significantly predicted trust in AI and usage intentions (Lalot & Bertram, 2025). In addition, prior studies found that people treat AI as real social actors, displaying attachment and even romantic love for such technologies (e.g., Gillath et al., 2021; Song et al., 2022).

There has been a few review efforts in consolidating research on trust in AI, albeit with a predominant focus on its antecedents (Glikson & Woolley, 2020; Kaplan et al., 2023; Li et al., 2024). In a narrative review, Glikson and Woolley (2020) proposed a theoretical framework for antecedents of trust in AI, which consisted of six dimensions – anthropomorphism, reliability, tangibility, transparency, intimacy, and task characteristics. In a meta-analytic review, Kaplan et al. (2023) categorized the predictors of trust in AI based on their relevance to AI, human, or context, and summarized the empirical findings for each category. These two reviews revealed some common insights. First, a consolidated framework helps to better organize the antecedents of trust in AI. Second, certain factors may moderate the relationship between trust in AI and its antecedents. For example, both reviews suggested that the way AI is presented (e.g., robotic, virtual, embedded) may serve as a moderator.

In addition, there is a methodological concern regarding research on trust in AI. Due to the popularity of this research across fields and disciplines, researchers have adopted different conceptualizations, operationalizations, and measurements of trust in AI in empirical investigations. For example, trust has been conceptualized as

an expectation or belief (Mcknight et al., 2011; Rotter, 1967), attitude (J. D. Lee & See, 2004), or willingness to act (Mayer et al., 1995; McKnight et al., 1998; Rousseau et al., 1998). Different conceptualizations of trust have led to different operationalizations and measurements of trust and trust in AI: trusting belief (Jian et al., 2000; McAllister, 1995; McKnight et al., 2002), trusting intention (Juravle et al., 2020; Komiak & Benbasat, 2006), or trusting behavior (K. Gupta et al., 2019; Schmidt et al., 2020). For more rigorous research endeavors in the future, it therefore becomes theoretically important to organize domain knowledge and consolidate these methodological approaches of trust in AI.

Narrative Review

Based on the discussions above, I address three research questions in the narrative review: (1) How is trust in AI conceptualized, operationalized and measured in current research? Is there a consensus on it? (2) What are the antecedents and consequences of trust in AI?

Review Methodology

Given the interdisciplinary nature of research on trust in AI, I conducted a comprehensive literature search across disciplines (Harari et al., 2020), such as business, management, psychology, communication, computer science, and ergonomics. First, I searched articles on trust in AI published as of April 2023 in multiple databases – *Web of Science*, *EBSCO*, *JSTOR*, *PsychINFO*, *ProQuest*, *IEEEXPlore*, and *EngineeringVillage*. A list of AI-related keywords was generated based on our definition of AI: *artificial*, intelligen*, AI, machine learning, expert*

system, intelligent agent, and *intelligent automation*.[2] These terms were crossed with *trust* to identify the articles on trust in AI. This search process yielded 1191 articles. After removing duplicates, a total of 1107 articles were left for further screening. Following prior recommendations, a detailed outline of the screening process is presented in Figure 1 in terms of a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flowchart (Moher et al., 2009).

To determine whether articles are eligible for the narrative review, I excluded articles that (1) were not written in languages other than English (n = 3), did not discuss or examine trust in AI (n = 755), and did not discuss or examine trust in AI at the individual level (n = 113), leaving 236 articles.

Review Findings

This section presents the insights from the narrative review, organized by the two research questions. First, I examine how trust in AI has been conceptually and empirically studied in prior studies. Second, I develop a nomological framework of trust in AI (Figure 2), outlining its core constructs, antecedents, and consequences.

Conceptualizations and Operationalizations of Trust in AI

Table 1 summarizes the conceptualizations and operationalizations of trust and its related constructs (i.e., propensity to trust, trustworthiness) across interpersonal and AI contexts. Consistent with prior research (J. D. Lee & See, 2004; McKnight et al., 2002; Rousseau et al., 1998), our review uncovered the diverse focuses when conceptualizing and operationalizing trust in AI. In general, most conceptualizations or definitions of trust in AI borrow from existing measures from interpersonal trust

[2] An asterisk (*) is an identifier used to find the synonyms that start with the word preceding it (e.g., for intelligen*, the search includes the words such as intelligent and intelligence). For discussions of other AI-related technologies, please refer to Appendix 1.

research, albeit with slight adjustments that reflect the unique characteristics of AI, particularly its capability or competence.

A stream of research conceptualizes trust in AI as a *belief*, representing the trustor's belief that AI will display expected attributes or behaviors. For example, Shin (2021) defined trust as "the belief that a vendor's services and/or reported results are reliable and trustworthy, and that the vendor will fulfill obligations in an exchange relationship with the user" (Shin, 2021, p. 4; see also Shin & Park, 2019). These belief-based definitions are typically related to perceptions of trustworthiness, which is an antecedent of trust in AI. While some empirical studies rely on Mayer et al.'s (1995) ABI framework to develop three-dimensional scales to measure trust beliefs (e.g., Hu et al., 2021; Shi et al., 2021; Suen & Hung, 2023), it is worth noting that trustworthiness and trust beliefs are essentially distinct constructs. Compared to trustworthiness, which encompasses specific attributes of the trustee (e.g., competence in planning routes), trust beliefs represent a broader and more general belief in and expectation of the trustee (e.g., performs as expected).

Alternatively, Lee and See (2004) proposed a relatively new approach to conceptualizing trust as an *attitude*, defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (p. 51). This perspective has been widely adopted in studies of trust in human-machine interactions (e.g., Ashoori & Weisz, 2019; Liu, 2021; Nasirian et al., 2017). It describes the basis of trust as performance and individual goals and emphasizes the role of context in which trust is cultivated. Trust attitude is typically measured using single-item indicators such as "How much do you trust …?" on a Likert scale or a percentage score (e.g., Juravle et al., 2020; Ozanne et al., 2022; Yin et al., 2019), or multi-dimensional scales capturing different attitudinal facets, such as

cognitive and affective facets (e.g., Komiak & Benbasat, 2006; McAllister, 1995; McKnight et al., 2011).

Studies that conceptualize trust as *intentions* and *behaviors* often follow prior frameworks (e.g., willingness to act; reliance-related behavior; Mayer et al., 1995; Dietz & Den Hartog, 2006) with the operationalizations tailored to the task or context. These studies assess whether the trustor is willing or planning to accept the information or recommendations provided by AI, share task responsibility, or delegate the decision-making power to AI at a future time point (e.g., Juravle et al., 2020; Komiak & Benbasat, 2006; S. Y. Zhang et al., 2021). In more applied settings, trust is reflected in observable behaviors, such as reliance on AI's suggestions and decisions (Gobel et al., 2022; T. Kim & Song, 2021; Yin et al., 2019; G. Zhang et al., 2023), viewing and monitoring patterns (e.g., Gobel et al., 2022; Ozanne et al., 2022; Sharan & Romano, 2020), or trust in subsequent tasks (Yokoi & Nakayachi, 2019). Trust behaviors can also be captured with physiological metrics, such as head movement (M. Gupta & Sinha, 2022), visual focus (Manchon et al., 2021), and brain activity (Montag et al., 2023).

Essentially, these varied conceptualizations all tap into the construct of trust in AI. Nevertheless, I note the different focuses of these conceptualizations in the nomological network to fully reflect these nuances.

Antecedents of Trust in AI

To summarize the antecedents of trust in AI investigated in the included articles, I classify them into four first-order categories based on their relevance to AI (trustee), human (trustor), task (trust object), and the broader context. This categorization draws conceptual support from Schilke et al. (2021), who describe trust as "A trusts B with respect to issue X" and highlight three components of trust – the

trustor (the actor placing trust), the *trustee* (the target of trust), and the *trust object* (the domain or activity in which trust is placed). Despite its simplicity, such categorization is grounded in established work in literature on trust (e.g., Schilke et al., 2021) and trust in AI (e.g., Kaplan et al., 2023; Lalot & Bertram, 2025; Li et al., 2024). While it does not explicitly discuss the broader context, I extend it by adding *context* as a distinct category to capture other situational factors (e.g., social, organizational, physical) that shape the trust process (Fulmer & Gelfand, 2012; J. D. Lee & See, 2004; Johns, 2006).

It is worth noting that Johns (2006) considers task context to be one dimension of discrete context, alongside social (e.g., social influence, social density) and physical elements (e.g., the built environment or decoration). However, I treat task-related factors as analytically distinct from context-related factors in this nomological network. *Task-related factors* include features and characteristic of the immediate domain or activity in which AI is expected to perform (i.e., trust object; Schilke et al., 2021), whereas the *context-related factors* refers to other "situational opportunities and constraints that affect the occurrence and meaning of organizational behavior as well as functional relationships between variables" (Johns, 2006, p. 386). This aims to allow for a clearer alignment with the trust process model and facilitates differentiation between the object of trust and the conditions under which trust develops.

Given the large number of antecedents identified from the included studies, I draw on established theoretical frameworks concerning technology acceptance and adoption (e.g., the Unified Theory of Acceptance and Use of Technology, Venkatesh et al., 2003; the Information System Successful Model, DeLone & McLean, 1992; Glikson & Woolley, 2020) to inform the coding and categorization of antecedents.

Table 2 provides definitions, representative constructs from the dataset, and corresponding sources for each antecedent.

AI-Related Factors. Our included articles reveal a series of antecedents of trust in AI that are relevant to AI's design characteristics, its performance, and its interaction with human trustors. Design characteristics, such as tangibility and anthropomorphism, have received extensive research attention in the human-AI interaction (e.g., Glikson & Woolley, 2020; Lalot & Bertram, 2025), as they are "engineerable" and can be tailored to enhance individual acceptance of AI (Li et al., 2025). Tangibility captures the capability of AI to be perceived or touched (e.g., physical presence or embodiment), and anthropomorphism reflects the perception that AI is perceived to possess human qualities like feelings (Glikson & Woolley, 2020).

AI's performance reflects its competency or ability to perform the assigned task for achieving an individual's goals. Prior models of technology adoption (e.g., Unified Theory of Acceptance and Use of Technology, Venkatesh et al., 2003, Venkatesh, 2022, Blut et al., 2022; Information System Successful Model, DeLone & McLean, 1992) have emphasized the critical role of this factor in facilitating users' acceptance and adoption of technologies. AI performance is associated with an individual's perceptions that AI is easy to use (effort expectancy), displays consistent behaviors over time (reliability), produces accurate results and suggestions (accuracy), delivers satisfactory performance, and/or helps the individual improve performance (performance expectancy).

Another set of desired characteristics of AI concerns the interaction process. This includes the disclosure of information (transparency), quality of information, and quality of interaction. Transparency reflects the extent to which the underlying operating rules and inner logics of AI are apparent to the trustor (Glikson & Woolley,

2020). Such information can be hidden as part of the technological black box or disclosed via various approaches differing in content, visualization, and comprehensibility. Disclosed information also varies in quality in terms of richness, precision, and personalization. In addition to information, interaction with AI consists of various levels of interactivity and immediacy, potentially affecting the physical or psychological closeness between both parties.

Human-Related Factors. Past studies have identified a rich set of human-related factors that affect the development of trust in AI. First, a human trustor's past experiences and expertise related to AI, technology, or the task itself, as well as the ability to understand AI-generated information and suggestions, are critical as they help the trustor better evaluate the trustworthiness of AI and decide whether to trust it or not. Emotional states, traits, and demographics are also found to be associated with trust in AI. Traits that are particularly relevant in the context of AI include self-efficacy with AI or task, attitudes toward AI and traditional technologies, propensity to trust, and big five personalities. Basic demographics like age, gender, and education have been recognized by prior studies to affect the acceptance of technologies (e.g., Venkatesh, 2022; Venkatesh et al., 2003). Particularly, in organizational settings, the trustor's job characteristics (e.g., function, experience, authority, and autonomy at work) have the potential to affect trust in AI as well.

Task-Related Factors. Glikson and Woolley (2020) emphasized task characteristics as one of the six factors to consider in the development of trust in AI. Research in this area has mostly focused on the performer (e.g., AI vs. human), the domain or area of the task, and specific features of the task, such as risk, complexity, and uncertainty. These factors reflect the trustor's expectations of potential error costs, cognitive or emotional loads, and unpredictability arising from the use of AI.

Context-Related Factors. Our review revealed several factors that capture the features of the broader organizational or social context in which the task happens. In organizational settings, leadership (e.g., leadership styles), culture (e.g., collectivism), regulation and policies can be leveraged to facilitate employees' trust in AI. In more general settings, social influence (i.e., the extent to which an individual perceives that important others believe he/she should use AI) serves as an important antecedent.

Consequences of Trust in AI

Our included papers predominantly focused on exploring the factors that benefit or harm trust in AI, with relatively less attention paid to the consequences of trust in AI. The first downstream consequence of trusting AI refers to individual perceptions and attitudes about AI and its products. Choung et al. (2022) found that trust in AI voice assistants significantly influenced the intention to use AI by affecting perceived usefulness and participants' attitudes toward AI. Malhotra and Ramalingam (2023) revealed that trust in AI was positively related to perceived animacy and intelligence of the product using AI, although the relationships did not reach statistical significance. Trust in AI also affects individuals' privacy concerns during their interactions with AI, which subsequently affect their information disclosure behaviors (Shin et al., 2022). The second set of consequences is behavioral – capturing whether trustors decide to use AI and how good the outcomes of their AI usages are (Bucinca et al., 2021; Leichtmann et al., 2023). Finally, a distinct set of research focuses on marketing-related outcomes, investigating the effect of trust in AI on product promotion effectiveness (C.-Y. Wang et al., 2020) and consumer-brand relationship (Jham et al., 2023).

In many studies, trust behavior is investigated as the ultimate outcome of the theoretical models. Yet I would like to emphasize the value of understanding more

consequences of trust in AI. This is particularly relevant as human-AI interactions can happen multiple times, lasting for a certain period. Investigating how individuals update their perceptions and evaluations of AI as well as how trust evolves over time would be of critical theoretical and practical value.

Potential Moderators

AI embodiment, or type of AI representation, has been indicated by previous reviews as a potential factor affecting the relationships between trust in AI and its antecedents. Kaplan et al. (2023) identified four types of common uses of AI – chatbots, robots, automated vehicles, and nonembodied, plain algorithms – based on their meta-analysis data. Glikson and Woolley (2020) organized their review based on three types of AI embodiments – robotic, virtual, and embedded – and argued that the development of trust between humans and AI would be different across various AI embodiments. Specifically, robotic AI refers to AI-enabled robots, virtual AI refers to AI-enabled virtual agents in which "AI has no physical presence, but a distinguished identity" (e.g., chatbot, avatar), and embedded AI refers to those "without a visual representation or a distinguished identity" (e.g., plain algorithm; Glikson & Woolley, 2020, p. 637, 639). Their review findings indicated that antecedents may have different effects on trust across three types of AI. For example, the low reliability of AI harms trust, and such a decrease may be more salient for embedded AI than robotic ones.

As mentioned in the previous section, one potential factor that may affect the effect sizes is how trust is measured. Trust measurement can be categorized specifically to four categories – trust beliefs, attitudes, intentions, and behaviors. Another categorization is simply based on its objectiveness – whether the measure is subjective, such as self-rated scales, or objective, such as behavior and physiological

signals.

In addition to moderators with theoretical meaning, study artifacts, such as "the particulars of its participants, methods, treatments, context, and the like" (Lipsey, 2019) may also play a role in affecting the relationships found in empirical studies. Samples of study artifacts include study design (e.g., correlational, experimental) and study setting (e.g., offline, online).

Discussion

From the narrative review, I develop a nomological framework of trust in AI, outlining the antecedents and consequences that have been investigated by prior research efforts. Our review findings also echoed prior researchers' statements regarding the lack of consensus in conceptualizing and operationalizing trust in AI. I argue that reviewing research without considering such differences may be harmful by covering up the potential differential effects on the theoretical relationships. For example, Lee and See (2004) posited that studying trust in technology as intentions or behaviors may potentially confuse it with other factors (e.g., workload, self-confidence) in translating beliefs and attitudes into behaviors.

Given this, I attempt to address two research questions in the meta-analysis: (1) What are the average effect sizes of the empirical relationships in the nomological network? (2) What are some potential moderators that affect the effect size found in empirical studies? (3) Which part of the nomological network lacks empirical investigation?

Meta-Analytic Review

Review Methodology

Literature Search and Eligibility Criteria

I further screened the articles included in the narrative review for the meta-

analysis based on the following criteria (see also Figure 1). First, I excluded articles that were not empirical, such as conceptual papers, reviews, and editorials (n = 30), and a meta-analysis (n = 1). Second, I excluded articles that did not empirically investigate trust in AI as one of the independent or dependent variables (n = 32). Then, one article with duplicate data is excluded (n = 1). Finally, we only excluded articles that did not report sample size and/or effect sizes that could be converted into Pearson's correlation coefficient (n = 67). As a result, the selection criteria led to the inclusion of 104 articles, including 120 independent studies with 592 effect sizes. The included articles are listed in Appendix 2.

Coding Procedure

I worked with one coauthor in coding the included articles. First, we coded basic information from each article, such as sample size (N), study number (e.g., 2 for Study 2), names of independent and dependent variables, reliabilities of variables (if measured in scales), and effect sizes (e.g., Pearson's r). When Pearson's r was not reported, we coded and transformed other available statistics (e.g., t, F, β, p values) to make the necessary data transformation, as described in the next section. Second, we coded potential moderators of the relationships between trust in AI and other variables, including AI embodiment (robot, virtual, embedded), trust measurement (subjective vs objective; belief vs attitude vs intention vs behavior), study design (correlational vs experimental), and study setting (online vs offline). Third, we coded any sample characteristics provided in the articles, such as sample source, distribution of gender, age, ethnicity, and education among participants.

To ensure the reliability of the coding, both authors first coded 30% of the articles independently, and the inter-rater reliability indicated a sufficient level of agreement (average Cohen's kappa = .88). The authors then met to discuss the

rationale for their own coding and resolved coding discrepancies. After the authors agreed upon the best coding, the rest of the articles were split between them and coded individually.

In addition to the above coding information, I categorized each variable into corresponding categories (AI, human, task, context) and factors (e.g., transparency, AI-related experience, task complexity, social influence) in Table 2. The trend of included papers published per year is presented in Figure 3.

Analytic Strategy

I applied Hunter and Schmidt's (2004) approach for random-effects meta-analysis. Random effects models assume variability of true effect sizes across samples, and Hunter and Schmidt's approach attempts to minimize the variability that is caused by methodological and statistical artifacts, such as sampling error, measurement error, and range restriction (Aguinis et al., 2011). This approach has been widely adopted by management research over the years (e.g., Judge et al., 2002; Kim et al., 2025; Lyubykh et al., 2022; Wang et al., 2014).

This meta-analysis utilizes Pearson's *r* as the effect size index. For each meta-analytic relationship (i.e., between trust in AI and a certain antecedent or consequence), I calculated the total number of studies (k), total sample size (N), mean correlation (r), mean correlation corrected for sampling error and measurement error (ρ), and standard deviations of r and ρ (SD_r, SD_ρ) in R using the *psychmeta* and *metafor* packages (Dahlke & Wiernik, 2019; Viechtbauer, 2010).

For those studies that did not report Pearson's *r*, I coded other effect sizes reported by the article (e.g., t, F, β, p values) or statistics that enabled me to calculate effect sizes (e.g., mean and standard deviations in each experimental condition, used for calculating Cohen's d), whichever available. Then, I transformed these effect sizes

into Pearson's *r* using the formulae provided by Rosenthal and DiMatteo (2001). As indicated in Figure 1, sixty-seven articles were excluded from our final analysis because the transformation was inapplicable based on existing formulae.

I corrected measurement error using Cronbach's alpha coefficient reported by each study. When studies did not report reliability because the variables were 1) experimentally manipulated, 2) behavioral measures, or 3) measured with a single-item measure, I imputed the value of 1.00 for the variables' reliability. When the variables were measured in multi-item scales without Cronbach's alpha being reported in the articles, I imputed the average available reliability estimates from other included studies[3] (for a similar approach of imputing mean reliabilities, see Judge et al., 2002, Lyubykh et al., 2022, Zhong et al., 2024, and Zhong et al., 2025).

In addition, I computed several indices to examine the accuracy of and the variability around the corrected mean correlation (ρ). Confidence interval (CI) reflects the accuracy of the estimate of ρ, or "the likely amount of error in our estimate of ρ due to sampling error" (Hunter & Schmidt, 2004, p.205). A 95% CI excluding zero indicates statistical significance of estimated ρ at the 5% significance level.

On the other hand, credibility intervals (CV) reflect the distribution of ρ across different samples – a CV excluding zero indicates generalizability of the effect across all samples (Hunter & Schmidt, 2004). CV also indicate between-study heterogeneity, as wider CVs indicate the presence of moderators. Nevertheless, there lacks a consensus in the constitution of width (Geyskens et al., 2009; Lyubykh et al., 2022), thus I also computed I^2 index (i.e., the percentage of variation in effect sizes due to

[3] Reliability imputation was conducted within each category of trust measurement (imputed alpha = 0.88 for trust beliefs and attitudes, 0.90 for trust intentions). Sensitivity analysis showed no substantial differences for mean corrected effect sizes before and after reliability imputations, except for interaction quality, emotion, and propensity to trust. Detailed sensitivity analysis for reliability imputation is presented in Appendix 3.

true heterogeneity), %Var (i.e., the percentage of variance from artifacts), and the *Q*-statistic for each estimated ρ (Higgins & Thompson, 2002). The I^2 values of 75, 50, and 25 indicate high, medium, and low heterogeneity, respectively. A %Var lower than 75% or a significant *Q* suggest the presence of moderators.

Finally, to assess the potential moderation effects of the multi-categorical variables (i.e., AI embodiment, trust measurement, study design, and study setting), I follow best practices guidelines (Aguinis et al., 2011; Kim et al., 2024, 2025) to conduct both subgroup comparisons and meta-regression analyses (Lipsey & Wilson, 2001), only when there was a minimum of three effect sizes in each subgroup (Hoffman et al., 2007; Lyubykh et al., 2022). A primary heuristic for detecting moderation effects is through comparing the confidence intervals of ρ in each subgroup (Kim et al., 2024, 2025). Then, meta-regression analyses regress the moderators on the estimated mean effect sizes in a random-effect model to assess the statistical significance of the moderator.

Review Findings

Main Effects

The average effect sizes calculated for each antecedent or consequence of trust in AI were displayed in Table 3.1-3.2, Figure 5 and 6.

AI-Related Factors[4]**.** Among AI-related factors, *information quality* (i.e., quality of the information that AI produces) appeared to be the most correlated factor to trust in AI (ρ = .37, SD_ρ = .25, 95% CI [.26, .48], 80% CV [.04, .70]), followed by *anthropomorphism* (i.e., perception of AI as having human qualities; ρ = .32, SD_ρ = .25, 95% CI [.22, .42], 80% CV [.00, .64]), *interaction quality* (i.e., quality of

[4] Excluding the outliers reduces the effect size of anthropomorphism to 0.28 (Youn & Jin, 2021), transparency to 0.25 (Tuncer & Ramirez, 2022). See Appendix 4 for more information.

interacting with AI; ρ = .31, SD_ρ = .31, 95% CI [.10, .53], 80% CV [-.11, .74]), *transparency* (i.e., apparency of AI's underlying rules and logic; ρ = .26, SD_ρ = .20, 95% CI [.21, .31], 80% CV [.01, .51]) and *performance* (i.e., AI's competency to perform the assigned task; ρ = .24, SD_ρ = .25, 95% CI [.18, .30], 80% CV [-.09, .56]). These results suggest that improvement in AI quality, appearance and transparency are all associated with enhanced trust in AI.

Human-Related Factors[5]. Among individual differences, *attitude towards technology* is significantly more effective in cultivating trust in AI (ρ = .65, SD_ρ = .33, 95% CI [.37, .93], 80% CV [.18, 1.11], as its confidence interval did not overlap with most of other factors. People also displayed higher trust in AI when they hold more favorable *attitudes towards AI* (ρ = .25, SD_ρ = .24, 95% CI [.13, .36], 80% CV [-.07, .56]), have higher *propensity to trust* (ρ = .25, SD_ρ = .24, 95% CI [.09, .41], 80% CV [-.09, .58]), have more *experience related to AI* (ρ = .20, SD_ρ = .26, 95% CI [.09, .31], 80% CV [-.14, .54]), *technology* (ρ = .09, SD_ρ = .00, 95% CI [.06, .15]), or the *task* (ρ = .08, SD_ρ = .09, 95% CI [.00, .16], 80% CV [-.05, .21]), as well as when they are more *agreeable* (ρ = .10, SD_ρ = .00, 95% CI [.06, .15]) and *emotionally stable* (ρ = .11, SD_ρ = .00, 95% CI [.06, .17]). The corrected average effect sizes of ability to understand (ρ = .51, SD_ρ = .25), emotion (ρ = .34, SD_ρ = .41), extraversion (ρ = .11 , SD_ρ = .07), openness to experience (ρ = .04, SD_ρ = .23), gender (ρ = .04, SD_ρ = .08), and education (ρ = .01, SD_ρ = .12) were positive yet not statistically significant (i.e., CI includes zero). On the other hand, age (ρ = -.01, SD_ρ = .10) and conscientiousness (ρ = -.05, SD_ρ = .11) were negatively associated with trust in AI, though the correlations were not significant.

[5] Excluding the outliers reduces the effect size of emotion to 0.07 (Chi et al., 2023), gender to 0.01 (Lacroux & Martin-Lacroux, 2022), age to -0.05 (Oksanen et al., 2020); and increases the effect size of education to 0.03 (Xiang et al., 2022). See Appendix 4 for more information.

Task-Related Factors. Among task-related factors, only the identity of *task performer* (i.e., whether AI or human performs the task to be trusted) was found significantly correlated with trust in AI (ρ = .28, SD_ρ = .28, 95% CI [.18, .39], 80% CV [-.09, .65]). The effect size of *human-AI similarity* is positive yet not significant (ρ = .46, SD_ρ = .49). Additionally, *task complexity* (ρ = .04, SD_ρ = .02), *risk* (ρ = .03, SD_ρ = .20)[6], and *uncertainty* (ρ = .00, SD_ρ = .00) had slightly positive or even no correlation to trust in AI.

Context-Related Factors. *Social influence* was found to be critical in cultivating trust in AI (ρ = .57, SD_ρ = .10, 95% CI [.49, .65], 80% CV [.44, .71])[7]. Its effect size is significantly larger than all other antecedents, except attitude toward technology and interaction quality. Other context-related factors are not discussed here due to insufficient number of included studies (k < 3).

Trustworthiness. Among the three dimensions of trustworthiness, perceived ability has the largest correlation with trust in AI (ρ = .70, SD_ρ = .10, 95% CI [.61, .79], 80% CV [.56, .85]), followed by perceived benevolence (ρ = .27, SD_ρ = .09, 95% CI [.13, .41], 80% CV [.14, .40]) and integrity (ρ = .20, SD_ρ = .21, 95% CI [.06, .34], 80% CV [-.08, .49]).

Consequences of Trust in AI. For most of the consequences identified in the narrative review, only one or two effect sizes (k < 3) were found in the meta-analysis. One important finding to note is that trust in AI was negatively correlated to performance (ρ = - .38, SD_ρ = .06, 95% CI [-.50, -.26], 80% CV [-.48, -.28]). This will be further discussed in the Discussion section.

[6] Excluding the outliers (Juravle et al., 2020) reduces the effect size of risk to 0.02. See Appendix 4 for more information.
[7] Excluding the outlier (Carbone, 2020) reduces the effect size of social influence to 0.55. See Appendix 4 for more information.

Moderating Effects

As mentioned in the Analytic Strategy section, for each meta-analyzed relationship, I investigate potential moderating effects only when the I^2 value, %Var, and significance of Q-statistic suggest the presence of moderators that explain the variance in effect sizes across studies. Specifically, I explored whether the variability in the overall effect sizes exists across different categories of four moderators: trust measurement (belief, attitude, intention, behavior), AI embodiment (robotic, virtual, embedded), study design (correlational vs. experimental), and study setting (online vs. offline).

Following best practices recommendations (Kim et al., 2025; Aguinis et al., 2011; Steel et al., 2021), I conducted both subgroup analyses and meta-regression to examine the effects of categorical moderators. Subgroup analysis involved comparing the corrected mean effect sizes (ρ) across moderator conditions. In line with established decision rules (Kim et al., 2025; Hamann et al., 2023), I evaluated between-group differences based on (1) whether the 95% CIs around ρ across groups overlapped, and (2) whether the ρ estimate of one condition fell outside the CI of other conditions. These heuristics provide descriptive insights into potential differences across moderator levels. To ensure the robustness of findings, I conduct random-effects meta-regression analyses using the *metafor* package when each subgroup contained at least three effect sizes (Lipsey & Wilson, 2001). Key results from subgroup and meta-regression analyses are shown in Table 4.1-4.4 and Figure 7.

Moderating Effect of Trust Measurement. Meta-regression analyses did not show significant moderating effects of trust measurement at the 95% confidence level. Nevertheless, as shown in Table 4.1, there are meaningful differences in CIs of performance, age, gender, conscientiousness, and task performer (AI vs. human)

under different types of trust measurements.

AI performance was significantly more positively related to trust in AI measured as *attitude* (ρ = .28, SD_ρ = .30, 95% CI [.19, .38]) than as *behavior* (ρ = .14, SD_ρ = .08, 95% CI [.08, .19]). Although the CI of *intention* (ρ = .23, SD_ρ = .21, 95% CI [.04, .41]) overlapped with the other two, its ρ estimate fell outside the CI of *behavior*.

Older (vs. younger) people displayed less trust in AI measured as *attitude* (ρ = -.06, SD_ρ = .08, 95% CI [-.11, .00]), but more trust when measured as *behavior* (ρ = .13, SD_ρ = .06, 95% CI [.03, .24]). In a different pattern, females (vs. male) and people with higher (vs. low) conscientiousness showed more trust *attitude* towards AI (gender: ρ = .09, SD_ρ = .09, 95% CI [.01, .17]; conscientiousness: ρ = .04, SD_ρ = .05, 95% CI [-.08, .15]), while less trust *behaviors* (gender: ρ = -.03, SD_ρ = .00, 95% CI [-.09, .02]; conscientiousness: ρ = -.12, SD_ρ = .21, 95% CI [.04, .41]). Though their CIs overlapped, their ρ estimates fell out of each other's CI.

The effect of task performer (AI vs. human) is significantly larger with *behavioral* measures of trust (ρ = .47, SD_ρ = .25, 95% CI [.30, .65]) than with *attitudinal* ones (ρ = .07, SD_ρ = .14, 95% CI [-.02, .16]). Although the CI of *intention* (ρ = .07, SD_ρ = .24, 95% CI [-2.10, 2.24]) overlapped with the other two, its ρ *estimate* fell outside the CI of *behavior* measure of trust.

Moderating Effect of AI Embodiment. The effect size of attitude towards AI was significantly larger for robotic AI (ρ = .77, SD_ρ = .18, 95% CI [.30, 1.23]) than for embedded AI (ρ = .12, SD_ρ = .13, 95% CI [.01, .24]). The ρ estimate for virtual AI (ρ = .28, SD_ρ = .26, 95% CI [.06, .50]) was not included in the CI for both robotic and embedded AI. In addition, meta-regression analysis showed significant between-group differences (F = 13.09, p < .001).

Similarly, the effect of AI-related experience was more positive for virtual AI (ρ = .39, SD_ρ = .29, 95% CI [.09, .70]) than for embedded ones (ρ = .09, SD_ρ = .18, 95% CI [-.01, .19]; F = 9.06, p < .01), despite the two CIs overlapped. Older (vs. younger) people placed more trust in virtual AI (ρ = .15, SD_ρ = .12, 95% CI [-.17, .47]) and less trust in embedded AI (ρ = -.06, SD_ρ = .05, 95% CI [-.11, -.01]; F = 6.14, p < .05).

Moderating Effect of Study Design. The effect sizes of AI performance significantly varied in different study designs – larger in *correlational* studies (ρ = .36, SD_ρ = .32, 95% CI [.25, .47]) than in *experimental* studies (ρ = .15, SD_ρ = .13, 95% CI [.09, .20]). The effects of interaction quality, transparency, and AI-related experience were also more positive in *correlational* (interaction quality: ρ = .42, SD_ρ = .32, 95% CI [.08, .77]; transparency: ρ = .47, SD_ρ = .27, 95% CI [.12, .81]; AI-related experience: ρ = .47, SD_ρ = .24, 95% CI [.16, .77]) than *experimental* studies (interaction quality: ρ = .09, SD_ρ = .10, 95% CI [-.08, .26], F = 5.63, p < .05; transparency: ρ = .25, SD_ρ = .19, 95% CI [.20, .30]; AI-related experience: ρ = .25 SD_ρ = .19, 95% CI [.20, .30], F = 17.97, p < .001), although the two CIs overlapped.

For *task risk*, however, the effect was slightly negative in *correlational* studies (ρ = - .03, SD_ρ = .37, 95% CI [-.34, .29]) while slightly positive in *experimental* studies (ρ = .04, SD_ρ = .15, 95% CI [-.08, .16], F = 4.76, p < .05).

Moderating effect of study setting. Last but not least, the effect sizes of AI performance, information quality, and task risk varied in online and offline settings. Task risk had a significantly more positive correlation with trust in AI in *offline* (ρ = .59, SD_ρ = .10, 95% CI [.24, .93]) than *online* settings (ρ = .02, SD_ρ = .18, 95% CI [-.09, .13]; F = 8.03, p < .05). The effect of information quality and AI performance were also more positive in offline (info quality: ρ = .64, SD_ρ = .10, 95% CI [.43, .84];

performance: ρ = .40, SD_ρ = .19, 95% CI [.20, .60]) than online settings (info quality: ρ = .34, SD_ρ = .25, 95% CI [.21, .47]; performance: ρ = .21, SD_ρ = .25, 95% CI [.14, .28]). Despite that the CIs of the two groups overlapped, the ρ estimate of offline studies was not included in the CIs of online studies.

Supplementary Analysis

The arguments of the Theory of Planned Behavior (Ajzen, 1991) depicts how individual behaviors are developed: beliefs form the basis of attitudes (i.e., the degree to which a person has a favorable or unfavorable evaluation or appraisal of the behavior; p. 188), attitudes serve as one of the predicting factors of behavioral intentions, which in turn expressed in actual behaviors when the situation permits. This theoretical perspective suggests the possibility that trustworthiness belief, trust attitude, trust intention, and trust behavior can be categorized as distinct stages in this development process. Some included studies measured them as different constructs and reported effect sizes for their correlations. Wherever data permitted, I calculated the average effect size of the relationships between these stages. Consistent with the Theory of Planned Behavior, these four factors were positively associated with each other. All three dimensions of trustworthiness were positively associated with trust attitudes (ability: ρ = .70, SD_ρ = .10, 95% CI [.61, .79], 80% CV [.56, .85]; benevolence: ρ = .27, SD_ρ = .09, 95% CI [.13, .41], 80% CV [.14, .40]; integrity: ρ = .20, SD_ρ = .21, 95% CI [.06, .34], 80% CV [-.08, .49]). There were also significant and positive correlations between trust attitude and trust intention (ρ = .40, SD_ρ = .43, 95% CI [.16, .63], 80% CV [-.19, .98]) and between trust attitude and trust behavior (ρ = .26, SD_ρ = .25, 95% CI [.04, .48], 80% CV [-.09, .61]).

Robustness Check

To ensure the robustness of our results, I perform several analyses to assess

potential publication bias: (a) examination of the funnel plot, (b) Egger's test of asymmetry, (c) trim and fill analysis, and (d) Rosenthal's fail-safe *N* [8]. First, the effect sizes of included studies were plotted against their standard errors in a funnel plot (see Figure 4). Results of Egger's test indicate that the funnel plot displayed a symmetrical shape (z = - .17, p = .086). Trim-and-fill analyses (Duval & Tweedie, 2000) also indicate that no missing studies are needed to make the funnel plot symmetrical (estimated number of missing studies = 0, SE = 12.89). Furthermore, I followed Hunter and Schmidt (2004) to compute fail-safe *N* using the Rosenthal approach, finding that 1,374,366 (p < .0001) unpublished studies with null effect sizes will be needed in order to make our estimated effect size non-significant. These results all suggest that this meta-analysis is unlikely to be susceptible to potential publication bias (also see Appendix 3).

I also examined whether outliers affected the meta-analytic results by obtaining influence diagnostics – externally studentized residuals (*rstudent*) and Cook's distance (Viechtbauer & Cheung, 2010) for each study in each meta-analyzed relationship. The *rstudent* values indicate how much a study's effect size deviates from the estimated ρ, while Cook's distance provides an overall measure of the study's influence on the estimate. Larger Cook's distance values (commonly > .50) suggest that omitting the study would substantially affect the estimate. These analyses resulted in the detection of seven outliers, one each for anthropomorphism, transparency, emotion, age, education, task risk, and social influence. A summary of the identified outliers, the relationships they influenced, their *rstudent* values, Cook's distances, and the associated changes in effect sizes is presented in Appendix 4.

[8] In this section, only test results for the overall dataset are reported. Detailed test results for each meta-analyzed relationship are presented in Appendix 3.

Discussion

In this meta-analysis, I provided a quantitative summary of the empirical relationships between trust in AI and its antecedents and consequences. Most of the antecedents related to AI (trustee), human (trustor), task (trust object), and context were found to have a positive correlation with trust in AI. Particularly, attitude toward technology (ρ = .65), social influence (ρ = .57), AI information quality (ρ = .37), and task performer (AI vs. human, ρ = .28) were the most influential factors within each of the antecedent categories (i.e., human, context, AI, and task-related).

In addition, I identified four theoretical and methodological moderators that have the potential to influence the direction and/or magnitude of the relationships between trust in AI and its antecedents. Results from subgroup analyses and meta-regression suggest that antecedents may influence trust in AI to a different extent depending on the embodiment of AI, the way of measuring trust in AI, and study design and settings. The moderating effect of AI embodiment corresponds with the latest research findings that physical embodiment of AI (e.g., tangible robots vs. intangible algorithms) can help to increase AI appreciation (Qin et al., 2025; Glikson & Woolley, 2020). We also probed larger effect sizes for AI performance, gender, and conscientiousness when trust in AI on trust attitudes than on trust behaviors, but an opposite pattern for age and task performer. Such findings indicate potential inconsistencies in people's attitudes and actual behaviors when trusting AI and suggest future studies to choose and clarify their operationalizations of trust in AI.

Compared with the large number of studies exploring antecedents of trust in AI, very few studies in our review focused on its subsequent consequences. Studies from two included articles revealed a surprising negative correlation between trust in AI and task performance. Specifically, Bucinca et al. (2021) found that in an AI-

assisted task of designing nutrition meals, participants who trusted AI more showed higher reliance on incorrect model predictions and had worse task performance. In a similar classification task, Leichtmann and colleagues (2023) revealed that participants with higher trust in AI were significantly more likely to perform wrong classifications. Both studies, conducted in the human-AI collaboration context, suggest a concern of overtrust in AI, particularly when AI provides less accurate suggestions. However, such negative correlation should be interpreted with caution considering research context and design. A recent meta-analysis on human-AI collaboration effectiveness suggests that human-AI synergy depends on a series of factors, such as the type of task (e.g., finite decision-making, creation tasks), type of AI (e.g., deep, shallow), as well as the relative performance of the human and AI alone (Vaccaro et al., 2024).

General Discussion

In Essay 1, I conducted a narrative review of 236 articles and developed a nomological network for trust in AI for a better understanding of the current research landscape. I then conducted a meta-analytic review of 104 articles, providing a quantitative summary of the empirical relationships identified in the nomological network.

Theoretical Implications

Essay 1 contributes to the literature on trust in AI in several aspects. First, it provides a comprehensive synthesis of the research landscape by integrating both narrative and meta-analytic approaches. The narrative review delineates how research on trust in AI builds upon traditional trust theories (e.g., interpersonal trust) and systematically compares the conceptualizations and operationalizations of trust across human, technological, and AI contexts. In doing so, it consolidates diverse definitions

and measurements, echo prior observations regarding the lack of consensus in the field (McKnight et al., 2002; Rousseau et al., 1998). By empirically testing whether these inconsistencies influence reported effect sizes, the study shows that different operationalizations of trust in AI may partially account for the observed effects of some antecedents, such as AI performance, age, gender, and performer of the task.

Second, this essay contributes a nomological framework that maps out the antecedents and outcomes of trust in AI identified in prior studies. This framework extends earlier categorizations (e.g., Glikson & Woolley, 2020; Kaplan et al., 2023; Li et al., 2024) by offering a more comprehensive and systematically categorized list of antecedents. It also highlights a number of theoretically important yet empirically understudied factors, particularly those related to organizational and social contexts. By surfacing these gaps, the framework provides a clear and actionable research agenda for organizational scholars seeking to advance the study of trust in AI.

Third, this essay applies the Hunter and Schmidt (2004) random-effects model to generate more accurate estimates of effect sizes by correcting for both sampling and measurement error. This approach enhances the reliability of the meta-analytic findings, particularly in light of the wide adoption of self-reported trust measures that are susceptible to measurement errors. The meta-analytic findings not only provide a quantitative summary of the nomological network of trust in AI but also offer cumulative evidence that can serve as a foundation for future research.

Finally, the model enables the examination of theoretical and methodological moderators that have been largely overlooked in prior reviews (e.g., Kaplan et al., 2023). Through subgroup analyses and meta-regressions, the findings identify key contingencies that moderate the effects of antecedents on trust in AI. Notably, AI embodiment emerged as a significant moderator, which is consistent with Glikson and

Woolley's (2020) theorization. Additionally, the analyses revealed the potential moderating roles of trust measurement, study design and study setting, each pointing to valuable directions for future research inquiry.

Practical Implications

First, this essay presents a comprehensive map of the key antecedents and outcomes associated with trust in AI and offers a quantitative summary of effect sizes identified in prior empirical studies. Together, these findings provide actionable insights for AI designers, developers, and policymakers by highlighting which factors – such as information quality, social influence, and users' attitude toward technology – are potentially more influential in fostering trust in AI, and therefore warrant prioritization in AI design and implementation.

Second, the identification of key moderators underscores the critical role of context in shaping user trust. For instance, certain antecedents may exert stronger effects for robotic AI rather than embedded, intangible algorithms. When organizations adopt and implement AI tools and systems, it is thus essential to align such implementation with the AI's form of embodiment, intended function, and target user group. Organizations should also be cautious when generalizing trust in AI findings from one research setting (e.g., lab-based studies) to another (e.g., field studies).

Finally, in the organizational context specifically, the findings emphasize the importance of attending to employee attitudes, personalities, and prior experiences when introducing AI technologies. Organizations should invest in structured training programs to enhance employees' technological literacy and proactively foster positive attitudes toward technologies in advance of AI implementation.

Limitations

The narrative review and meta-analysis have several methodological limitations. First, the review scope is constrained by the time of data collection and sources that are utilized, as we did not include unpublished papers, papers in other languages, and papers published after April 2023. We acknowledge the importance of comprehensiveness and timeliness of review and conduct supplementary analyses to address concerns wherever possible. First, tests of funnel plot and Rosenthal's fail-safe *N* did not indicate potential publication bias. Next, following prior research (e.g., Qin et al., 2025), we conduct meta-regression analysis with publication year as the moderator to find that most of the mean effect sizes did not significantly vary across their time of publication (see Appendix 5). Given these results, we did not attempt to call for unpublished studies and gray literature that did not undergo the peer-review process, as including them may introduce additional concerns to the reliability of findings (Bi et al., 2025; Ferguson & Brannick, 2012).

Additionally, our inclusion of purely English-language articles may have introduced language bias and geographical limitations (Bi et al., 2025). Although the included studies had a diverse set of samples from non-English speaking countries, such as China, Japan, South Korea, and Germany, we remind researchers of the potential differences in psychological patterns towards AI from people in different cultures (e.g., Yam et al, 2023), and even differences in cultural tendencies of AI models themselves (e.g., Lu et al., 2025).

Another methodological limitation stems from the availability of reported sample sizes and effect sizes. For this reason, we are not able to test the contingencies of some antecedents, as the *k* of their effect sizes are too small to be included in subgroup analyses and meta-regressions. We therefore remind researchers to interpret estimates based on small number of samples with caution.

Second, the timeframe of our literature search constrained our ability to discuss the most recent technological advancements in AI technologies, as our data collection concluded prior to the rapid emergence of generative AI technologies (GenAI; e.g., ChatGPT) in academic research. Recent research has recognized the distinctions of GenAI from traditional AI technologies, noting that GenAI not only processes more complex inputs but also produces new text or visual outputs and performs creative tasks (B. C. Lee & Chung, 2024). These capabilities go beyond self-learning and performance optimization traditionally associated with AI. Future research should explore whether users exhibit different reactions to GenAI technologies (Dasborough, 2023) or rely on different factors to form their evaluation of AI trustworthiness and develop trust in such technologies.

Third, the categorization of different types of trust measures was empirically challenging, as some studies incorporated multiple types of trust measures in a single scale. For example, a scale may capture trust beliefs ("… seems to be a very reliable artificial agent"), while capturing trust intentions ("if I were to work with …, I can trust") at the same time. In future research, a more fine-grained approach to categorizing trust measurements is needed to more accurately investigate the effect of measurements on effect sizes obtained from empirical studies.

Fourth, although this Essay primarily adopts Mayer et al.'s (1995) conceptualization of trust and ABI framework of trustworthiness, there have been other approaches to theorizing trust. Particularly, Lewicki & Bunker (1996) took a transformational approach to studying trust. They emphasized the dynamic nature of trust beyond one-time transactional exchanges and proposed a three-stage process, where trust evolves from calculus-based trust to knowledge-based, and identification-based trust. Several empirical studies suggest that the strength of predictors on trust in

AI may vary across time or different stages of trust development (Alam & Mueller, 2021; Glikson & Woolley, 2020). Future research would benefit from taking the *time* or *process* element into consideration in theoretical models as well as research designs (for a longitudinal meta-analysis, see Bi et al., 2025).

Future Research Directions

A closer examination of the nomological network and the meta-analysis results (Figure 6) revealed several theoretical relationships that remain underexplored empirically.

Among the antecedents of trust in AI, tangibility has largely served as a research context rather than being studied directly alongside other antecedents such as transparency and performance, despite its theorized importance in Glikson and Woolley's (2020) review. This might be due to the practical difficulty of manipulating tangibility, as the embodiment of AI technologies or products is typically determined early in the development process. Although this essay was able to code AI embodiment as a moderator and assess its effect, future research is encouraged to explore the theoretical mechanisms explaining why people trust different forms of AI to varying degrees.

Our findings also showed individual differences in forming trust in AI. Specifically, characteristics such as age and conscientiousness were found to be negatively associated with trust in AI. This offers valuable opportunities for organizational researchers to investigate *why* certain employees are less inclined to develop trust in AI. For instance, Tang et al. (2022) argue that conscientious employees experienced a complementarity mismatch from AI usage, when AI's capability to autonomize decision-making overlaps with their preference for structure, order, and organization. Similarly, future research exploring the mechanism through

which older employees develop trust would be especially beneficial in the context of an increasingly aging workforce.

Research on task-related antecedents has primarily emphasized the roles of task performer (human vs. AI) and task risk, while devoting relatively less attention to other important task characteristics, such as uncertainty and complexity. Future studies should gather empirical evidence on these underexplored characteristics and identify additional task features that may shape trust in AI. For instance, Qin and colleagues (2025) propose a capability-personalization framework that categorizes tasks based on two dimensions – whether AI outperforms humans and whether personalization is necessary in the decision context. Similarly, Li and Bitterly (2024) suggest that the level of empathy required by a task may serve as a boundary condition for trust in AI.

Additionally, few studies included in this review have systematically examined the organizational, social, and even physical contexts in which trust in AI is formed. This presents a promising avenue for organizational researchers to explore how trust develops within workplace settings shaped by coworkers, teams, organizational practices and cultures, and broader social dynamics. For example, Erengin et al. (2024) found that an employee's trust in an AI teammate was significantly influenced by a trustworthy human teammate's trust perceptions. Other studies have revealed both benefits (e.g., inflated performance ratings; He et al., 2024) and concerns (e.g., diminished perceptions of competence and motivation; Reif et al., 2025) associated with AI usage from coworkers and supervisors. Whether and how such beliefs and social evaluations shape employees' trust trajectories warrants further empirical investigation.

Moreover, the meta-analytic findings identified AI embodiment as a critical

contingency for developers and AI-adopting organizations to consider. Beyond this, we encourage future research to examine the interplay between trustee, trustor, trust object and contextual factors. A more intergrated approach that considers how these elements interact will provide a deeper understanding of the mechanisms underlying trust in AI.

Finally, the workplace outcomes associated with trust in AI are of critical interest to organizations and managers, yet with limited empirical investigations. Kong and colleagues (2023) found that trust in AI was positively related to task performance, creative performance, and employee satisfaction with both life and career. However, AI adoption in the workplace is not always voluntary or initiated by employees. In contexts where AI implementation is mandated, it remains an open question whether employees who trust AI necessarily achieve better performance and experience greater well-being. Future research should investigate whether the effect of trust in AI persist under conditions of low control over AI use.

References

Aguinis, H., Gottfredson, R. K., & Wright, T. A. (2011). Best-practice recommendations for estimating interaction effects using meta-analysis. *Journal of Organizational Behavior*, *32*(8), 1033–1043.

Ajenaghughrure, I. B., da Costa Sousa, S. C., & Lamas, D. (2020, June). Risk and Trust in artificial intelligence technologies: A case study of Autonomous Vehicles. In *2020 13th international conference on human system interaction (HSI)* (pp. 118–123). IEEE.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211.

Alam, L., & Mueller, S. (2021). Examining the effect of explanation on satisfaction and trust in AI diagnostic systems. *BMC Medical Informatics & Decision Making*, *21*(1), 1–15. https://doi.org/10.1186/s12911-021-01542-6

Ashoori, M., & Weisz, J. D. (2019). In AI we trust? Factors that influence trustworthiness of AI-infused decision-making processes. *arXiv Preprint:1912.02675*.

Bi, S., Maes, M., Stevens, G. W. J. M., de Heer, C., Li, J.-B., Sun, Y., & Finkenauer, C. (2025). Trust and subjective well-being across the lifespan: A multilevel meta-analysis of cross-sectional and longitudinal associations. *Psychological Bulletin.* Advance online publication. https://doi.org/10.1037/bul0000480

Bigras, E., Jutras, M. A., Sénécal, S., Léger, P. M., Black, C., Robitaille, N., ... & Hudon, C. (2018). In AI we trust: characteristics influencing assortment planners' perceptions of AI based recommendation agents. In *HCI in Business, Government, and Organizations: 5th International Conference, HCIBGO 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15-20, 2018,*

Proceedings 5 (pp. 3-16). Springer International Publishing.

Black, J. S., & van Esch, P. (2020). AI-enabled recruiting: What is it and how should a manager use it? *Business Horizons*, *63*(2), 215–226. https://doi.org/10.1016/j.bushor.2019.12.001

Blau, P. (1964). *Power and exchange in social life*. New York: J Wiley & Sons.

Blut, M., Chong, A. Y. L., Tsigna, Z., & Venkatesh, V. (2022). Meta-analysis of the unified theory of acceptance and use of technology (UTAUT): Challenging its validity and charting a research agenda in the red ocean. *Journal of the Association for Information Systems*, *23*(1), 13–95.

Bucinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction, Volume 5, Issue CSCW1*, *5*, 1–21. https://doi.org/10.1145/3449287

Choung, H., David, P., & Ross, A. (2022). Trust in ai and its role in the acceptance of ai technologies. *International Journal of Human-Computer Interaction*. https://doi.org/10.1080/10447318.2022.2050543

Cronin, M. A., & George, E. (2023). The why and how of the integrative review. *Organizational Research Methods*, *26*(1), 168–192. https://doi.org/10.1177/1094428120935507

Dahlke, J. A., & Wiernik, B. M. (2019). Psychmeta: An R package for psychometric meta-analysis. *Applied Psychological Measurement, 43*(5), 415–416.

Dasborough, M. T. (2023). Awe-inspiring advancements in AI: The impact of ChatGPT on the field of Organizational Behavior. *Journal of Organizational Behavior*, *44*(2), 177–179. https://doi.org/10.1002/job.2695

DeLone, W. H., & McLean, E. R. (1992). Information systems success: The quest for

the dependent variable. *Information Systems Research*, *3*(1), 60–95.

Dietz, G., & Den Hartog, D. N. (2006). Measuring trust inside organisations. *Personnel Review, 35*(5), 557–588. https://doi.org/10.1108/00483480610682299

Dirks, K. T., & de Jong, B. (2022). Trust within the workplace: A review of two waves of research and a glimpse of the third. *Annual Review of Organizational Psychology and Organizational Behavior*, *9*(1), 247–276. https://doi.org/10.1146/annurev-orgpsych-012420-083025

Duval, S. J., & Tweedie, R. L. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*(2), 455–463.

Erengin, T., Briker, R., & de Jong, S. B. (2024). You, me, and the AI: The role of third-party human teammates for trust formation toward AI teammates. *Journal of Organizational Behavior*. https://doi.org/10.1002/job.2857

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods, 17*(1), 120–128. https://doi.org/10.1037/a0024445

Ferràs-Hernández, X. (2018). The future of management in a world of electronic brains. *Journal of Management Inquiry*, *27*(2), 260–263.

Fulmer, C. A., & Gelfand, M. J. (2012). At what level (and in whom) we trust: Trust across multiple organizational levels. *Journal of Management*, *38*(4), 1167–1230.

Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and TAM in online shopping: An integrated model. *MIS Quarterly*, 51–90.

Geyskens, I., Krishnan, R., Steenkamp, J. B. E. M., & Cunha, P. V. (2009). A review and evaluation of meta-analysis practices in management research. *Journal of Management, 35*(2), 393–419. https://doi.org/10.1177/0149206308328501

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660.

Gobel, K., Niessen, C., Seufert, S., & Schmid, U. (2022). Explanatory machine learning for justified trust in human-AI collaboration: Experiments on file deletion recommendations. *Frontiers in Artificial Intelligence*, *5*, 919534. https://doi.org/10.3389/frai.2022.919534

Gupta, K., Hajika, R., Pai, Y. S., Duenser, A., Lochner, M., & Billinghurst, M. (2019). In AI we trust: Investigating the relationship between biosignals, trust and cognitive load in VR. In Proceedings of the 25th ACM symposium on virtual reality software and technology (pp. 1-10).

Gupta, M., & Sinha, N. (2022). Wearable technology and women empowerment in the technology industry: An inductive-thematic analysis. *Journal of Information Technology Research*, *15*(1). https://doi.org/10.4018/JITR.299387

Hamann, M., Halw, O., & Guenther, T. W. 2023. Meta-analysis of the corporate planning–organizational performance relationship: A research note. *Strategic Management Journal*, *44*, 1803–1819.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517–527.

Harari, M. B., Parola, H. R., Hartwell, C. J., & Riegelman, A. (2020). Literature searches in systematic reviews and meta-analyses: A review, evaluation, and recommendations. *Journal of Vocational Behavior*, *118*, 103377.

https://doi.org/10.1016/j.jvb.2020.103377

Hasan, R., Shams, R., & Rahman, M. (2021). Consumer trust and perceived risk for voice-controlled artificial intelligence: The case of Siri. *Journal of Business Research*, *131*, 591–597. https://doi.org/10.1016/j.jbusres.2020.12.012

He, G., Yam, K. C., Zhao, P., Dong, X., Zheng, L., & Qin, X. (2025). Leaders inflate performance ratings for employees who use robots to augment their performance. *Human Resource Management*, *64*(2), 543–563.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

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

Hoffman, B. M., Papas, R. K., Chatkoff, D. K., & Kerns, R. D. (2007). Meta-analysis of psychological interventions for chronic low back pain. *Health Psychology, 26*(1), 1–9.

Hu, P., Lu, Y., & Gong, Y. (2021). Dual humanness and trust in conversational AI: A person-centered approach. *Computers in Human Behavior*, *119*, 106727. https://doi.org/10.1016/j.chb.2021.106727

Huang, M.-H., & Rust, R. T. (2018). Artificial Intelligence in Service. *Journal of Service Research*, *21*(2), 155–172. https://doi.org/10.1177/1094670517752459

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage.

IBM Global AI Adoption Index 2022. (2022, May). IBM. https://www.ibm.com/downloads/cas/GVAGA3JP

Jham, V., Malhotra, G., & Sehgal, N. (2023). Consumer-brand relationships with AI anthropomorphic assistant: Role of product usage barrier, psychological distance and trust. *International Review of Retail, Distribution & Consumer Research*, *33*(2), 117–133. https://doi.org/10.1080/09593969.2023.2178023

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.

Johns, G. (2006). The essential impact of context on organizational behavior. *Academy of Management Review*, *31*(2), 386–408.

Johnson, D., & Grayson, K. (2005). Cognitive and affective trust in service relationships. *Journal of Business Research*, *58*(4), 500–507.

Judge, T. A., Heller, D., & Mount, M. K. (2002). Five-factor model of personality and job satisfaction: a meta-analysis. *Journal of Applied Psychology, 87*(3), 530–541.

Juravle, G., Boudouraki, A., Terziyska, M., & Rezlescu, C. (2020). Trust in artificial intelligence for medical diagnoses. *Progress in Brain Research*, *253*, 263–282.

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

Kim, J., Chang, H., & Bell, B. S. (2025). Organizational-level training and performance: A meta-analytic investigation. *Journal of Management*, 01492063251327588.

Kim, J. (J.), Vaulont, M. J., Zhang, Z., & Byron, K. (2024). Looking inside the black box of gender differences in creativity: A dual-process model and meta-analysis. *Journal of Applied Psychology, 109*(12), 1861–1900. https://doi.org/10.1037/apl0001205

Kim, T., & Song, H. (2021). How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair. *Telematics & Informatics*, *61*, 101595. https://doi.org/10.1016/j.tele.2021.101595

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

Kong, H., Yin, Z., Baruch, Y., & Yuan, Y. (2023). The impact of trust in AI on career sustainability: The role of employee–AI collaboration and protean career orientation. *Journal of Vocational Behavior*, *146*, 103928. https://doi.org/10.1016/j.jvb.2023.103928

Lacroux, A., & Martin-Lacroux, C. (2022). Should I trust the Artificial Intelligence to recruit? Recruiters' perceptions and behavior when faced with algorithm-based recommendation systems during resume screening. *Frontiers in Psychology*, *13*, 895997. https://doi.org/10.3389/fpsyg.2022.895997

Lalot, F., & Bertram, A.-M. (2025). When the bot walks the talk: Investigating the foundations of trust in an artificial intelligence (AI) chatbot. *Journal of Experimental Psychology: General, 154*(2), 533–551. https://doi.org/10.1037/xge0001696

Landers, R. N., Auer, E. M., Dunk, L., Langer, M., & Tran, K. N. (2023). A simulation of the impacts of machine learning to combine psychometric employee selection system predictors on performance prediction, adverse impact, and number of dropped predictors. *Personnel Psychology*, *76*(4), 1037–1060. https://doi.org/10.1111/peps.12587

Lee, B. C., & Chung, J. (2024). An empirical investigation of the impact of ChatGPT

on creativity. *Nature Human Behaviour*, *8*(10), 1906–1914. https://doi.org/10.1038/s41562-024-01953-1

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

Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., & Mara, M. (2023). Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior*, *139*, 107539. https://doi.org/10.1016/j.chb.2022.107539

Lewandowsky, S., Mundy, M., & Tan, G. P. A. (2000). The dynamics of trust: Comparing humans to automation. *Journal of Experimental Psychology: Applied, 6*(2), 104–123. https://doi.org/10.1037/1076-898X.6.2.104

Lewicki, R. J., & Bunker, B. B. (1996). Developing and maintaining trust in work relationships. In R. Kramer & T. R. Tyler (Eds.), *Trust in organizations: Frontiers of theory and research* (pp. 114–139). Thousand Oaks, CA: Sage

Li, M., & Bitterly, T. B. (2024). How perceived lack of benevolence harms trust of artificial intelligence management. *Journal of Applied Psychology, 109*(11), 1794–1816. https://doi.org/10.1037/apl0001200

Li, Y., Wu, B., Huang, Y., & Luan, S. (2024). Developing trustworthy artificial intelligence: Insights from research on interpersonal, human-automation, and human-AI trust. *Frontiers in Psychology*, *15*, 1382693. https://doi.org/10.3389/fpsyg.2024.1382693

Lipsey, M. W. (2019). Identifying potentially interesting variables and analysis opportunities. In *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 141–152). Russel Sage Foundation.

Lipsey, M. W., & Wilson, D. B. (2001). Practical meta-analysis. SAGE Publications.

Liu, B. J. (2021). In AI we trust? Effects of agency locus and transparency on uncertainty reduction in human-AI interaction. *Journal of Computer-Mediated Communication*, *26*(6), 384–402. https://doi.org/10.1093/jcmc/zmab013

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Lu, J. G., Song, L. L., & Zhang, L. D. (2025). Cultural tendencies in generative AI. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-025-02242-1

Luhmann, N. (1968). Vertrauen, Ein mechanismus der reduction sozialer komplexität. Lucius & Lucius Verlagsgesellschaft, Stuttgart. Trad. It.(2002), La Fiducia. *Il Mulino, Bologna*.

Lyubykh, Z., Turner, N., Hershcovis, M. S., & Deng, C. (2022). A meta-analysis of leadership and workplace safety: Examining relative importance, contextual contingencies, and methodological moderators. *Journal of Applied Psychology*, *107*(12), 2149–2175. https://doi.org/10.1037/apl0000557

Malhotra, G., & Ramalingam, M. (2023). Perceived anthropomorphism and purchase intention using artificial intelligence technology: Examining the moderated effect of trust. *Journal of Enterprise Information Management*, *38*(2), 401–423. https://doi.org/10.1108/JEIM-09-2022-0316

Manchon, J. B., Bueno, M., & Navarro, J. (2021). Calibration of trust in automated driving: A matter of initial level of trust and automated driving style? *Human Factors*, *65*(8), 1613–1629. https://doi.org/10.1177/00187208211052804

Maslej, N., Fattorini, L., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Ngo, H., Niebles, J. C., Parli, V., Shoham, Y., Wald, R., Clark, J., &

Perrault, R. (2023). *The AI Index 2023 Annual Report*. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University.

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.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709–734.

McAllister, D. J. (1995). Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal*, *38*(1), 24–59.

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

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

McKnight, D. H., Cummings, L. L., & Chervany, N. L. (1998). Initial trust formation in new organizational relationships. *Academy of Management Review*, *23*(3), 473–490.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, *6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Montag, C., Klugah-Brown, B., Zhou, X., Wernicke, J., Liu, C., Kou, J., Chen, Y., Haas, B. W., & Becker, B. (2023). Trust toward humans and trust toward

artificial intelligence are not associated: Initial insights from self-report and neurostructural brain imaging. *Personality Neuroscience*, *6*, e3. https://doi.org/10.1017/pen.2022.5

Nasirian, F., Ahmadian, M., & Lee, O.-K. D. (2017). AI-based voice assistant systems: Evaluating from the interaction and trust perspectives. *23rd Americas Conference on Information Systems Proceedings*. https://aisel.aisnet.org/amcis2017/AdoptionIT/Presentations/27

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

Oksanen, A., Savela, N., Latikka, R., & Koivula, A. (2020). Trust toward robots and artificial intelligence: An experimental approach to human–technology interactions online. *Frontiers in Psychology*, *11*, 568256. https://doi.org/10.3389/fpsyg.2020.568256

Ozanne, M., Bhandari Aparajita, Bazarova, N. N., & DiFranzo Dominic. (2022). Shall AI moderators be made visible? Perception of accountability and trust in moderation systems on social media platforms. *Big Data & Society*, *9*(2), 20539517221115666. https://doi.org/10.1177/20539517221115666

Pitardi, V., & Marriott, H. R. (2021). Alexa, she's not human but... Unveiling the drivers of consumers' trust in voice-based artificial intelligence. *Psychology & Marketing*, *38*(4), 626–642. https://doi.org/10.1002/mar.21457

Qin, X., Zhou, X., Chen, C., Wu, D., Zhou, H., Dong, X., ... & Lu, J. G. (2025). AI aversion or appreciation? A capability–personalization framework and a meta-analytic review. *Psychological Bulletin*, *151*(5), 580–599.

Reif, J. A., Larrick, R. P., & Soll, J. B. (2025). Evidence of a social evaluation penalty for using AI. *Proceedings of the National Academy of Sciences*, *122*(19), e2426766122.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52*(1), 59–82.

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality*, *35*(4), 651–665. https://doi.org/10.1111/j.1467-6494.1967.tb01454.x

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

Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors*, *58*(3), 377–400.

Schilke, O., Reimann, M., & Cook, K. S. (2021). Trust in social relations. *Annual Review of Sociology*, *47*(1), 239–259. https://doi.org/10.1146/annurev-soc-082120-082850

Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*, *29*(4), 260–278.

Sharan, N. N., & Romano, D. M. (2020). The effects of personality and locus of control on trust in humans versus artificial intelligence. *Heliyon*, *6*(8), e04572. https://doi.org/10.1016/j.heliyon.2020.e04572

Shi, S., Gong, Y., & Gursoy, D. (2021). Antecedents of trust and adoption intention toward artificially intelligent recommendation systems in travel planning: A

heuristic–systematic model. *Journal of Travel Research*, *60*(8), 1714–1734. https://doi.org/10.1177/0047287520966395

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, *146*, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior*, *98*, 277–284.

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), 103595. https://doi.org/10.1016/j.im.2022.103595

Suen, H.-Y., & Hung, K.-E. (2023). Building trust in automatic video interviews using various AI interfaces: Tangibility, immediacy, and transparency. *Computers in Human Behavior*, *143*, 107713. https://doi.org/10.1016/j.chb.2023.107713

Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review*, *61*(4), 15–42. https://doi.org/10.1177/0008125619867910

Tang, P.M., Koopman, J., McClean, S. T., Zhang, J. H., Li, C. H., De Cremer, D., ... & Ng, C. T. S. (2022). When conscientious employees meet intelligent machines: An integrative approach inspired by complementarity theory and role theory. *Academy of Management Journal*, *65*(3), 1019–1054.

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, *8*, 2293–2303.

Venkatesh, V. (2022). Adoption and use of AI tools: A research agenda grounded in

UTAUT. *Annals of Operations Research*, *308*(1), 641–652.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425–478. https://doi.org/10.2307/30036540

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, *1*(2), 112–125. https://doi.org/10.1002/jrsm.11

Wang, C.-Y., Song, Y., Wu, C.-Y., & Yang, P.-T. (2020). *The moderating effect of artificial intelligence phobia on the relationship between trust and product promotion effectiveness: An exploratory study*. 356–359. https://doi.org/10.1145/3377571.3377594

Wang, D., Waldman, D. A., & Zhang, Z. (2014). A meta-analysis of shared leadership and team effectiveness. *Journal of Applied Psychology, 99*(2), 181–198.

Yam, K. C., Tan, T., Jackson, J. C., Shariff, A., & Gray, K. (2023). Cultural differences in people's reactions and applications of robots, algorithms, and artificial intelligence. *Management and Organization Review*, *19*(5), 859–875.

Yin, M., Vaughan, J. W., & Wallach, H. (2019). Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA. *279*, 1–12. https://doi.org/10.1145/3290605.3300509

Yokoi, R., Eguchi, Y., Fujita, T., & Nakayachi, K. (2021). Artificial Intelligence is trusted less than a doctor in medical treatment decisions: Influence of perceived care and value similarity. *International Journal of Human-Computer Interaction*,

37(10), 981–990. https://doi.org/10.1080/10447318.2020.1861763

Yokoi, R., & Nakayachi, K. (2019). The effect of shared investing strategy on trust in artificial intelligence. *Japanese Journal of Experimental Social Psychology*, *59*(1), 46–50. https://doi.org/10.2130/jjesp.1819

Zhang, G., Chong, L., Kotovsky, K., & Cagan, J. (2023). Trust in an AI versus a Human teammate: The effects of teammate identity and performance on Human-AI cooperation. *Computers in Human Behavior*, *139*, 107536.

Zhang, S. Y., Meng, Z. X., Chen, B. B., Yang, X., & Zhao, X. R. (2021). Motivation, social emotion, and the acceptance of artificial intelligence virtual assistants-trust-based mediating effects. *Frontiers in Psychology*, *12*, 728495. https://doi.org/10.3389/fpsyg.2021.728495

Zhong, Y., Sluss, D. M., & Badura, K. L. (2024). Subordinate-to-supervisor relational identification: A meta-analytic review. *Journal of Applied Psychology, 109*(9), 1431–1460.

Zhong, R., Yao, J., Wang, Y., Lyubykh, Z., & Robinson, S. L. (2025). Workplace aggression and employee performance: A meta-analytic investigation of mediating mechanisms and cultural contingencies. *Journal of Applied Psychology, 110*(4), 536–574. https://doi.org/10.1037/apl0001244

Tables and Figures

Table 1. Conceptualizations and operationalizations of trust in humans and AI

		Trust in Humans	Trust in AI
Propensity to trust	Conceptualization	A generalized expectation about the trustworthiness of others; a stable within-party factor (Mayer et al., 1995)	
	Operationalization	• One should be very cautious with strangers (reversed) • Most people can be counted on to do what they say they will do • These days, you must be alert or someone is likely to take advantage of you (reversed)	
Trustworthiness beliefs	Conceptualization	Evaluations of whether the trustee can be trusted based on the trustee's characteristics and actions (e.g., ability, benevolence, integrity; Mayer et al., 1995) • Ability: group of skills, competencies, and characteristics that enable a party to have influence within some specific domain • Benevolence: the extent to which a trustee is believed to want to do good to the trustor, aside from an egocentric profit motive • Integrity: perception that the trustee adheres to a set of principles that the trustor finds acceptable	
	Operationalization	• Ability: X is very capable of performing its job • Benevolence: X is very concerned about my welfare • Integrity: X tries hard to be fair in dealings with others	• Ability: The AI is very capable and proficient in … • Benevolence: The AI will … in my best benefits • Integrity: The AI is truthful in its dealing with me
Trust (as beliefs & expectations)	Conceptualization	• An expectancy held by an individual or a group that the word, promise, verbal or written statement of another individual or group can be relied upon (Rotter, 1967: 651) • Confident positive expectations regarding another's conduct in a context of risk (Lewicki et al., 1998) • A subjective, aggregated, and confident set of beliefs about the other party and one's relationship with her/him (Dietz & Den Hartog, 2006)	• The belief that the other side of the relationship, i.e., technology will work in a functional, helpful and reliable way, providing positive results (McKnight et al., 2011) • Belief that a vendor's services and/or reported results are reliable and trustworthy, and that the vendor will fulfill obligations in an exchange relationship with the user (Shin & Park, 2019; Shin, 2021)
	Operationalization	• I believe that X will keep its promises • I expect X to be honest in our interactions	• I am confident in the system • The system is truthful in its dealings with me
Trust (as attitude)	Conceptualization	The extent to which a person is confident in, and willing to act on the basis of, the words, actions and decisions, of another (McAllister, 1995)	The attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability (Lee & See, 2004: 51)

		Trust in Humans	Trust in AI
	Operationalization	Affect-based • We have a sharing relationship. We can both freely share our ideas, feelings, and hopes Cognition-based • This person approaches his/her job with professionalism and dedication • Given this person's track record, I see no reason to doubt his/her competence and preparation for the job	Overall attitude • To what extent do you trust this AI? • I trust the recommendations by AI • This AI can be trusted/relied on/entrusted with work Cognitive vs. affective (McAllister, 1995; Komiak & Benbasat, 2006) • I feel secure/comfortable/content about relying on the AI system to ... • AI approaches the task with professionalism and dedication
Trust (as intention)	Conceptualization	• The willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party (Mayer et al., 1995: 712) • One's willingness to depend on another party based on its perceived characteristics (Rousseau et al., 1998) • One believes in, and is willing to depend on, another party (McKnight et al., 1998: 474)	
	Operationalization	• I would be willing to let X have complete control over my ... • I would be comfortable giving X a task or problem which was critical to me, even if I could not monitor their actions • If I had my way, I wouldn't let X have any influence over issues that are important to me (reversed)	• Would you follow X's recommendation? • Probability (0-100%) to adopt X's recommendation? • I am willing to use the AI to help me choose/decide/recommend … • I will try/ plan to/ intend to use X in the future
Trust (as behavior)	Conceptualization	Trust informed risk-taking behaviors (Mayer et al., 1995; Dietz & Den Hartog, 2006) • Reliance-related • Disclosure-related	
	Operationalization	• Reliance on suggestions/decisions (e.g., number of times, percentage, frequency) • Monitoring • Disclosure of sensitive information • Amount of money invested in a partner in a trust game	

Notes. Operationalization section only presents sample items; X refers to the trustee.

Table 2. Antecedents of trust in AI, definitions, sample variables, and sources

Antecedent	Definition	Sample variables	Source
AI-related			
Performance	AI's competency to perform the assigned task for achieving individual's goals	Accuracy, system quality, perceived usefulness, performance efficacy	UTAUT, ISSM
Design characteristics			
Tangibility	The capability of AI to be perceived or touched	Tangibility	G&W(2020)
Anthropomorphism	Individual's perception of AI as having human qualities, such as feelings	Anthropomorphism, voice humanness, social presence	G&W(2020)
Interaction characteristics			
Transparency	The extent to which the underlying operating rules and inner logics of AI are apparent	Transparency, explanation (type), provision of heuristics, interpretability	G&W(2020)
Information quality	The quality of the information that the AI produces	Information quality, information richness, personalization, preciseness	ISSM
Interaction quality	The quality of interacting with the AI	Interaction quality, immediacy, social interactivity	ISSM, G&W(2020)
Human-related			
Experience/ expertise			
AI-related experience	Individual's experience and/or expertise with AI	Familiarity of AI, AI experience	UTAUT
Technology-related experience	Individual's experience and/or expertise with traditional technologies	Familiarity with technology, expertise with IT	UTAUT
Task-related experience	Individual's past experience related to the task performed by AI	Domain-specific expertise, issue familiarity	UTAUT
Technology-related education background	Individual's educational background related to traditional technologies	Technology/engineering degree	UTAUT
Ability to understand	Individual's ability to understand AI output	Explainability, understanding	
State			
Emotion	Individual's emotional states toward AI	Positive affect, discomfort, fear, emotional reactance	

Antecedent	Definition	Sample variables	Source
Trait			
AI efficacy	Individuals' belief in their capabilities to use or interact with AI systems properly	AI self-efficacy, robot use self-efficacy	
Task-specific efficacy	Individuals' belief in their capabilities to perform a task	Self-competence in recruitment	
Attitude towards AI	Individual's attitude toward AI technology	AI attitude, implicit theories of AI, general trust in AI, AI acceptance, perceived AI's threat	UTAUT-AI
Attitude toward technology	Individual's attitude toward traditional technologies (e.g., computer, automation)	Technology attachment, personal innovativeness, affinity for technology	UTAUT-AI
Propensity to trust	Individual's general propensity to trust others	Generalized trust level, disposition to trust, trust propensity, trusting stance	
Personality	The enduring configuration of characteristics and behavior that comprises an individual's unique adjustment to life (APA dictionary)	Agreeableness, openness to experience, conscientiousness, extraversion, neuroticism/emotional stability	
Demographics			
Age	/	Age	UTAUT
Gender	/	Gender (male=0, female=1)	UTAUT
Education	Level of education completed by individuals	Education	UTAUT
Job characteristics			
Work experience	The length of time when individuals work for an organization	Work experience	
Work authority	Perceived level of authority at work	Authority of work position	
Work autonomy	Perceived amount of autonomy at work	Autonomy at work	
Task-related			
Task performer	Whether it is AI or human that perform the task to be trusted	Identity decision-maker, agent, doctor, etc.	
Task area	The area or field to which the task relates	Scenario	
Task risk	The expectations of adverse results and error costs arising from AI use (Blut et al., 2022)	Risk, error cost, privacy risk, amount at stake	UTAUT-Meta
Task complexity	The work the technology is performing, such as whether it deals with largely technical or interpersonal judgments	Task complexity, task difficulty, cognitive load	G&W(2020)
Uncertainty	The perceived uncertainty or unpredictability of the situation	Uncertainty, predictability of the situation	

Antecedent	Definition	Sample variables	Source
Context			
Human-AI similarity	Similarity between human and AI	Congruence with self, voice similarity, shared strategy	
Leadership	Traits or behaviors characteristic of organization leader (APA dictionary)	Management commitment, authoritarian leadership	
Organizational culture	A distinctive pattern of fundamental values, beliefs, and assumptions shared by members of an organization	Organizational collectivism	
Organizational regulation & policy	A set of rules, guidelines, and regulations that govern organizational activities	Regulatory protection, responsible corporate privacy	
Social influence	Individual's perception that important others believe he/she should use the AI	Social influence	UTAUT

Note. UTAUT refers to the unified theory of acceptance and use of technology (UTAUT; Venkatesh et al., 2003); UTAUT-AI refers to UTAUT in the AI context (Venkatesh, 2022); UTAUT-Meta refers to a meta-analysis of UTAUT (Blut et al., 2022); ISSM refers to the information system successful model (ISSM; DeLone & McLean, 1992); G&W (2020) refers to a review of trust in artificial intelligence (Glikson & Woolley, 2020).

Table 3.1. Summary of effect size statistics for antecedents of trust in AI

Variable	*k*	*N*	*r*	SD_r	ρ	SD_ρ	95% CI	80% CV	%Var	tau^2	*Q*	I^2
AI-related factors												
Information quality	23	4,566	0.32	0.23	0.37	0.25	[0.26, 0.48]	[0.04, 0.70]	7.81	0.06	281.57***	92.19
Anthropomorphism	27	10,816	0.29	0.23	0.32	0.25	[0.22, 0.42]	[0.00, 0.64]	3.95	0.06	659.04***	96.06
Interaction quality	11	2,293	0.29	0.28	0.31	0.31	[0.10, 0.53]	[-0.11, 0.74]	4.84	0.05	206.47***	95.16
Performance	68	24,906	0.23	0.24	0.24	0.25	[0.18, 0.30]	[-0.09, 0.56]	4.23	0.06	1585.97***	95.78
Transparency	66	13,268	0.25	0.20	0.26	0.20	[0.21, 0.31]	[0.01, 0.51]	11.24	0.04	578.23***	88.76
Human-related factors												
Attitude toward technology	8	2,603	0.58	0.28	0.65	0.33	[0.37, 0.93]	[0.18, 1.11]	1.54	0.10	453.75***	98.46
Ability to understand	3	410	0.50	0.26	0.51	0.25	[-0.14, 1.16]	[0.03, 0.98]	6.20	0.06	32.28***	93.80
Emotion	6	2,702	0.30	0.39	0.34	0.41	[-0.09, 0.77]	[-0.27, 0.95]	1.29	0.17	388.75***	98.71
Propensity to trust	12	3,413	0.21	0.22	0.25	0.24	[0.09, 0.41]	[-0.09, 0.58]	6.78	0.05	162.27***	93.22
Attitude towards AI	20	6,775	0.22	0.22	0.25	0.24	[0.13, 0.36]	[-0.07, 0.56]	5.42	0.05	350.57***	94.58
AI-related experience	24	6,924	0.19	0.24	0.20	0.26	[0.09, 0.31]	[-0.14, 0.54]	5.37	0.07	428.29***	94.63
Emotional stability	6	2,081	0.09	0.04	0.11	0.00	[0.06, 0.17]	[0.11, 0.11]	153.82	0.00	3.25	-53.82
Extraversion	5	1,833	0.09	0.08	0.11	0.07	[-0.01, 0.22]	[-0.01, 0.22]	40.18	0.01	9.96*	59.82
Agreeableness	5	1,833	0.08	0.03	0.10	0.00	[0.06, 0.15]	[0.10, 0.10]	343.50	0.00	1.16	-243.50
Technology-related experience	4	1,196	0.09	0.04	0.09	0.00	[0.02, 0.17]	[0.09, 0.09]	178.46	0.00	1.68	-78.46
Task-related experience	10	3,122	0.08	0.11	0.08	0.09	[0.00, 0.16]	[-0.05, 0.21]	29.48	0.01	30.53***	70.52
Openness	5	1,833	0.03	0.20	0.04	0.23	[-0.26, 0.35]	[-0.31, 0.40]	6.60	0.05	60.61***	93.40
Gender	17	3,893	0.04	0.10	0.04	0.08	[-0.01, 0.10]	[-0.06, 0.15]	45.15	0.01	35.44**	54.85
Education	8	2,451	0.02	0.12	0.01	0.12	[-0.10, 0.13]	[-0.16, 0.18]	22.55	0.01	31.04***	77.45
Age	22	5,686	-0.01	0.11	-0.01	0.10	[-0.07, 0.04]	[-0.14, 0.12]	31.74	0.01	66.16***	68.26
Conscientiousness	13	2,273	-0.039	0.12	-0.05	0.11	[-0.14, 0.03]	[-0.19, 0.09]	43.92	0.01	27.32**	56.08
Task-related factors												
Human-AI similarity	4	888	0.45	0.46	0.46	0.49	[-0.32, 1.25]	[-0.34, 1.27]	1.27	0.22	236.88***	98.73
Task performer	31	15,434	0.27	0.28	0.28	0.28	[0.18, 0.39]	[-0.09, 0.65]	2.21	0.03	1359.77***	97.79
Task complexity	4	1,041	0.03	0.07	0.04	0.02	[-0.07, 0.15]	[0.01, 0.07]	92.79	0.00	3.23	7.21
Task risk	18	10,779	0.03	0.19	0.03	0.20	[-0.07, 0.13]	[-0.23, 0.29]	4.52	0.04	376.48***	95.48

Variable	*k*	*N*	*r*	SD_r	ρ	SD_ρ	95% CI	80% CV	%Var	tau^2	*Q*	I^2
Task uncertainty	4	754	0.00	0.00	0.00	0.00	[-0.01, 0.00]	[0.00, 0.00]	73642.21	-0.01	0.00	-73542.21
Context-related factors												
Social influence	9	3,497	0.50	0.10	0.57	0.10	[0.49, 0.65]	[0.44, 0.71]	16.56	0.01	48.230***	83.44
Trustworthiness												
Ability	8	2,389	0.64	0.10	0.70	0.10	[0.61, 0.79]	[0.56, 0.85]	11.76	0.01	59.52***	88.24
Benevolence	6	759	0.24	0.10	0.27	0.09	[0.13, 0.41]	[0.14, 0.40]	53.44	0.01	9.36	46.56
Integrity	12	2,623	0.18	0.20	0.20	0.21	[0.06, 0.34]	[-0.08, 0.49]	11.35	0.04	96.93***	88.65

Note. k = the number of studies providing effect sizes for the meta-analysis; N = total sample size; r = uncorrected correlation; SD_r = standard deviation of r; ρ = mean correlation corrected for measurement error; SD_ρ = standard deviation of ρ; CI = confidence interval around ρ; CV = credibility interval around ρ; %Var = percentage of variance accounted for by study artifacts; Q = Q-statistic of homogeneity; I^2 = percentage of variation in effect sizes due to true heterogeneity. Within each category, variables are ordered by respective ρ.
$^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$; two-tailed.

Table 3.2. Summary of effect size statistics for consequences of trust in AI

Variable	*k*	*N*	*r*	SD_r	ρ	SD_ρ	95% CI	80% CV	%Var	tau^2	*Q*	I^2
Consequences												
Performance	4	1,218	-0.37	0.09	-0.38	0.06	[-0.50, -0.26]	[-0.48, -0.28]	42.66	0.00	7.03†	57.34

Note. k = the number of studies providing effect sizes for the meta-analysis; N = total sample size; r = uncorrected correlation; SD_r = standard deviation of r; ρ = mean correlation corrected for measurement error; SD_ρ = standard deviation of ρ; CI = confidence interval around ρ; CV = credibility interval around ρ; %Var = percentage of variance accounted for by study artifacts; Q = Q-statistic of homogeneity; I^2 = percentage of variation in effect sizes due to true heterogeneity.
$^{\dagger}p < .10$; two-tailed.

Table 4.1. Moderating effect of trust measurement

Antecedent	Trust Measurement	*k*	*N*	Subgroup Analysis			*F*
				ρ	*SDρ*	95%CI	
Performance	Attitude	42	15182	0.28	0.30	[0.19, 0.38]	
	Intention	8	1845	0.23	0.21	[0.04, 0.41]	
	Behavior	13	6412	0.14	0.08	[0.08, 0.19]	
							0.81
Age	Attitude	14	3976	-0.06	0.08	[-0.11, 0.00]	
	Intention	1	240				
	Behavior	6	993	0.13	0.06	[0.03, 0.24]	
							1.48
Gender	Attitude	9	2460	0.09	0.09	[0.01, 0.17]	
	Intention	1	240				
	Behavior	7	1193	-0.03	0.00	[-0.09, 0.02]	
							1.42
Conscientiousness	Attitude	6	1109	0.04	0.05	[-0.08, 0.15]	
	Intention	0					
	Behavior	7	1164	-0.12	0.09	[-0.24, 0.00]	
							1.05
Task Performer (AI vs human)	Attitude	14	4805	0.07	0.14	[-0.02, 0.16]	
	Intention	2	653	0.07	0.24	[-2.10, 2.24]	
	Behavior	10	7891	0.47	0.25	[0.30, 0.65]	
							1.17

Note. Only subgroup analyses with meaningful between-group differences are presented. k = the number of studies providing effect sizes for the meta-analysis; N = total sample size; ρ = mean correlation corrected for measurement error; SD_ρ = standard deviation of ρ; CI = confidence interval around ρ.

Table 4.2. Moderating effect of AI embodiment

Antecedent	AI Embodiment	*k*	*N*	Subgroup Analysis			*F*
				ρ	*SDρ*	95%CI	
AI-related experience	Robotic	0					
	Virtual	6	2,451	0.39	0.29	[0.09, 0.70]	
	Embedded	17	4,075	0.09	0.18	[-0.01, 0.19]	
							9.06^{**}
Attitude towards AI	Robotic	3	213	0.77	0.18	[0.30, 1.23]	
	Virtual	8	3,644	0.28	0.26	[0.06, 0.50]	
	Embedded	8	2,672	0.12	0.13	[0.01, 0.24]	
							13.09^{***}
Age	Robotic	0					
	Virtual	3	1,024	0.15	0.12	[-0.17, 0.47]	
	Embedded	15	3,560	-0.06	0.05	[-0.11, -0.01]	
							6.14^{*}

Note. Only subgroup analyses with meaningful between-group differences are presented. k = the number of studies providing effect sizes for the meta-analysis; N = total sample size; ρ = mean correlation corrected for measurement error; SD_ρ = standard deviation of ρ; CI = confidence interval around ρ.
$^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$; two-tailed.

Table 4.3. Moderating effect of study design

Antecedent	Study Design	*k*	*N*	Subgroup Analysis			*F*
				ρ	*SDρ*	95%CI	
Performance	Correlational	38	11549	0.36	0.32	[0.25, 0.47]	
	Experimental	29	13173	0.15	0.13	[0.09, 0.2]	
							0.27
Interaction quality	Correlational	6	1577	0.42	0.32	[0.08, 0.77]	
	Experimental	5	716	0.09	0.10	[-0.08, 0.26]	
							5.63*
Transparency	Correlational	5	659	0.47	0.27	[0.12, 0.81]	
	Experimental	61	12609	0.25	0.19	[0.2, 0.3]	
							1.51
AI-related experience	Correlational	5	2073	0.47	0.24	[0.16, 0.77]	
	Experimental	19	4851	0.09	0.17	[0, 0.18]	
							17.97***
Task risk	Correlational	8	2061	-0.03	0.37	[-0.34, 0.29]	
	Experimental	9	8534	0.04	0.15	[-0.08, 0.16]	
							4.76*

Note. Only subgroup analyses with meaningful between-group differences are presented. k = the number of studies providing effect sizes for the meta-analysis; N = total sample size; ρ = mean correlation corrected for measurement error; SD_ρ = standard deviation of ρ; CI = confidence interval around ρ.
$^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$; two-tailed.

Table 4.4. Moderating effect of study setting

Antecedent	Study Setting	*k*	*N*	Subgroup Analysis			*F*
				ρ	*SDρ*	95%CI	
Performance	Online	51	22001	0.21	0.25	[0.14, 0.28]	
	Offline	7	638	0.40	0.19	[0.20, 0.60]	
							2.52
Information quality	Online	18	3970	0.34	0.25	[0.21, 0.47]	
	Offline	4	412	0.64	0.10	[0.43, 0.84]	
							2.11
Task risk	Online	14	10375	0.02	0.18	[-0.09, 0.13]	
	Offline	3	220	0.59	0.10	[0.24, 0.93]	
							8.03*

Note. Only subgroup analyses with meaningful between-group differences are presented. k = the number of studies providing effect sizes for the meta-analysis; N = total sample size; ρ = mean correlation corrected for measurement error; SD_ρ = standard deviation of ρ; CI = confidence interval around ρ.
*$p < .05$; two-tailed.

Figure 1. Flow of the literature search and screening procedure

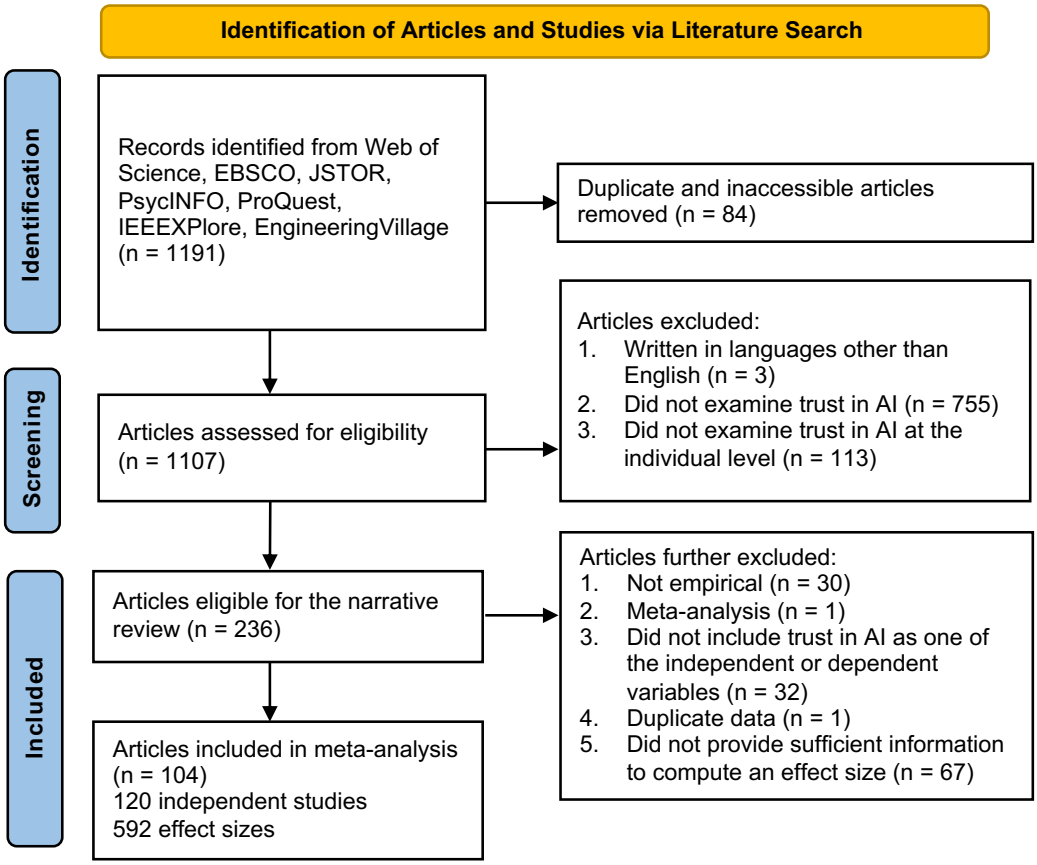


Figure 2. Nomological framework of trust in AI

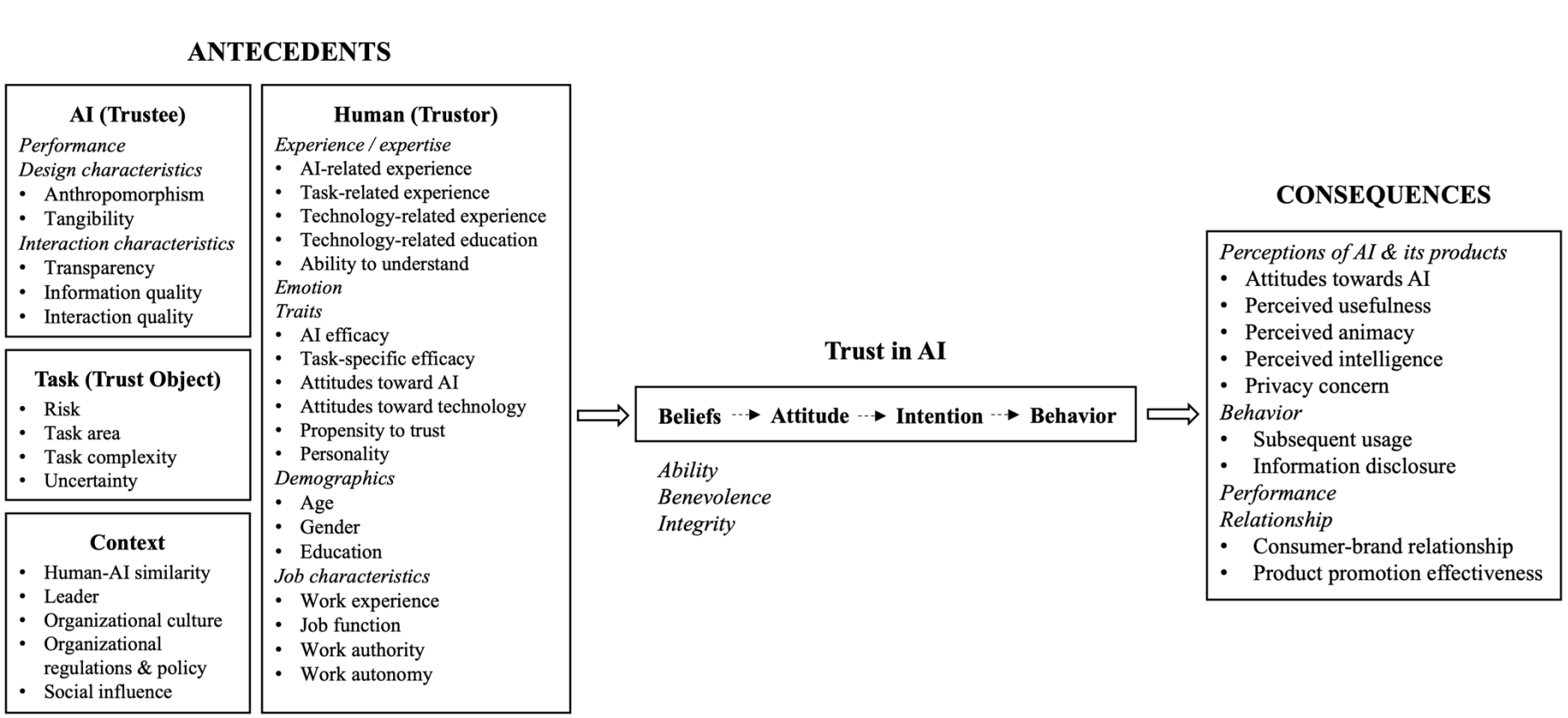
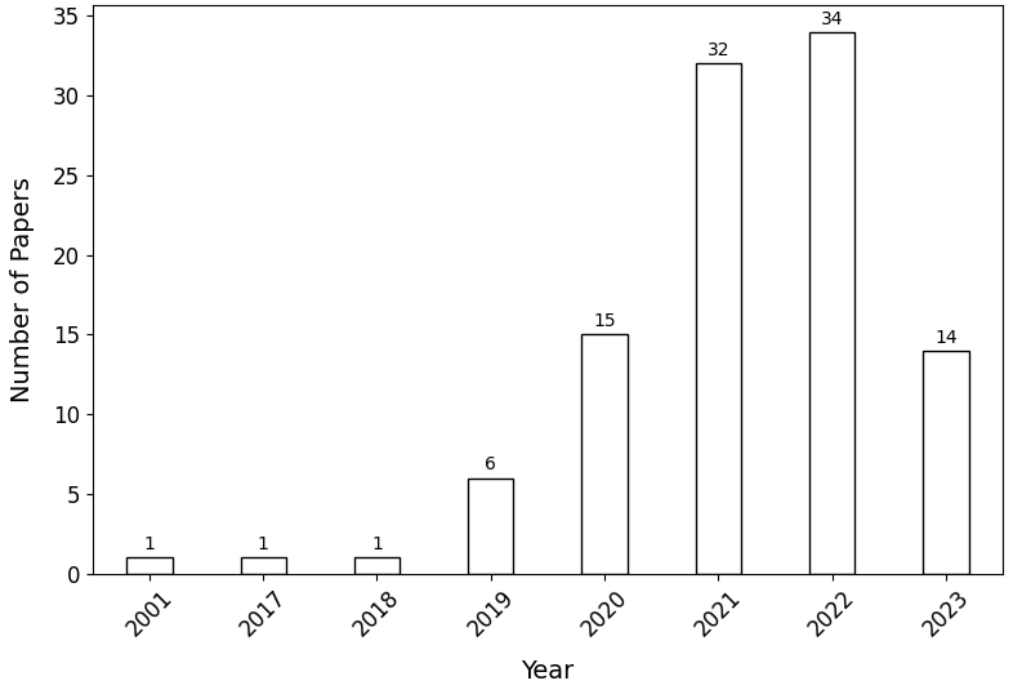
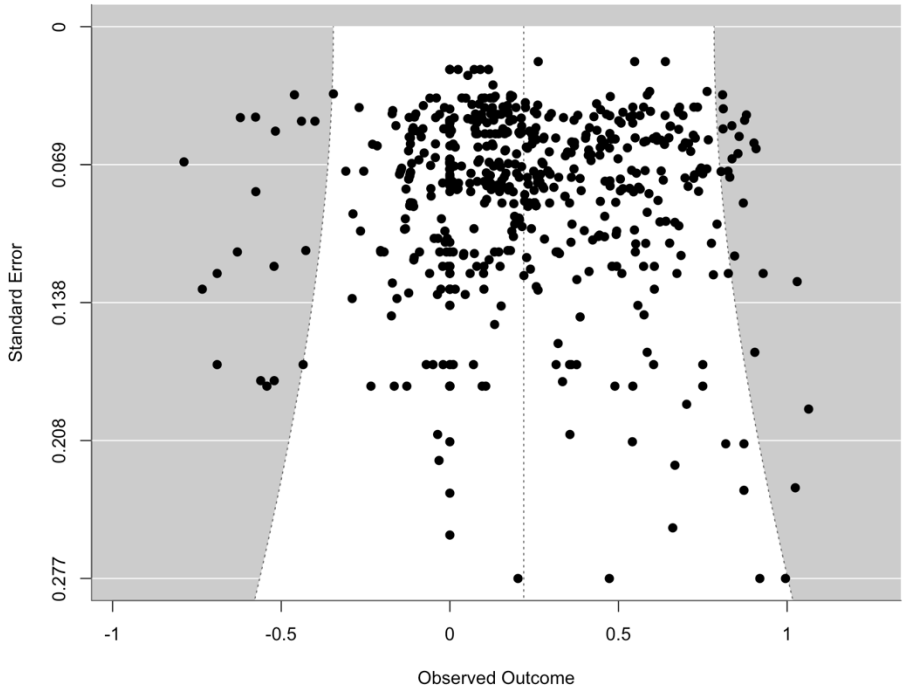


Figure 3. Number of included papers published per year



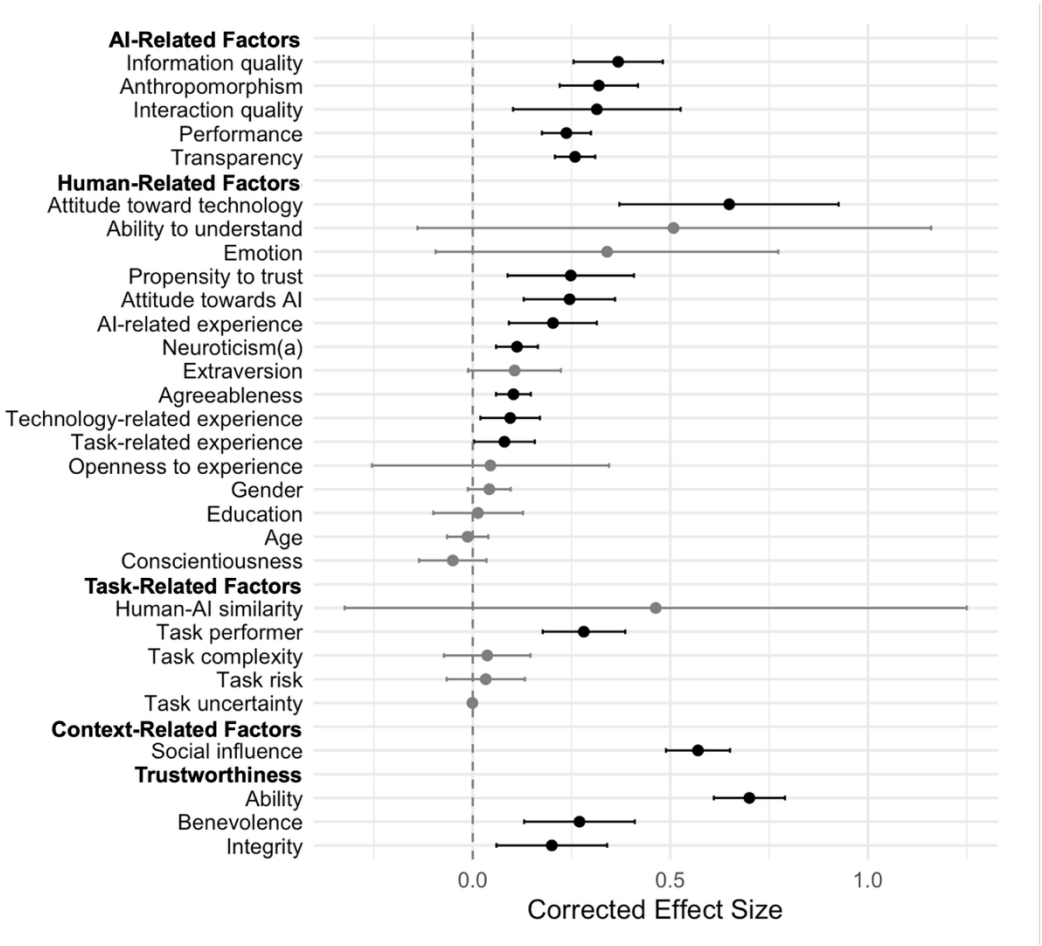
Note. The x-axis refers to the year when the included papers were published. The y-axis refers to the number of papers included in the meta-analysis.

Figure 4. Funnel plot of all included studies



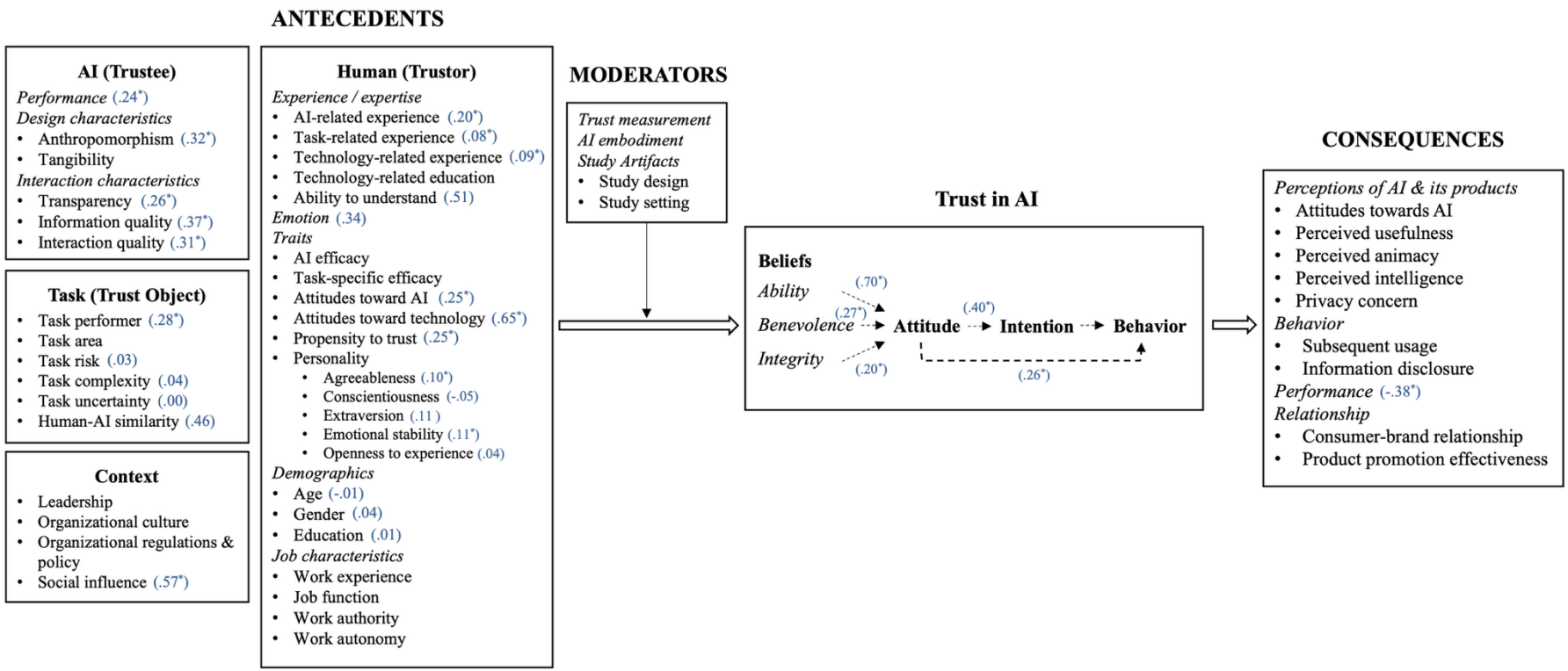
Note. The values on the x-axis (observed outcome) represent effect sizes from the studies included in the meta-analysis. The y-axis refers to the standard error of the effect sizes.

Figure 5. Forest plot of effect sizes of antecedents of trust in AI



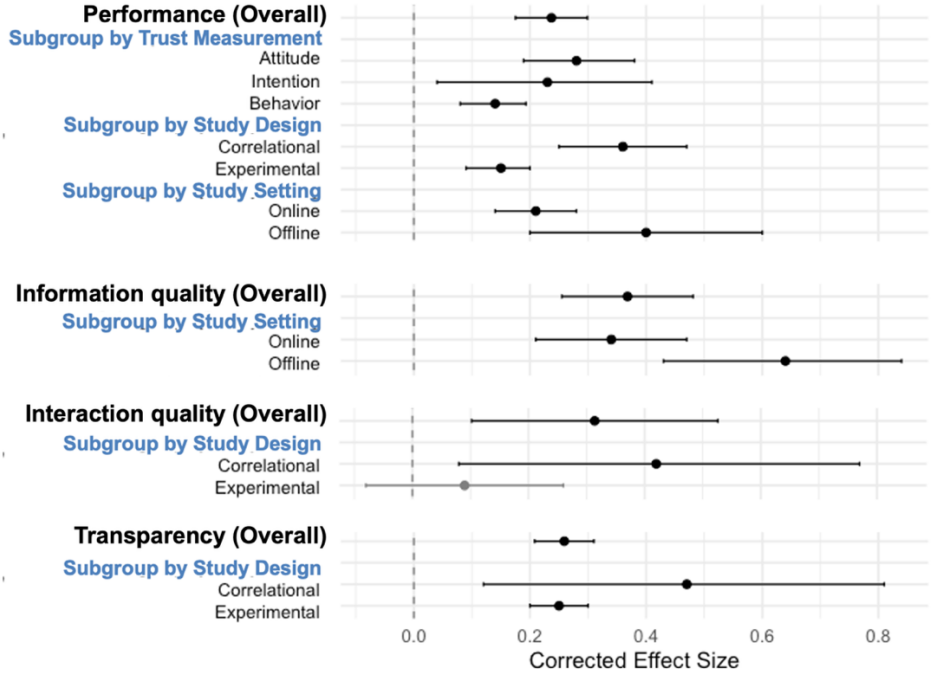
Note. Each point represents a corrected correlations (ρ) between an antecedent and trust in AI. Error bars indicate 95% confidence intervals (CI). Points and error bars shown in black denote significant effects (95% CI excludes zero), while those in gray denote non-significant effects (95% CI includes zero). Only antecedents with $k > 2$ are present.

Figure 6. Nomological framework of trust in AI with corrected average effect sizes

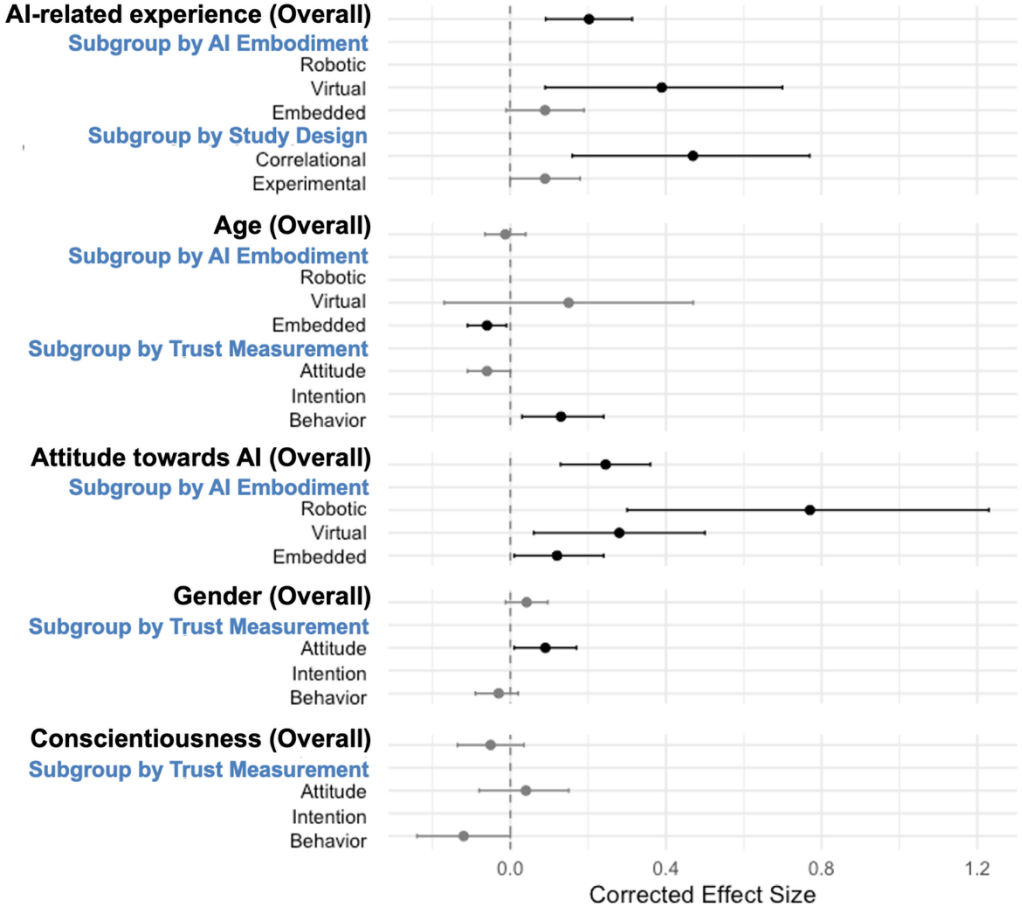


Note. For each antecedent or consequence with $k > 2$, the average effect size (ρ) in relevance to trust in AI is present in paratheses.

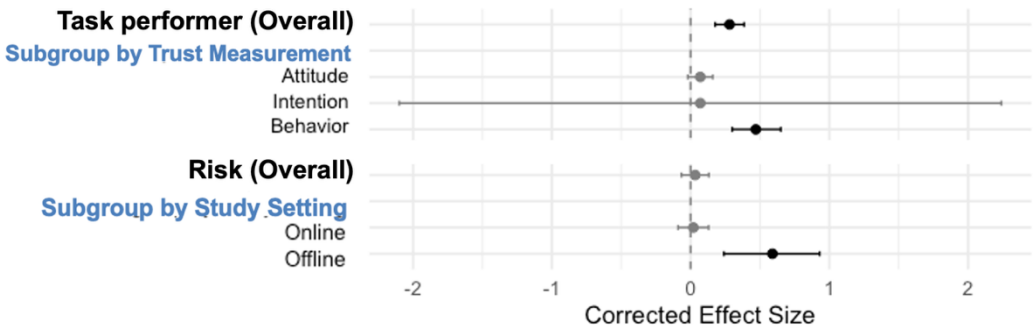
Figure 7. Forest plot of subgroup analysis



(a) AI-related factors (subgroup)



(b) Human-related factors (subgroup)



(c) Task-related factors (subgroup)

Appendices

Appendix 1. Summary of AI-related technologies

Technology	Definition	Relationship with AI	Examples	Search terms
Technologies Mentioned in Search Terms				
Artificial intelligence (AI)	Machines capable of performing cognitive functions commonly attributed to the human mind, such as learning, reasoning, and decision making (Qin et al., 2025); Machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments (OECD, 2019); A new generation of technologies capable of interacting with the environment by (a) gathering information, …, evaluating the results of their actions and improving their decision systems to achieve specific objectives (Ferràs-Hernández, 2018; Glikson & Woolley, 2020); The software-based technology that permits automated machines to sense their surroundings and intelligently make decisions based on the available data (Kaplan et al., 2023)	/	/	artificial* intelligen*, AI
Machine learning (ML)	Computer systems that automatically improve through experience (Jordan & Mitchell, 2015)	Subset of AI focused on data-driven learning	Driver allocation on ride-hailing platforms	machine learning
Expert system	A sophisticated computer program designed to mimic human reasoning and problem-solving skills by utilizing a vast knowledge base and a set of rules or algorithms (EBSCO)	Early AI using IF-THEN rules, sometimes called symbolic AI	Rule-based diagnostic systems in healthcare	expert system
Intelligent automation	The integration of AI with automation technologies	AI-driven	Smart warehouse auto-sorting parcels	intelligent automation
Intelligent agent	A software program designed to make decisions, respond to its environment, and take actions to achieve specific goals, often employing artificial intelligence techniques (EBSCO)	AI-driven	Digital personal assistants	intelligent agent

Technology	Definition	Relationship with AI	Examples	Search terms
Other Technologies not Explicitly Mentioned in Search Terms				
Generative artificial intelligence (GenAI)	Branch of AI that focuses on creating content, including audio, images, text, and videos. It employs advanced algorithms to produce unique outputs based on user prompts… is rooted in machine learning (EBSCO)	Subset of AI, rooted in ML	ChatGPT, DeepSeek	/
Natural Language Processing (NLP)	A filed of AI concerned with enabling computers to understand, interpret, and generate human language	Subset of AI focused on	Voice-to-text transcription	/
Deep learning (DL)	A type of ML in which multilayered (or "deep") artificial neural networks allow a computer system to "earn" from experience, rather than rely wholly on pre-programmed knowledge (EBSCO)	Subset of ML using deep neural networks with multiple layers	Speech and facial recognition	/
Large-language model (LLM)	A neural network trained on vast amounts of text data to predict and generate language (OpenAI)	Subset of DL, also subset of NLP	GPT-4, DeepSeek-R1	/
Robot	Mechanical devices designed to perform tasks independently, often guided by programming and mathematical algorithms (EBSCO)	Some are AI-driven	Service robots, warehouse robots	/
Algorithm	A process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer (Oxford dictionary) A defined sequence of steps designed to accomplish a specific task, often implemented in programming languages so that computers can understand and execute them (EBSCO)	Some are AI-driven	Sorting algorithms	/
Automation	The execution by a machine agent of a function that was previously carried out by a human (Parasuraman & Riley, 1997)	Some are AI-driven	Conveyor belts	/

References:

EBSCO. https://www.ebsco.com/research-starters/

Ferràs-Hernández, X. (2018). The future of management in a world of electronic brains. *Journal of Management Inquiry*, *27*(2), 260–263.

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.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

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

Qin, X., Zhou, X., Chen, C., Wu, D., Zhou, H., Dong, X., ... & Lu, J. G. (2025). AI aversion or appreciation? A capability–personalization framework and a meta-analytic review. *Psychological Bulletin, 151*(5), 580–599.

Appendix 2. List of articles included in the meta-analysis

Ajenaghughrure, I.B., da Costa Sousa, S. C., & Lamas, D. (2020). Risk and Trust in artificial intelligence technologies: A case study of Autonomous Vehicles. *2020 13th International Conference on Human System Interaction (HSI)*, 118–123. https://doi.org/10.1109/HSI49210.2020.9142686

Angerschmid, A., Theuermann, K., Holzinger, A., Chen, F., & Zhou, J. (2022). Effects of Fairness and Explanation on Trust in Ethical AI. *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, *13480*, 51–67. https://doi.org/10.1007/978-3-031-14463-9_4

Aoki, N. (2021). The importance of the assurance that "humans are still in the decision loop" for public trust in artificial intelligence: Evidence from an online experiment. *Computers in Human Behavior*, *114*, 106572. https://doi.org/10.1016/j.chb.2020.106572

Bansal, G., Smith-Renner, A. M., Bucinca, Z., Wu, T., Holstein, K., Hullman, J., & Stumpf, S. (2022). Workshop on Trust and Reliance in AI-Human Teams. *CHI EA '22: CHI Conference on Human Factors in Computing Systems Extended Abstracts*, *116*, 1–6. https://doi.org/10.1145/3491101.3503704

Bayer, S., Gimpel, H., & Markgraf, M. (2022). The role of domain expertise in trusting and following explainable AI decision support systems. *Journal of Decision Systems*, *32*(1), 110–138. https://doi.org/10.1080/12460125.2021.1958505

Bigras, E., Jutras, M. A., Senecal, S., Leger, P. M., Black, C., Robitaille, N., Grande, K., & Hudon, C. (2018). In AI we trust: Characteristics influencing assortment planners' perceptions of AI based recommendation agents. *HCI in Business, Government, and Organizations (HCIBGO 2018)*, 3–16. https://doi.org/10.1007/978-3-319-91716-0_1

Bilal, H., & Várallyai, L. (2021). Artificial intelligence in talent acquisition, do we trust it? *Journal of Agricultural Informatics*, *12*(1), 41–51. https://doi.org/10.17700/jai.2021.12.1.594

Böckle, M., Yeboah-Antwi, K., & Kouris, I. (2021). Can you trust the black box? The effect of personality traits on trust in AI-enabled user interfaces. In Degen, H., Ntoa, S. (eds), *Artificial Intelligence in HCI. HCII 2021. Lecture Notes in Computer Science*, *vol 12797*. https://doi.org/10.1007/978-3-030-77772-2_1

Branley-Bell, D., Whitworth, R., & Coventry, L. (2020). User trust and understanding of explainable AI: Exploring algorithm visualisations and user biases. In Kurosu, M. (eds), *Human-Computer Interaction. Human Values and Quality of Life. HCII 2020. Lecture Notes in Computer Science*, *vol 12183*. https://doi.org/10.1007/978-3-030-49065-2_27

Bruzzese, T., Gao, I., Dietz, G., Ding, C., & Romanos, A. (2020). Effect of confidence indicators on trust in AI-generated profiles. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*, 1–8. https://doi.org/10.1145/3334480.3382842

Carbone, S. C. (2020). *Quantitative influences of trust and unified use and acceptance factors on AI adoption in healthcare* (2448616931) [Ph.D., Capella University]. ProQuest Dissertations & Theses Global. https://www.proquest.com/dissertations-theses/quantitative-influences-trust-unified-use/docview/2448616931/se-2

Chanda, T., Hauser, K., Hobelsberger, S., Bucher, T.-C., Carina Nogueira Garcia, Wies, C., Kittler, H., Tschandl, P., Navarrete-Dechent, C., Podlipnik, S.,

Chousakos, E., Crnaric, I., Majstorovic, J., Alhajwan, L., Foreman, T., Peternel, S., Sarap, S., Özdemir, İ., Barnhill, R. L., … Brinker, T. J. (2023). *Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma* (2790189025). https://www.proquest.com/working-papers/dermatologist-like-explainable-ai-enhances-trust/docview/2790189025/se-2

Chen, Y., Prentice, C., Weaven, S., & Hisao, A. (2022). The influence of customer trust and artificial intelligence on customer engagement and loyalty—The case of the home-sharing industry. *Frontiers in Psychology*, *13*, 912339. https://doi.org/10.3389/fpsyg.2022.912339

Chen, Y.-N. K., & Wen, C.-H. R. (2021). Impacts of attitudes toward government and corporations on public trust in artificial intelligence. *Communication Studies*, *72*(1), 115–131. https://doi.org/10.1080/10510974.2020.1807380

Cheng, X., Zhang, X., Cohen, J., & Mou, J. (2022). Human vs. AI: Understanding the impact of anthropomorphism on consumer response to chatbots from the perspective of trust and relationship norms. *Information Processing & Management*, *59*(3), 102940. https://doi.org/10.1016/j.ipm.2022.102940

Chi, O. H., Chi, C. G., Gursoy, D., & Nunkoo, R. (2023). Customers acceptance of artificially intelligent service robots: The influence of trust and culture. *International Journal of Information Management*, *70*, 102623. https://doi.org/10.1016/j.ijinfomgt.2023.102623

Chi, O. H., Jia, S., Li, Y., & Gursoy, D. (2021). Developing a formative scale to measure consumers' trust toward interaction with artificially intelligent (AI) social robots in service delivery. *Computers in Human Behavior*, *118*, 106700. https://doi.org/10.1016/j.chb.2021.106700

Chiou, E. K., Salehi, P., Blasch, E., Sung, J., Cohen, M. C., Pan, A., Mancenido, M., Mosallanezhad, A., Ba, Y., & Bhatti, S. (2022). Trust in AI-enabled decision support systems: Preliminary validation of MAST criteria. *3rd IEEE International Conference on Human-Machine Systems, ICHMS 2022*. https://doi.org/10.1109/ICHMS56717.2022.9980623

Choi, S., Jang, Y., & Kim, H. (2022). Influence of pedagogical beliefs and perceived trust on teachers' acceptance of educational artificial intelligence tools. *International Journal of Human-Computer Interaction*, *39*(4), 910–922. https://doi.org/10.1080/10447318.2022.2049145

Choudhury, A., Asan, O., & Medow, J. E. (2022). Effect of risk, expectancy, and trust on clinicians' intent to use an artificial intelligence system—Blood Utilization Calculator. *Applied Ergonomics*, *101*, 103708. https://doi.org/10.1016/j.apergo.2022.103708

Choung, H., David, P., & Ross, A. (2022). 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

Diprose, W. K., Buist, N., Hua, N., Thurier, Q., Shand, G., & Robinson, R. (2020). Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. *Journal of the American Medical Informatics Association*, *27*(4), 592–600. https://doi.org/10.1093/jamia/ocz229

Elofson, G. (2001). Developing trust with intelligent agents: An exploratory study. In C. Castelfranchi & YH. Tan (Eds.), *Trust and Deception in Virtual Societies* (pp. 125–138). Springer. https://doi.org/10.1007/978-94-017-3614-5_6

Gautam, R. (2022). Trust in AI – Similarities with and differences from trust in humans [M.S., University of Delaware]. In *ProQuest Dissertations and Theses*

(2776551346). ProQuest Dissertations & Theses Global. https://www.proquest.com/dissertations-theses/trust-i-similarities-with-differences-humans/docview/2776551346/se-2?accountid=12665

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*, 106607. https://doi.org/10.1016/j.chb.2020.106607

Gobel, K., Niessen, C., Seufert, S., & Schmid, U. (2022). Explanatory machine learning for justified trust in human-AI collaboration: Experiments on file deletion recommendations. *Frontiers in Artificial Intelligence*, *5*, 919534. https://doi.org/10.3389/frai.2022.919534

Gupta, K., Hajika, R., Pai, Y. S., Duenser, A., Lochner, M., & Billinghurst, M. (2019). *In AI we trust: Investigating the relationship between biosignals, trust and cognitive load in VR*. 25th ACM Symposium on Virtual Reality Software and Technology. https://doi.org/10.1145/3359996.3364276

Gutzwiller, R. S., & Reeder, J. (2021). Dancing with algorithms: Interaction creates greater preference and trust in machine-learned behavior. *Human Factors*, *63*(5), 854–867. https://doi.org/10.1177/0018720820903893

Harnkham, N. (2023). Artificial intelligence in medicine: Measuring respondents' trust in artificial intelligence for better patient care [D.B.A., William Howard Taft University]. In *ProQuest Dissertations and Theses* (2800671825). ProQuest Dissertations & Theses Global. https://www.proquest.com/dissertations-theses/artificial-intelligence-medicine-measuring/docview/2800671825/se-2?accountid=12665

Heuer, H., & Breiter, A. (2020). How fake news affect trust in the output of a machine learning system for news curation. In van Duijn, M., Preuss, M., Spaiser, V., Takes, F., Verberne, S. (eds), *Disinformation in Open Online Media (MISDOOM 2020)*, *vol 12259*. https://doi.org/10.1007/978-3-030-61841-4_2

Honeycutt, D. R., Nourani, M., & Ragan, E. D. (2020). Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *8*, 63–72. https://doi.org/10.1609/hcomp.v8i1.7464

Hu, P., Lu, Y., & Gong, Y. (2021). Dual humanness and trust in conversational AI: A person-centered approach. *Computers in Human Behavior*, *119*, 106727. https://doi.org/10.1016/j.chb.2021.106727

Huo, W., Zheng, G., Yan, J., Sun, L., & Han, L. (2022). Interacting with medical artificial intelligence: Integrating self-responsibility attribution, human–computer trust, and personality. *Computers in Human Behavior*, *132*, 107253. https://doi.org/10.1016/j.chb.2022.107253

Ingrams, A., Kaufmann, W., & Jacobs, D. (2022). In AI we trust? Citizen perceptions of AI in government decision making. *Policy and Internet*, *14*(2), 390–409. https://doi.org/10.1002/poi3.276

Jakesch, M., French, M., Xiao, M., Hancock, J. T., & Naaman, M. (2019). AI-mediated communication: How the perception that profile text was written by ai affects trustworthiness. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, *239*, 1–13. https://doi.org/10.1145/3290605.3300469

Jham, V., Malhotra, G., & Sehgal, N. (2023). Consumer-brand Relationships with AI Anthropomorphic Assistant: Role of Product Usage Barrier, Psychological Distance and Trust. *International Review of Retail, Distribution & Consumer*

Research, *33*(2), 117–133. https://doi.org/10.1080/09593969.2023.2178023
Juravle, G., Boudouraki, A., Terziyska, M., & Rezlescu, C. (2020). Trust in artificial intelligence for medical diagnoses. *Progress in Brain Research*, *253,* 263–282.
Kandoth, S., & Shekhar, S. K. (2022). Social influence and intention to use AI: the role of personal innovativeness and perceived trust using the parallel mediation model. *Forum Scientiae Oeconomia*, *10*(3), 131–150. https://doi.org/10.23762/FSO_VOL10_NO3_7
Kim, T., & Song, H. (2021). How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair. *Telematics & Informatics*, *61*, 101595. https://doi.org/10.1016/j.tele.2021.101595
Kim, T., & Song, H. (2023). "I believe AI can learn from the error. Or can it not?": The effects of implicit theories on trust repair of the intelligent agent. *International Journal of Social Robotics*, *15*(1), 115–128. https://doi.org/10.1007/s12369-022-00951-5
Kyung, N., & Kwon, H. E. (2022). Rationally trust, but emotionally? The roles of cognitive and affective trust in laypeople's acceptance of AI for preventive care operations. *Production & Operations Management*, 00, 1–20. https://doi.org/10.1111/poms.13785
Lacroux, A., & Martin-Lacroux, C. (2022). Should I trust the artificial intelligence to recruit? Recruiters' perceptions and behavior when faced with algorithm-based recommendation systems during resume screening. *Frontiers in Psychology*, *13*, 895997. https://doi.org/10.3389/fpsyg.2022.895997
Langer, M., König, C. J., Back, C., & Hemsing, V. (2022). Trust in artificial intelligence: Comparing trust processes between human and automated trustees in light of unfair bias. *Journal of Business and Psychology*, *38*(3), 493–508. https://doi.org/10.1007/s10869-022-09829-9
Lapinska, J., Kadzielawski, G., Sudolska, A., Gorka, J., Escher, I., & Brzustewicz, P. (2021). Employee trust in artificial intelligence in chemical industry companies. *Przemysl Chemiczny*, *100*(2), 127–131. https://doi.org/10.15199/62.2021.2.1
Lee, O.-K. D., Ayyagari, R., Nasirian, F., & Ahmadian, M. (2021). Role of interaction quality and trust in use of AI-based voice-assistant systems. *Journal of Systems and Information Technology*, *23*(2), 154–170. https://doi.org/10.1108/JSIT-07-2020-0132
Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., & Mara, M. (2023). Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior*, *139*, 107539. https://doi.org/10.1016/j.chb.2022.107539
Li, J., Zhou, Y., Yao, J., & Liu, X. (2021). An empirical investigation of trust in AI in a Chinese petrochemical enterprise based on institutional theory. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-92904-7
Liu, B. (2021). In AI we trust? Effects of agency locus and transparency on uncertainty reduction in human–AI interaction. *Journal of Computer-Mediated Communication*, *26*(6), 384–402. https://doi.org/10.1093/jcmc/zmab013
Lowe, M., & Hyun, N. K. (2020). Vocal similarity, trust and persuasion in human-AI agen t interactions. *ACR North American Advances, 48*, 907.
Lundberg, H., Mowla, N. I., Abedin, S. F., Thar, K., Mahmood, A., Gidlund, M., & Raza, S. (2022). Experimental analysis of trustworthy in-vehicle intrusion detection system using eXplainable artificial intelligence (XAI). *IEEE Access*,

10, 102831–102841. https://doi.org/10.1109/ACCESS.2022.3208573
Luo, Y., Li, X., & Ye, Q. (2023). The Impact of Privacy Calculus and Trust on User Information Participation Behavior in AI-based Medical Consultation-The Moderating Role of Gender. *Journal of Electronic Commerce Research*, *24*(1), 48–67.
Ma, S., Lei, Y., Wang, X., Zheng, C., Shi, C., Yin, M., & Ma, X. (2023). Who should I trust: AI or myself? Leveraging human and AI correctness likelihood to promote appropriate trust in AI-assisted decision-making. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–19).
Maehigashi, A. (2022). The nature of trust in communication robots: Through comparison with trusts in other people and AI systems. *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 900–903. https://doi.org/10.1109/HRI53351.2022.9889521
Maier, T., Menold, J., & Mccomb, C. (2022). The relationship between performance and trust in AI in E-Finance. *Frontiers in Artificial Intelligence*, *5*, 891529 (17 pp.). https://doi.org/10.3389/frai.2022.891529
Malhotra, G., & Ramalingam, M. (2022). Perceived anthropomorphism and purchase intention using artificial intelligence technology: Examining the moderated effect of trust. *Journal of Enterprise Information Management*, *ahead-of-print*. https://doi.org/10.1108/JEIM-09-2022-0316
Manchon, J. B., Bueno, M., & Navarro, J. (2021). Calibration of trust in automated driving: A matter of initial level of trust and automated driving style? *Human Factors*, *65*(8), 1613–1629. https://doi.org/10.1177/00187208211052804
Micocci, M., Borsci, S., Thakerar, V., Walne, S., Manshadi, Y., Finlay, E., Mullarkey, D., Buckle, P., & Hanna, G. B. (2021). Attitudes towards trusting artificial intelligence insights and factors to prevent the passive adherence of GPs: A pilot study. *Journal of Clinical Medicine*, *10*(14), 3101. https://doi.org/10.3390/jcm10143101
Min, F., Zou, F., He, Y., & Jiang, X. (2021). Research on Users' Trust of Chatbots Driven by AI: An Empirical Analysis Based on System Factors and User Characteristics. *2021 IEEE International Conference on Consumer Electronics and Computer Engineering*, 55–58. https://doi.org/10.1109/ICCECE51280.2021.9342098
Molina, M. D., & Sundar, S. S. (2022). When AI moderates online content: Effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication*, *27*(4), 1–12.
Montag, C., Klugah-Brown, B., Zhou, X., Wernicke, J., Liu, C., Kou, J., Chen, Y., Haas, B. W., & Becker, B. (2023). Trust toward humans and trust toward artificial intelligence are not associated: Initial insights from self-report and neurostructural brain imaging. *Personality Neuroscience*, *6*. https://doi.org/10.1017/pen.2022.5
Nakashima, H., Mantovani, D., & Machado, C. (2022). Users' trust in black-box machine learning algorithms. *Revista de Gestão*, *31*(2), 237–250. https://doi.org/10.1108/REGE-06-2022-0100
Nasirian, F., Ahmadian, M., & Lee, O.-K. D. (2017). AI-based voice assistant systems: Evaluating from the interaction and trust perspectives. *AMCIS 2017 Proceedings*, *2017-August*. https://aisel.aisnet.org/amcis2017/AdoptionIT/Presentations/27
Okamura, K., & Yamada, S. (2020). Adaptive trust calibration for human-AI collaboration. *PLoS ONE*, *15*(2), 1–20.

https://doi.org/10.1371/journal.pone.0229132
Oksanen, A., Savela, N., Latikka, R., & Koivula, A. (2020). Trust toward robots and artificial intelligence: An experimental approach to human–technology interactions online. *Frontiers in Psychology*, *11*, 568256. https://doi.org/10.3389/fpsyg.2020.568256
Ozanne, M., Bhandari Aparajita, Bazarova, N. N., & DiFranzo Dominic. (2022). Shall AI moderators be made visible? Perception of accountability and trust in moderation systems on social media platforms. *Big Data & Society*, *9*(2), 20539517221115666. https://doi.org/10.1177/20539517221115666
Pitardi, V., & Marriott, H. R. (2021). Alexa, she's not human but... Unveiling the drivers of consumers' trust in voice-based artificial intelligence. *Psychology & Marketing*, *38*(4), 626–642. https://doi.org/10.1002/mar.21457
Prakash, A. V., Joshi, A., Nim, S., & Das, S. (2023). Determinants and consequences of trust in AI-based customer service chatbots. *Service Industries Journal*, 1–34. bth. https://doi.org/10.1080/02642069.2023.2166493
Ribes, D., Henchoz, N., Portier, H., Defayes, L., Phan, T.-T., Gatica-Perez, D., & Sonderegger, A. (2021). Trust indicators and Explainable AI: A study on user perceptions. *Human-Computer Interaction – INTERACT 2021*, *12933*, 662–671. https://doi.org/10.1007/978-3-030-85616-8_39
Sassmannshausen, T., Burggraf, P., Wagner, J., Hassenzahl, M., Heupel, T., & Steinberg, F. (2021). Trust in artificial intelligence within production management—An exploration of antecedents. *Ergonomics*, *64*(10), 1333–1350. https://doi.org/10.1080/00140139.2021.1909755
Schelble, B. G., Lopez, J., Textor, C., Zhang, R., McNeese, N. J., Pak, R., & Freeman, G. (2022). 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
Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*, *29*(4), 260–278. https://doi.org/10.1080/12460125.2020.1819094
Selten, F., Robeer, M., & Grimmelikhuijsen, S. (2023). "Just like I thought": Street-level bureaucrats trust AI recommendations if they confirm their professional judgment. *Public Administration Review*, *83*(2), 263–278. https://doi.org/10.1111/puar.13602
Sharan, N. N., & Romano, D. M. (2020). The effects of personality and locus of control on trust in humans versus artificial intelligence. *Heliyon*, *6*(8), e04572. https://doi.org/10.1016/j.heliyon.2020.e04572
Shi, S., Gong, Y., & Gursoy, D. (2021). Antecedents of trust and adoption intention toward artificially intelligent recommendation systems in travel planning: A heuristic–systematic model. *Journal of Travel Research*, *60*(8), 1714–1734. https://doi.org/10.1177/0047287520966395
Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, *146*, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551
Shin, D., Zaid, B., Biocca, F., & Rasul, A. (2022). In Platforms We Trust?Unlocking the Black-Box of News Algorithms through Interpretable AI. *Journal of Broadcasting & Electronic Media*, *66*(2), 235–256.
Smit, D., Eybers, S., & Smith, J. (2022). A data analytics organisations perspective on trust and AI adoption. In Jembere, E., Gerber, A.J., Viriri, S., Pillay, A. (eds) *Artificial Intelligence Research. SACAIR 2021. Communications in Computer*

and Information Science, *vol 1551*, 47–60. https://doi.org/10.1007/978-3-030-95070-5_4

Stevens, A. F. (2022). Trustworthiness of artificial intelligence technology in healthcare transformation: Assessing clinician trust and acceptance of artificial intelligence [Ph.D., Northcentral University]. In *ProQuest Dissertations and Theses* (2758659205). ProQuest Dissertations & Theses Global. https://www.proquest.com/dissertations-theses/trustworthiness-artificial-intelligence/docview/2758659205/se-2?accountid=12665

Suen, H.-Y., & Hung, K.-E. (2023). Building trust in automatic video interviews using various AI interfaces: Tangibility, immediacy, and transparency. *Computers in Human Behavior*, *143*, 107713. https://doi.org/10.1016/j.chb.2023.107713

Sullivan, Y., de Bourmont, M., & Dunaway, M. (2022). Appraisals of harms and injustice trigger an eerie feeling that decreases trust in artificial intelligence systems. *Annals of Operations Research*, *308*(1/2), 525–548. https://doi.org/10.1007/s10479-020-03702-9

Textor, C., Zhang, R., Lopez, J., Schelble, B. G., McNeese, N. J., Freeman, G., Pak, R., Tossell, C., & de Visser, E. J. (2022). Exploring the relationship between ethics and trust in human–Artificial Intelligence teaming: A mixed methods approach. *Journal of Cognitive Engineering and Decision Making*, *16*(4), 252–281. https://doi.org/10.1177/15553434221113964

Tuncer, S., & Ramirez, A. (2022). Exploring the role of trust during human-AI collaboration in managerial decision-making processes. *24th International Conference on Human-Computer Interaction (HCII 2022)*, *LNCS, vol 13518*, 541–557. https://doi.org/10.1007/978-3-031-21707-4_39

Weisman, W. D., & Peña, J. F. (2021). Face the uncanny: The effects of doppelganger talking head avatars on affect-based trust toward artificial intelligence technology are mediated by uncanny valley perceptions. *CyberPsychology, Behavior & Social Networking*, *24*(3), 182–187. https://doi.org/10.1089/cyber.2020.0175

Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & Andre, E. (2019). "Do you trust me?": Increasing user-trust by integrating virtual agents in explainable AI interaction design. *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents (IVA'19)*, 7–9. https://doi.org/10.1145/3308532.3329441

Wischnewski, M., & Kramer, N. (2022). Can AI reduce motivated reasoning in news consumption? Investigating the role of attitudes towards AI and prior-opinion in shaping trust perceptions of news. *1st International Conference on Hybrid Human-Artificial Intelligence (HHAI 2022)*, *354*, 184–198. https://doi.org/10.3233/FAIA220198

Wong, L., Tan, G., Ooi, K., & Dwivedi, Y. (2023). The role of institutional and self in the formation of trust in artificial intelligence technologies. *Internet Research*, *34*(2), 343–370. https://doi.org/10.1108/INTR-07-2021-0446

Woodcock, C., Mittelstadt, B., Busbridge, D., & Blank, G. (2021). The impact of explanations on layperson trust in artificial intelligence-driven symptom checker apps: Experimental study. *Journal of Medical Internet Research*, *23*(11), e29386. https://doi.org/10.2196/29386

Wu, J.-J., Khan, H. A., Chien, S.-H., & Wen, C.-H. (2022). Effect of customization, core self-evaluation, and information richness on trust in online insurance service: Intelligent agent as a moderating variable. *Asia Pacific Management*

Review, *27*(1), 18–27. https://doi.org/10.1016/j.apmrv.2021.04.001
Xiang, H., Zhou, J., & Xie, B. (2022). AI tools for debunking online spam reviews? Trust of younger and older adults in AI detection criteria. *Behaviour & Information Technology*, *42*(5), 478–497. https://doi.org/10.1080/0144929x.2021.2024252
Xu, Y., Huang, Y., Wang, J., & Zhou, D. (2022). How do employees form initial trust in artificial intelligence: Hard to explain but leaders help. *Asia Pacific Journal of Human Resources*, *62*(3), e12402.
Yang, F., Huang, Z., Scholtz, J., & Arendt, D. L. (2020). How do visual explanations foster end users' appropriate trust in machine learning? *IUI'20: Proceedings of the 25th International Conference on Intelligent User Interfaces*, 189–201. https://doi.org/10.1145/3377325.3377480
Yin, M., Vaughan, J. W., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. *CHI'19: CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300509
Yokoi, R., Eguchi, Y., Fujita, T., & Nakayachi, K. (2021). Artificial Intelligence Is Trusted Less than a Doctor in Medical Treatment Decisions: Influence of Perceived Care and Value Similarity. *International Journal of Human-Computer Interaction*, *37*(10), 981–990. https://doi.org/10.1080/10447318.2020.1861763
Yokoi, R., & Nakayachi, K. (2019). The effect of shared investing strategy on trust in artificial intelligence. *Japanese Journal of Experimental Social Psychology*, *59*(1), 46–50. https://doi.org/10.2130/jjesp.1819
Youn, S., & Jin, S. V. (2021). "In AI we trust?" The effects of parasocial interaction and technopian versus luddite ideological views on chatbot-based customer relationship management in the emerging "feeling economy." *Computers in Human Behavior*, *119*. https://doi.org/10.1016/j.chb.2021.106721
Yu, L., & Li, Y. (2022). Artificial intelligence decision-making transparency and employees' trust: The parallel multiple mediating effect of effectiveness and discomfort. *Behavioral Sciences*, *12*(5), 127. https://doi.org/10.3390/bs12050127
Zarifis, A., Kawalek, P., & Azadegan, A. (2021). Evaluating if trust and personal information privacy concerns are barriers to using health insurance that explicitly utilizes AI. *Journal of Internet Commerce*, *20*(1), 66–83.
Zhang, G., Chong, L., Kotovsky, K., & Cagan, J. (2023). Trust in an AI versus a Human teammate: The effects of teammate identity and performance on Human-AI cooperation. *Computers in Human Behavior*, *139*, 107536. https://doi.org/10.1016/j.chb.2022.107536
Zhang, S. Y., Meng, Z. X., Chen, B. B., Yang, X., & Zhao, X. R. (2021). Motivation, social emotion, and the acceptance of artificial intelligence virtual assistants-trust-based mediating effects. *Frontiers in Psychology*, *12*, 728495. https://doi.org/10.3389/fpsyg.2021.728495
Zhang, Y., Vera Liao, Q., & Bellamy, R. K. E. (2020). *Efect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making*. 295–305. https://doi.org/10.1145/3351095.3372852
Zierau, N., Flock, K., Janson, A., Sollner, M., & Leimeister, J. M. (2021). The influence of AI-based chatbots and their design on users' trust and information sharing in online loan applications. *Hawaii International Conference on System Sciences (HICSS)*, 5483–5492.

Appendix 3. Tests of imputation and publication bias for each meta-analyzed relationship

Variables	*k* *before imputation*	*N* *before imputation*	*ρ* *before imputation*	*k*	*N*	*ρ*	Trim-and-Fill *(No. of missing studies)*	Egger's Test	Fail-Safe *k*	*FSk* > 5*k*+10?
AI-related factors										
Information quality	16	3,148	0.39	23	4,566	0.37	0 (*SE*=2.45)	t = 0.87 (p = 0.395)	4651	Yes
Anthropomorphism	23	8,278	0.33	27	10,816	0.32	3 (*SE*=3.44)	t = -0.57 (p = 0.574)	9970	Yes
Interaction quality	8	1,690	0.42	11	2,293	0.31	0 (*SE*=2.18)	t = -0.63 (p = 0.544)	759	Yes
Performance	55	20,306	0.22	68	24,906	0.24	0 (*SE*=4.66)	t = 2.97 (p = 0.004)	43126	Yes
Transparency	36	9,899	0.26	66	13,268	0.26	0 (*SE*=4.78)	t = 1.72 (p = 0.091)	20189	Yes
Human-related factors										
Attitude toward technology	8	2,603	0.65	8	2,603	0.65	0 (*SE* = 1.90)	t = -1.33 (p = 0.231)	3305	Yes
Ability to understand	2	340	0.53	3	410	0.51	0 (*SE*=1.50)	t = -0.28 (p = 0.825)	143	Yes
Emotion	4	1,626	0.06	6	2,702	0.34	0 (*SE* = 1.77)	t = -0.81 (p = 0.463)	522	Yes
Propensity to trust	10	2,363	0.12	12	3,413	0.25	0 (*SE* = 2.28)	t = -2.21 (p = 0.051)	430	Yes
Attitude towards AI	20	6,775	0.25	20	6,775	0.25	0 (*SE* = 2.28)	t = 1.86 (p = 0.079)	3410	Yes
AI-related experience	15	5,142	0.14	24	6,924	0.20	6 (*SE* = 3.25)	t = -0.85 (p = 0.403)	1,829	Yes
Emotional stability	5	1,870	0.11	6	2,081	0.11	1 (*SE* = 1.81)	t = 1.16 (p = 0.311)	37	No
Extraversion	4	1,622	0.08	5	1,833	0.11	0 (*SE* = 1.68)	t = 0.41 (p = 0.707)	26	No
Agreeableness	4	1,622	0.10	5	1,833	0.10	0 (*SE* = 1.68)	t = 0.56 (p = 0.616)	17	No
Technology-related experience	3	798	0.11	4	1,196	0.09	2 (*SE* = 1.47)	t = -2.44 (p = 0.135)	4	No
Task-related experience	7	2,734	0.10	10	3,122	0.08	1 (*SE* = 2.20)	t = -3.99 (p = 0.004)	14	No
Openness	4	1,622	0.12	5	1,833	0.04	0 (*SE* = 1.12)	t = -1.05 (p = 0.371)	0	No
Gender	14	3,726	0.04	17	3,893	0.04	5 (*SE* = 2.71)	t = -0.60 (p = 0.557)	1	No
Education	7	2,053	0.03	8	2,451	0.01	0 (*SE* = 1.84)	t = -1.48 (p = 0.189)	0	No
Age	18	5,121	-0.01	22	5,686	-0.01	5 (*SE* = 3.14)	t = -0.32 (p = 0.752)	0	No
Conscientiousness	9	1,895	-0.06	13	2,273	-0.05	0 (*SE* = 2.30)	t = 0.87 (p = 0.405)	0	No

Variables	*k* *before imputation*	*N* *before imputation*	*ρ* *before imputation*	*k*	*N*	*ρ*	Trim-and-Fill *(No. of missing studies)*	Egger's Test	Fail-Safe *k*	*FSk* > 5*k*+10?
Task-related factors										
Human-AI similarity	4	888	0.46	4	888	0.46	1 (*SE* = 1.61)	t = -5.00 (p = 0.038)	241	Yes
Task performer	26	14,350	0.29	31	15,434	0.28	6 (*SE* = 3.71)	t = -3.92 (p = 0.001)	4,866	Yes
Task complexity	3	226	0.04	4	1,041	0.04	2 (*SE* = 1.51)	t = 1.52 (p = 0.267)	0	No
Task risk	16	10,642	0.04	18	10,779	0.03	0 (*SE* = 2.61)	t = 1.59 (p = 0.131)	421	Yes
Task uncertainty	4	754	0.00	4	754	0.00	0 (*SE* = 1.38)	t = -1.56e^{+15} (p <.001)	0	No
Context-related factors										
Social influence	7	2,298	0.59	9	3,497	0.57	0 (*SE* = 1.84)	t = 1.33 (p = 0.224)	3,581	Yes
Trustworthiness										
Ability	3	586	0.70	8	2,389	0.70	0 (*SE* = 1.90)	t = -1.09 (p = 0.317)	3,535	Yes
Benevolence	4	613	0.26	6	759	0.27	1 (*SE* = 1.82)	t = -0.75 (p = 0.497)	79	Yes
Integrity	6	1,429	0.26	12	2,623	0.20	0 (*SE* = 2.14)	t = 1.59 (p = 0.144)	465	Yes
Performance (outcome)	2	820	-0.42	4	1,218	-0.37	1 (*SE* = 1.58)	t = 4.64 (p = 0.044)	240	Yes

Note. "Before imputation" refers to *k*, *N*, and *ρ* when no reliability imputation is not conducted. In trim-and-fill analyses, a small or zero estimated number of missing studies indicate the symmetry of the funnel plot. In Egger's test, a non-significant *t* value also suggests funnel plot symmetry. Symmetry of funnel plot implies that included studies are distributed evenly around the mean effect size, reducing concerns about systematic publication bias. Fail-Safe *N* indicates the number of unpublished studies with null effect sizes that would be needed to make the estimated effect size non-significant; 5*k*+10 is a commonly used criterion for evaluating whether the Fail-Safe *N* is sufficiently large to alleviate concerns about publication bias (*k* = number of observed studies).

Appendix 4. Summary of outlier analyses

Antecedent	Outlier	*rstudent*	Cook's distance	*ρ*	95%CI	*ρ* after removal	95%CI after removal	*Δρ*
Information quality	Bigras et al. 2018 (*n* =20)	5.02	.69	.37	[.26, .48]	.37	[.25, .48]	0
Anthropomor-phism	Youn & Jin 2021 (*n*=602)	3.18	.26	.32	[.22, .42]	.28	[.18, .37]	- .04
Transparency	Tuncer & Ramirez 2022 (*n*=134)	3.7	.29	.26	[.21, .31]	.25	[.20, .31]	- .01
AI performance	Yin et al. 2019 (*n*=13)	6.55	.29	.24	[.18, .30]	.24	[.17, .30]	0
Emotion	Chi et al. 2023 (*n*=986)	3.49	.57	.34	[- .09, .77]	.07	[- .23, .38]	- .27
Gender	Lacroux & Martin-Lacroux 2022 (*n*=560)	3.26	.57	.04	[- .01, .10]	.01	[- .03, .06]	- .03
Education	Xiang et al. 2022 (*n*=80)	-3.41	.75	.01	[- .02, .10]	.03	[- .06, .12]	+ .02
Age	Oksanen et al. 2020 (*n*=720)	3.14	.53	- .01	[- .10, .13]	- .05	[- .09, .00]	- .04
Risk	Juravle et al. 2020 (Study 1, *n*=176)	5.77	1.87	.03	[- .07, .13]	.02	[- .07, .11]	- .01
Social influence	Carbone 2020 (*n*=213)	3.09	.69	.57	[.49, .65]	.55	[.49, .62]	- .02

Note. rstudent refers to externally studentized residuals, indicating how much a study's effect size deviates from *ρ*; Cook's distance provides an overall measure of the study's influence on the *ρ* estimate, and larger values suggest that omitting the study would substantially alter the estimate; *Δρ* represents the change in *ρ* after removing the outlier(s).

Appendix 5. Meta-regression analysis based on publication year

Variable	*k*	*F* (test of moderator)	*p*
AI-related factors			
Information quality	23	1.14	> 0.05
Anthropomorphism	27	0.98	> 0.05
Interaction quality	11	1.48	> 0.05
Performance	68	1.30	> 0.05
Transparency	66	3.18	0.013
Human-related factors			
Attitude toward technology	8	0.39	> 0.05
Ability to understand	3	NA	NA
Emotion	6	1.95	> 0.05
Propensity to trust	12	0.41	> 0.05
Attitude towards AI	20	0.22	> 0.05
AI-related experience	24	1.49	> 0.05
Emotional stability	6	0.05	> 0.05
Extraversion	5	6.98	> 0.05
Agreeableness	5	1.14	> 0.05
Technology-related experience	4	4.40	> 0.05
Task-related experience	10	0.17	> 0.05
Openness	5	43.1	0.023
Gender	17	0.24	> 0.05
Education	8	4.02	0.091
Age	22	2.85	0.067
Conscientiousness	13	0.30	> 0.05
Task-related factors			
Human-AI similarity	4	NA	NA
Task performer	31	0.42	> 0.05
Task complexity	4	2.46	> 0.05
Task risk	18	4.91	0.012
Task uncertainty	4	NA	NA
Context-related factors			
Social influence	9	7.21	0.029

Note. k = the number of studies providing effect sizes for the meta-analysis; F = omnibus test of the moderator effect; p = the p value associated with the F test; values above 0.05 indicate that effect sizes are not contingent on publication year.

Essay 2 and Essay 3: Unpacking the Effect of AI Transparency on Trust in AI

General Introduction

As artificial intelligence (AI) increasingly penetrates people's work and daily lives, considerable efforts have been invested in developing governance frameworks for AI's functionality and ethics, among which AI transparency is often highlighted as a crucial component. For instance, the European Union's AI Act, the first regulation on AI, features a section specifically outlining transparency requirements, such as disclosing summaries of copyrighted data used for training. Transparency and explainability were also part of the eleven governance principles published by the Singapore government.

In academic research, AI transparency has also received extensive attention across various disciplines, particularly due to the significance of "explainable AI (XAI)" in algorithm design and its influence on user responses. It is defined in management research as "the level to which the underlying operating rules and inner logics of the technology are apparent to the users" (Glikson & Woolley, 2020, p. 631). Enhancing AI transparency is widely regarded as beneficial in various aspects. It fosters human trust in AI (Glikson & Woolley, 2020; Langer & König, 2023), mitigates the risk of error and misuse, facilitates the distribution of responsibility, and shows respect for people (Blackman & Ammanath, 2022). Additionally, research has shown that AI transparency is associated with users' satisfaction, adoption intentions, and actual usage behaviors (Bayer et al., 2021; Bigras et al., 2018; S. Yu et al., 2023).

Transparency is viewed as more critical for AI technologies than for more traditional technologies such as automation. The latter is often viewed as deterministic and acts following pre-set guidelines, while AI functions in highly complex and partially random environments, leading to non-deterministic, multi-layered behaviors.

This introduces uncertainty and risk in human-AI interactions, which are critical elements of trust. Glikson and Woolley (2020) suggested that trust is a particularly relevant outcome of transparency.

Despite extensive research efforts, included papers in Essay 1 reported inconsistent findings regarding the effect of AI transparency on trust. On one hand, transparent AI has been found to facilitate more favorable attitudes, higher intentions to trust in AI, and more acceptance behaviors (Kyung & Kwon, 2022; Shin, 2021; Shin et al., 2022; L. Yu & Li, 2022). For instance, illustrating decision heuristics (Elofson, 2001), presenting data points or keywords important to the decision-making process (Schmidt et al., 2020b; Zhou et al., 2019), and providing textual or visual explanations for the decisions (Gobel et al., 2022; Leichtmann et al., 2023; Selten et al., 2023) were all found to increase human trust in AI. On the other hand, some studies found a negative or non-significant relationship between AI transparency and trust. Bayer et al. (2021) found that explanations of an AI-based decision support system were related to less reliance on AI's suggestions, although the difference was not statistically significant. Nakashima et al. (2022) found that people displayed high trust in machine learning algorithms, regardless of the existence of explanation artifacts. In addition, educating users about how machine learning works could not significantly affect trust (Leichtmann et al., 2023).

I argue that such inconsistency in the effect of AI transparency stems from the fact that studies have encompassed different dimensions and facets of this construct, leading to a lack of convergence on its conceptual and operational meaning. To gain deeper insights into how AI transparency has been conceptualized and empirically examined, I reviewed the relevant articles included in the meta-analysis in Essay 1 (see Table 1). To categorize different facets of AI transparency, I draw on the

framework of Organizational Transparency (Schnackenberg et al., 2021; Schnackenberg & Tomlinson, 2016), which proposes three facets underlying transparency: information accuracy, disclosure, and clarity.

Building on these findings, Essay 2 and Essay 3 develop a multi-facet conceptualization of AI transparency, comprising four core facets: (a) AI information accuracy (i.e., the extent to which the information delivered by AI is accurate), (b) AI information disclosure (i.e., the extent to which information about how and/or why AI reaches a particular decision, recommendation, or prediction is disclosed), (c) AI information clarity (i.e., the extent to which the information delivered by AI is understandable for job seekers, even when they have little technical knowledge), and (d) AI information personalization (i.e., the extent to which information delivered by AI is customized according to job seekers' unique characteristics and/or preferences).

To empirically test the relationships of these facets on the formation of trust in AI, I design two between-subject experiments within the context of AI-mediated career assessment and feedback sessions. The first experiment focuses on AI information accuracy and disclosure, given their primary role in determining the presence and reliability of AI-provided information, which is more content-focused. The second experiment focuses on AI information clarity and personalization, exploring how the manner and style of information communication (more process-focused) shape trust formation. Separating these facets across two studies allows the balancing of theoretical clarity with practical considerations such as experimental complexity, participant burden, and statistical power.

Theoretical Background

Information Uncertainty in Human-AI Interactions

Generally, it is believed that there is a certain level of *information uncertainty* embedded in the interactions with AI, which means users "have less information available than they ideally would like to have in order to be able to confidently form a given social judgment" (Van Den Bos, 2009, p. 198). According to the Uncertain Reduction perspective (Berger, 1986; Berger & Calabrese, 1975), such uncertainty involves two elements – *explanation* and *prediction*. Individuals, when interacting with another party, would seek explanations of the other's behavior (e.g., why, what it meant), trying to reduce the number of alternative explanations; they also seek to develop a priori predictions about the alternative acts that the other party might take.

Originally developed for research on interpersonal relationships, information uncertainty between the two interacting parties and uncertainty reduction also applies to the context of organizational behaivor (e.g., Lind & Van den Bos, 2002) and AI (e.g., Gobel et al., 2022; Liu, 2021). Information uncertainty is particularly salient for human-AI interactions, especially during initial interactions, given the complexity and "black box" nature of AI's functioning process. Just as humans are inherently motivated to make sense of the events perceived in their environments (Heider, 1958), AI users are motivated to reduce the uncertainty in their interactions with AI by gathering more information about AI's inner logic, understand and evaluate AI-generated information, results, and recommendations, and subsequently make better plans to achieve their goals (Berger & Calabrese, 1975; Liu, 2021).

AI Transparency: Beyond Information Disclosure

One important approach to reduce such information uncertainty is through building transparency (Venkatesh et al., 2016). This enables users to obtain information and make sense of how and why AI presents certain results or makes specific suggestions. In the context of AI, transparency is defined as the level to

which “the underlying operating rules and inner logics of the technology” (Glikson & Woolley, 2020, p. 631) or “the internal mechanics of a system (e.g., AI’s mind)” (De Freitas et al., 2023) are apparent, observable, and understandable by humans. A transparent AI informs users about its functionality, capabilities, and accuracy, usually by outlining its decision-making process and providing illustrations for its results and decisions. Such information equips users with basic-level knowledge, regardless of their technological expertise, that helps to make sense of and evaluate AI-generated results during the human-AI interaction process.

Table 1 revealed that research predominantly attempted to enhance AI transparency by offering *explanations* differing in amount (e.g., placebo, brief, detailed), content (e.g., past examples, rules), and way of presentation (e.g., visual, textual). For instance, Elofson (2001) found that intelligent agents providing their decision heuristics (versus not providing) were trusted more and preferred when making hiring decisions. Similarly, an AI interviewer providing explanations about its functionality was trusted more by job interviewees (Suen & Hung, 2023). Alam and Mueller (2021) found that local explanations that were based on specific cases, compared with global explanations based on general cases, were associated with higher user satisfaction, trust, and understanding of AI. Additionally, local explanations were more effective when presented visually rather than in terms of text.

While disclosure of information is the first step in creating transparency, I argue that relatively less attention has been paid to the *quality* of such information – how accurate is the provided information (Schmidt et al., 2020; Tuncer & Ramirez, 2022; Zhang et al., 2020)? How is the information relevant to users’ individual preferences and characteristics (Shi et al., 2021; Wu et al., 2022; Yokoi et al., 2021)? How well the provided information was understood and digested by the recipients (De

Freitas et al., 2023; Glikson & Woolley, 2020; Shin, 2021)? For instance, De Freitas and colleagues (2023) conceptualized transparency as not only making the inner mechanics of the system *observable*, but also making them *understandable* to users. Glikson and Woolley (2020) emphasized in their review the importance of offering explanations that are "*understandable* to users, even when they have *little* technical knowledge" (p. 631). Relevant to this are the concepts of interpretability, comprehensibility, and explainability – capturing whether the inner process is hard to examine, comprehend, and explain (Ashoori & Weisz, 2019; Lundberg et al., 2022; Shin, 2021; Tuncer & Ramirez, 2022). Additionally, in an organizational context where AI helped with task assignments, Yu and Li (2022) distinguished the concept of AI decision-making transparency from employee perceived transparency – the former refers to whether information about AI's working mode is released, while the latter focuses on the availability of information subjectively perceived by employees.

From Information Disclosure to Information Quality

In Essay 2 and Essay 3, I propose a broader approach to studying AI transparency – conceptualizing it as the perceived quality of AI-provided information, which encompasses multiple elements, each explaining a fundamental aspect of transparency.

The idea of information quality has been introduced and emphasized in "research on transparency and technology acceptance. In this section, I discuss how the Information System Success Model (DeLone & McLean, 1992; Delone & McLean, 2003) and the framework of Organizational Transparency (Schnackenberg et al., 2021; Schnackenberg & Tomlinson, 2016) contribute particularly to illustrating the relationship between transparency, information quality, and trust.

Information System Success Model (ISSM)

This model focuses on unveiling the factors contributing to the effectiveness of information systems. They argue for the serial nature of information communication: the information system (a) generates information, (b) communicates the information product to the recipient, and (c) the recipient is influenced or not influenced by the information product. Stage (a) concerns *system quality*, the desired characteristics of the information system itself. Stage (b) is related to *information quality*, the desired characteristics of the information output produced by the system. Finally, stage (c) includes the receipt of information output (i.e., use) and its influence on the recipient, including *user satisfaction*, *individual impact*, and *organizational impact*. Specifically, *information quality* is captured by a set of desired characteristics such as accuracy, reliability, clarity, understandability, precision, timeliness, completeness, relevance, and personalization of output (DeLone & McLean, 1992, 2004). Most research investigates this construct from the perspective of the information recipient, thus the specific set of desired characteristics is chosen rather subjectively depending on the application scenarios (DeLone & McLean, 1992). However, as we discuss in the next section, our framework of AI transparency consists of four basic facets that appear to be universally valued by AI users.

The Framework of Organizational Transparency

Developed by Schnackenberg and Tomlinson (2016), this framework reflects one of the primary efforts in synthesizing research on transparency and articulating its multidimensional nature. They define *transparency* as a *perception* of information quality, "the perceived quality of intentionally shared information from a sender" (Schnackenberg & Tomlinson, 2016, p. 1788), in the organizational context. They further contend that transparency is *multidimensional*, consisting of three distinct dimensions rather than being unidimensional as is often presumed. The three

dimensions are: (a) *disclosure*, the perception that relevant information is made available to the recipient, (b) *clarity*, the perceived degree of lucidity and comprehensibility of the communicated by the sender, and (c) *accuracy*, the perception that the provided information is correct and precise to the extent possible, considering the nature of the sender-recipient relationship. Additionally, it is worth noting that they found in a subsequent survey that transparency perception was capable of explaining the variance in important outcomes, such as trustworthiness (i.e., ability, benevolence, integrity).

A Typology of Multifaceted AI Transparency

Building upon the review in Essay 1 and drawing on the ISSM and framework of organizational transparency, I propose a typology of AI transparency that consists of four key facets: *disclosure*, *accuracy*, *clarity*, and *personalization*. Each facet reflects perceptions of a desired characteristic of the information output provided by the AI model. A categorization of the transparency-related variables for each Essay 1 article based on their relevance to the four facets can be found in Table 1.

AI Information Disclosure

Disclosure is defined as whether information about how and/or why AI reaches a particular decision, recommendation or prediction is *made available* to its users. As discussed above, AI users are inherently motivated to make sense of the output, decisions, or suggestions provided by AI. This includes understanding the process by which the model arrives at its decisions, such as the heuristics it employs, the underlying operating principles, and the inner mechanics that govern its functioning. For example, AI-based tools can disclose the important training data (Zhou et al., 2019), words (Schmidt et al., 2020), textual features, and behavioral features (Xiang et al., 2022) that are critical to the decision-making process.

AI Information Accuracy

AI information accuracy captures whether the information delivered by AI is *correct* to the extent possible given the sender-recipient relationship. It reflects users' *subjective* perceptions of whether the provided information is reliable or credible. Importantly, it differs from AI effectiveness, which concerns AI's capability to achieve its functional goals and objective performance outcomes. In empirical studies, AI accuracy is frequently manipulated or communicated through objective metrics like accuracy rate (Kim et al., 2021) and confidence scores (Ashoori & Weisz, 2019; Chanda et al., 2024; Tuncer & Ramirez, 2022; Zhang et al., 2020). However, within the framework of AI transparency, AI information accuracy is distinct in that it captures *subjective* perceptions held by the users.

AI Information Clarity

Clarity refers to the extent to which the information delivered by AI is understandable for recipients, even those with limited technical knowledge. Yet it is less clear in the literature the specific indicators of interpretability, explainability, or understandability. One potential approach to improving clarity for non-technical audiences is to use more layman's language over technical jargons. Another approach from the perspective of AI users argues that it is also important for users to enhance their ability to properly use and evaluate AI products, which is conceptualized as users' AI literacy (Leichtmann et al., 2023; Wang et al., 2022).

AI Information Personalization

Finally, this facet refers to the extent to which the information delivered by AI is customized according to the recipient's unique characteristics and/or preferences. This enables users to feel that their uniqueness is not neglected and their preferences are taken into consideration by the AI. For instance, AI recommendation agents are

able to tailor product recommendations that satisfy customers' personal demands and preferences (Min et al., 2021; Wu et al., 2022) or provide medical diagnoses that are based on the patient's unique symptoms and situations (Yokoi et al., 2021).

Overview

The following sections are organized as follows: Essay 2 and Essay 3 each hypothesizes a theoretical model featuring two facets of AI transparency – AI information disclosure and accuracy in Essay 2, and AI information clarity and personalization in Essay 3. For each Essay, sections on hypothesis development, methods and results, and discussions will be presented. In the end, a general discussion section is provided to summarize the findings from Essay 2 and Essay 3, discuss their theoretical implications, practical implications, and limitations, as well as outline future research directions.

Essay 2: Role of AI Information Disclosure and Accuracy in Facilitating Trust in AI

Hypothesis Development

Transparency and Trust

Transparency is positively associated with trust in various contexts from manufacturing, inter-organizational communications, to e-governance (Edmonds et al., 2019; Schnackenberg et al., 2021; Venkatesh et al., 2016). A recent meta-analysis showed that, in interpersonal relationships, transparency of the trustee was positively related to perceived trustworthiness ($r = 0.48$). In the context of technology acceptance and adoption, transparency is believed to be one of the critical factors predicting development of human trust in automation (Schaefer et al., 2016), robots (Hancock et al., 2011), recommendation agent (Benbasat & Wang, 2005) and AI (Glikson & Woolley, 2020). Greater transparency indicates more relevant information is available to the users, which then alleviates the information uncertainty or asymmetry perceived during the human-technology interaction processes (Xiao & Benbasat, 2007). This enables users to better understand and navigate their interactions with these technologies, and to have richer evidence for forming evaluations of their trustworthiness and deciding subsequent attitudinal and behavioral responses.

I propose that AI transparency exerts a direct influence on users' beliefs in AI's trustworthiness, which, in turn, affects the formation of trust attitudes as well as adoption intentions and behaviors. To unpack the effect of AI transparency on perceived trustworthiness, I draw on Mayer et al.'s (1995) framework of Ability-Benevolence-Integrity, which indicates that a major portion of trustworthiness is explained by three factors – ability, benevolence, and integrity. In the context of AI,

ability refers to AI's competencies, capabilities, and attributes that enable it to perform effectively within a specific domain. *Benevolence* reflects the extent to which AI is believed to act in users' best interests, beyond egocentric profit motives. Finally, *integrity* captures AI's adherence to principles that are considered acceptable by the users.

In the next section, I discuss how AI information disclosure and AI information accuracy, as facets of AI transparency, influence the three dimensions of AI trustworthiness. It is worth investigating the effects of three trustworthiness dimensions separately, as they capture different components – ability deals more with "can-do", while benevolence and integrity concern more about "will-do" (Colquitt et al., 2007).

AI Information Disclosure and Perceived Trustworthiness

Prior research indicates two possible effects of information disclosure. On one hand, AI that discloses information about the process, criteria, and even constraints of its decision-making process may be perceived as *more* trustworthy. Such disclosure signals the existence of a clear, logical, and explainable system that guides AI's functioning process and showcases AI's capability to communicate with the user, increasing the perception of AI's ability. Besides, disclosure, as often viewed as voluntary, conveys AI's willingness to disclose rather than to conceal, and its intention to reduce the information asymmetry and take responsibility for its decisions (Kyung & Kwon, 2022). Such disclosed information will allow information recipients to make informed decisions (Tomlinson & Schnackenberg, 2022), which further contributes to users' evaluations of AI's benevolence. It also indicates that the AI is honest and adheres to the principle of openness and information sharing (Schnackenberg & Tomlinson, 2016).

On the other hand, mere disclosure may not be enough to cultivate trustworthiness. De Freitas et al. (2023) proposed that not all explanations are equally effective at improving attitudes towards AI. This might also be due to the reason that individuals vary in their abilities to understand and make sense of AI-provided information. If the packaging of information is hard to understand or lacks legitimacy labels, users may cast doubt on the ability and motivation of AI to disclose such information. Suspicion of the possibility that AI deliberately obscures or intentionally manipulates the communication process is even harmful to the formation of trustworthiness perceptions.

> ***Hypothesis 1.*** *AI information disclosure increases perceived AI trustworthiness in terms of (a) ability, (b) benevolence, and (c) integrity.*

> ***Hypothesis 1 (alternative).*** *AI information disclosure is not significantly related to perceived AI trustworthiness in terms of (a) ability, (b) benevolence, and (c) integrity.*

AI Information Accuracy and Perceived Trustworthiness

Accurate information from AI will primarily affect the development of ability beliefs, as it provides proof of AI's competence to complete assigned tasks effectively. Conversely, information perceived as less accurate raises doubts about AI's capabilities. This skepticism is particularly detrimental in the case of AI, given that ability is often regarded as the most critical factor in the three dimensions of trustworthiness in research on technology trust, or even the sole criterion determining AI's trustworthiness when AI is presumed to lack agency (McKnight et al., 2011).

Although less theorized in prior research, the provision of accurate information from AI may also enhance users' perceptions of its benevolence and integrity. Voluntarily sharing accurate information signals care for the users' benefits

and shows an aversion to engaging in manipulative practices (Schnackenberg & Tomlinson, 2016). In contrast, the delivery of inaccurate information by AI undermines the inference of AI's intention, leading to doubt on its commitment to be beneficial, honest, truthful and unbiased.

> ***Hypothesis 2.*** *AI information accuracy increases perceived AI trustworthiness in terms of (a) ability, (b) benevolence, and (c) integrity.*

Interaction of AI Information Disclosure and Accuracy

I further propose that the effect of AI information accuracy on shaping trustworthiness perceptions may depend on the level of AI information disclosure.

On one hand, when AI-provided information is perceived as accurate, additional disclosed information provides transparency into the AI's processes and logic, which helps to reinforce perceptions of trustworthiness. On the other hand, things become complicated when users perceive the AI-provided information as inaccurate. Specifically, they will experience uncertainty about the underlying cause – whether the inaccuracy results from a technical limitation, a random error, or their own misunderstanding. This uncertainty can erode trust, as users lack sufficient cues to attribute the cause of the perceived inaccuracy.

In this context, additional disclosed information can paradoxically reduce users' suspicion of the AI that provides seemingly inaccurate information. By offering detailed explanations or signals of openness, disclosure can create an impression that the AI is reliable and acting in good faith. As a result, users may be more inclined to attribute the perceived inaccuracy to their own misunderstanding or to acceptable variation in the outputs, rather than to flaws in the AI's functionality or motives. This ultimately would lead to *overtrust* in the inaccurate AI. In contrast, in the absence of

disclosure, users have fewer cues to explain the perceived inaccuracy, so they are more likely to doubt the AI's functionality or motives.

Given this, while AI information disclosure may reinforce perception of trustworthiness, it may also reduce suspicion and attenuates the negative impact of perceived inaccuracy (or low accuracy). As such, the direction of this moderation effect remains an open empirical question. Therefore, I only theorize the interaction effect between AI information accuracy and AI information disclosure without hypothesizing the direction of the moderation effect:

> ***Hypothesis 3.*** *AI information accuracy interacts with AI information disclosure to influence perceived AI trustworthiness.*

Perceived Trustworthiness, Trust in AI, and Consequences

The Theory of Planned Behavior (TPB) (Ajzen, 1991) describes a development process flowing from beliefs to attitudes, and behavior. In our research context, a human trustor first forms trustworthiness beliefs, which represent cognitive appraisals of specific attributes (e.g., ability, benevolence, integrity) of AI based on clues before and during interaction. Beliefs serve as a foundation for *attitudes* that reflect an individual's general evaluation of AI as a whole. Attitudes in turn predict behavioral *intentions*, which translate into actual *behaviors* when the situation permits.

Following this rationale, I propose that perceived AI trustworthiness will contribute to the formation of trust in AI and elicit subsequent behavioral responses. The distinction between AI trustworthiness and trust in AI allows for examining whether various beliefs consistently translate into generalized trust in AI, eventually informing actual behavior and future behavioral intentions.

Trust in AI

Aligning with the above arguments, Mayer et al. (1995) proposed in their framework that trust is a function of perceived ability, benevolence, and integrity, and the three distinct and related factors account for the within-trustor variation in trust for others (Mayer & Davis, 1999). In parallel, AI users first form their cognitive judgments and evaluations of the AI's attributes associated with their ability, benevolence, and integrity, and subsequently generate more holistic attitudes about whether the AI can be trusted.

In addition, I proposed that trustworthiness in terms of ability may exert a stronger influence on the formation of trust attitudes in AI. AI and other technologies are designed and perceived primarily as functional tools or decision aids, especially in the present research context of AI career assessment. While I acknowledge the potential emotional and relational values AI can offer, competence forms the foundation for technologies to operate and produce outputs. This functional orientation suggests that people tend to weigh ability more heavily than benevolence or integrity when deciding whether the AI can be trusted. Prior research provides empirical support for this argument – Essay 1's meta-analysis revealed that each of the trustworthiness beliefs was, on average, positively related to trust in AI. Notably, the average effect size of ability ($\rho = .73$) was substantially larger than that of benevolence ($\rho = .23$) and integrity ($\rho = .20$).

Given the above arguments, I propose the following hypotheses:

> ***Hypothesis 4.*** *Perceived AI trustworthiness in terms of (a) ability, (b) benevolence, and (c) integrity is positively related to trust in AI, with perceived ability expected to have the strongest influence.*
> ***Hypothesis 5.*** *Perceived AI trustworthiness mediated the effect of (a) AI information disclosure and (b) AI information accuracy on trust in AI.*

Consequences of Trust in AI

Trust behavior, reflecting whether the users of AI adopt it for usage, accept its suggestions and recommendations, or rely on it to make decisions, is a critical consequence that develops from trustworthiness beliefs and trust attitudes. Research has found that higher levels of perceived trustworthiness and trust in AI were associated with more acceptance and reliance. For example, people who perceive AI as more trustworthy or place more trust in AI displayed a higher intention to follow AI's recommendations and a lower tendency to double-check AI's output (Gobel et al., 2022; Selten et al., 2023; Yokoi & Nakayachi, 2019).

Hypothesis 6. *Trust in AI is positively related to trust behavior.*

Another consequence that has been emphasized in research on technology acceptance and adoption is *trust appropriateness* – whether users' trust fits with the technology's capabilities, such as following a correct AI or *not* following an incorrect AI (Edmonds et al., 2019; J. D. Lee & See, 2004). Contrary to the *appropriate trust* are situations of *overtrust* (or, misuse) and *distrust/ undertrust* (or, disuse) – overtrust indicates that users follow an incorrect AI, while *distrust/ undertrust* reflects situations when users fail to follow a correct AI. It is less clear from the current literature about the determining factors for trust appropriateness; thus I do not hypothesize the direction of the effect of trust in AI on trust appropriateness:

Hypothesis 7. *Trust in AI is significantly related to trust appropriateness.*

Finally, I propose that trust in AI also affects users' behavioral intentions when foreseeing future interactions with AI. As indicated by Mayer and Davis (1999), such intentions reflect a willingness to engage in risk-taking activities with AI, such

as sharing sensitive information and permitting the trustee to control over issues that are important to the trustor (in our research context, choosing careers). For example, Shi et al. (2021) found that users' trust in an AI-based recommendation system led to higher intention to adopt it as a decision aid or delegate certain tasks (e.g., making travel plans) to AI. Thus, trust in AI is expected to be positively related to trust intentions regarding future interactions.

Hypothesis 8. *Trust in AI is positively related to future trust intentions.*

The overall theoretical framework of Essay 2 is presented in Figure 1.

Methods and Results

Sample and Procedure

I designed a 2 (disclosure vs. no disclosure) by 2 (information accuracy: high vs. low) between-subject experiment to test the hypotheses. Power analysis using G*Power suggests a minimum sample size of 128 to reach a medium effect size of .25 with sufficient statistical power of .80 at a significance level of .05.

Three hundred and seventy-six undergraduate students from a large Asian university were recruited to participate in our online experiment in exchange for one course credit. After removing 61 responses that failed the attention check, 315 (83.78%) were kept in the final sample for analysis. 61.6% of participants were female, 93% were Asian, and 55.9% were looking for a job at the time of the study. They had a mean age of 20.22 (SD = 1.58) and an average internship or work experience of 1.1 years (SD = 1.32). None of the participants had missing responses on the focal variables.

Figure 2 presents the overall study procedure. Upon study registration,

participants were briefed on a career assessment session mediated by an AI career advisor. Before the session, participants provided information about their demographics, personalities, and familiarity with AI and AI-mediated career assessments. Then, they were randomly assigned to one of four experimental conditions and directed to the corresponding online platform developed and hosted by the research team to complete the career assessment (see Figure 3 for sample screenshots). At the end of the session, participants received a career assessment report provided by the AI career advisor, including test results, an analysis of career interests, and four recommended jobs. Participants were asked to choose only one job from the recommendation list to obtain more detailed information and received the information after the choice. Upon session completion, participants completed a post-session survey on manipulation check questions, reason of choice, and their trust in the AI career advisor. Finally, participants will be debriefed about the purpose of the study. Throughout the study, a mouse-tracking algorithm was embedded into the webpage to track participants' viewing time and mouse trajectories.

Career Assessment Test

In real-life career assessment and feedback sessions, individuals typically undertake psychometric tests to identify their personal values and vocational preferences and explore suitable career options with guidance from career coaches or employers. These career options are identified by matching individuals' vocational interests with the task requirements and demands associated with specific jobs or vocations. One of the most widely used career assessments is based on Holland's RIASEC model of vocational interests, which categorizes individuals' vocational interests along six dimensions: realistic, investigative, artistic, social, enterprising, and conventional.

In this study, the short form of the *O*NET Interest Profiler (IP)* was employed as the career assessment test for two reasons. First, hosted on the *Occupation Information Network (O*NET)* website, IP is a widely recognized and freely accessible resource for career exploration, planning and coaching. Particularly, it is one of the recommended career assessment tools in the university and the country where our participants were located. Second, the 60-item IP short form is one of the shortest validated interest assessments available (Nye, 2022; Rounds et al., 2010). Compared to the 180-item full version, it reduces the cognitive burden on participants while maintaining acceptable response quality in an online, computerized setting. Participants will be shown 60 work activities, with 10 items representing each RIASEC type. For each work activity, participants rate their level of interest on a 5-point Likert scale (0 = "strongly dislike," 1 = "dislike," 2 = "unsure," 3 = "like," 4 = "strongly like"). Upon completion, participants received six scores, one for each RIASEC type, with a score range of 0 to 40.

To ensure the realism and functionality of the self-developed online platform, I invited seven academic researchers in the fields of organizational behavior, strategy, marketing, and information systems to go through the platform, and modify it based on their feedback and suggestions.

Career Assessment Report & Manipulations

A career assessment report includes the assessment results, manipulation information based on their assigned conditions, and a list of recommended jobs (Figure 3b-d). *Accuracy* was manifested in terms of the assessment results and job recommendations, and *disclosure* was manipulated in terms of the text paragraphs displayed after viewing assessment results and before viewing recommended jobs.

On the assessment result page, participants viewed a bar chart visualizing their

career assessment scores and brief descriptions of the six dimensions. Each participant's six RIASEC scores were summed from their responses and ranked to create a two-letter vocational interest profile, with the first letter indicating their highest-scoring dimensions. Recommended jobs were drawn systematically from the O*NET 28.0 database (see Appendix 1), using the relevant interest profile to select representative occupations corresponding to the highest-rated dimensions.

In the *high accuracy* condition, the two-letter vocational interest profile reflected participants' actual responses to the assessment test without any deviations. For example, a participant scoring highest in the Social (S) and Enterprising dimensions (E) and lowest in the Realistic (R) and Investigative dimensions (I) would receive scores that accurately represented their true assessment results (i.e., high in SE, low in RI). The participant would then be recommended representative jobs in the SE category, such as "Recreation Workers".

In contrast, in the *low accuracy* condition, prior to ranking and profile creation, the two highest-scoring and two lowest-scoring dimensions were intentionally swapped. Scores of the four middle-scoring dimensions were left unchanged. For instance, a participant scoring highest in SE and lowest in RI will be told they scored highest in RI and lowest in SE instead, which deviates from their earlier responses to the assessment test. The swapped ranking was used both for generating the bar chart visualization and for identifying the job recommendations. The participant would then be recommended representative jobs in the RI category, such as "Automotive Engineer". Notably, participants did not receive any explicit indication that the swapped rankings and vocational interest profile differed from their actual responses in the assessment test. This swapping process was conducted automatically by the experimental platform immediately after the assessment test was

completed and prior to the display of test results. All participants viewed only the manipulated scores and recommendations without access to raw item-level responses.

AI information disclosure will be manipulated as the existence of additional text paragraphs between assessment results and job recommendations (see Appendix 2). In the *disclosure* condition, participants will receive information about how the AI career advisor came up with the job recommendations before they view the four recommended jobs. In the *no disclosure* condition, participants will only view the four recommended jobs from the AI career advisor without receiving additional information.

Measures

All items were rated on a seven-point Likert scale (1 = *strongly disagree*, 7 = *strongly agree*).

AI Trustworthiness in terms of ability, benevolence, and integrity were each assessed with three items adapted from Mayer and Davis (1999) and McKnight (2002). Sample items include "The AI career advisor is very capable of providing career assessment and advice" for *ability* (α = .89), "I believe that the AI career advisor will act in my best interest" for *benevolence* (α = .93), and "The AI career advisor is sincere and genuine" for *integrity* (α = .77). In addition, I measured overall AI trustworthiness with three items from Pitardi and Marriott (2021). A sample item is "I feel that the AI career advisor is trustworthy" (α = .92).

Trust in AI was measured with two self-generated items, "Overall, I trust the AI career advisor" and "I trust the jobs recommended by the AI career advisor" (α = .94).

In this study, I operationalized *trust behavior* as whether participants chose the jobs recommended by the AI career advisor. This binary variable was coded as one if

participants chose so, and coded as zero if participants chose jobs that were not specifically recommended by the AI career advisor.

I operationalize *trust appropriateness* as the accuracy of choice. This binary variable was coded as one if participants chose the job reflecting their true vocational interests and coded as zero if not. In the two High-Accuracy conditions, the values of trust behaviors and trust appropriateness are equal.

Future trust intentions are measured in two terms – *willingness to act on AI's job recommendations* (e.g., "I would not hesitate to apply to the jobs the AI career advisor recommended to me"; α = .87) and *willingness to give information to the AI career advisor* in a follow-up session (e.g., "If there is a follow-up career advice session, I will be willing to provide my resume to the AI career advisor"; α = .78), each with two items adapted from McKnight et al. (2002).

In addition, I measured participants' age and gender (male = 0, female = 1). I also measured *familiarity with AI* and *propensity to trust*, as they were found by prior research to associate with our focal variables. Familiarity with AI was measured with five items developed by Yu et al. (2024). A sample item is "I have experience with AI" (α = .81). Propensity to trust was measured with four items (M. K. Lee & Turban, 2001; Xiang et al., 2022). A sample item is "It is easy for me to trust a person/thing" (α = .93).

Results

Descriptive Statistics

Table 3 presents the descriptive statistics, reliabilities, and correlations of focal variables. Perceived AI trustworthiness in terms of *ability*, *benevolence*, and *integrity* were all positively related to trust in AI ($r_{ability}$ = 73, $r_{benevolence}$ = 37, $r_{integrity}$ = .59; all p < .001), willingness to follow ($r_{ability}$ = .58, $r_{benevolence}$ = .35, $r_{integrity}$ = .49; all p < .001),

and willingness to give information ($r_{ability}$ = .44, $r_{benevolence}$ = .22, $r_{integrity}$ = .38, all p < .001). Only ability (r = .18, p = .002) and integrity (r = .15, p = .008) were positively associated with choice accuracy. Perceived AI trustworthiness was not significantly related to choice of AI recommendation. Trust attitude is positively correlated with all outcome variables, including choice of AI recommendation (r = .13, p = .023), choice accuracy (r = .21, p < .001), willingness to follow (r = .73, p < .001) and willingness to give information (r = .53, p < .001). Choice accuracy was positively related to subsequent willingness to follow (r = .17, p = .002) and willingness to give information (r = .14, p = .015), while choice of AI recommendation had no significant relationship with the two.

Confirmatory Factor Analysis

I conducted confirmatory factor analysis (CFA) in *R* (version 4.0.3, R Core Team, 2020) using the *lavaan* package (v.0.6.7) to verify our measurement model. Chi-square difference tests indicated that the six-factor model (i.e., three trustworthiness dimensions, trust in AI, willingness to follow, and willingness to give information) demonstrated a better fit to the data ($\chi^2_{(75)}$ = 140.07, *RMSEA* = .05, *CFI* = .98, *TLI* = .97, SRMR = .04) than several alternative models, showing the discriminant validity of the focal variables (see Table 2). All items loaded significantly and strongly on their intended factors (standardized loadings ranges from .70 to .95; see Appendix 3 Table A1). Inter-factor correlations ranged from .23 to .84, with several correlations exceeding .70[9], suggesting moderate to high shared variance between some constructs (see Appendix 3 Table A2).

To further examine the discriminant validity of key constructs, I used the

[9] Inter-factor correlation is .837 between perceived ability and integrity; .798 between perceived ability and trust in AI; and .808 between trust in AI and willingness to follow.

Fornell–Larcker criterion to compare the Average Variance Extracted (AVE) values for each factor to their squared correlations with other factors. Under this criterion, perceived integrity (AVE = .53) suffers from insufficient discriminant validity, as its AVE was lower than its squared correlation with perceived ability. Discriminant validity was supported for all other variables. The full matrix of inter-factor correlations and AVE is presented in Appendix 3.

Manipulation Check

To test the effectiveness of the manipulations of AI information disclosure and AI information accuracy, participants were asked to respond to four items for the Accuracy manipulation and three items for the Disclosure manipulation on a seven-point Likert scale (1 = *strongly disagree*, 7 = *strongly agree*). Sample items include "The job recommendations I received were accurate" for perceived AI information accuracy (α = .92) and "I received information from the AI career advisor about its underlying operating rules and inner logic" for perceived AI information disclosure (α = .83).

Results of independent samples t-test showed support for the effectiveness of both manipulations. Participants in the Disclosure condition perceived more AI information being disclosed (M_{dis} = 4.77, SD = 1.16) than those in the No Disclosure condition (M_{dis} = 4.50, SD = 1.32), albeit the mean difference did not reach statistical significance ($t_{(313)}$ = 1.93, p = .054). Participants in the High Accuracy conditions perceived a significantly higher level of AI information accuracy ($M_{highacc}$ = 4.77, SD = 1.27) than those in the Low Accuracy conditions (M_{lowacc} = 3.66, SD = 1.41; $t_{(313)}$ = 7.35, p < .001).

Analytic Strategy

I conduct all hypothesis test analyses in SPSS (version 29.0.2.0). H1 and H2 were tested with independent samples t-test. H3 was tested with the two-way Analysis of Variance (ANOVA). H4 and H8 were tested using linear regression analyses, while H6 and H7 were tested with binary logistic regression analyses. Finally, H5 was tested using SPSS PROCESS Macro (Hayes, 2012).

Hypothesis Tests

Regarding Hypothesis 1, AI information disclosure did not show a significant effect on perceived AI trustworthiness in terms of ability (ΔM = .10, $t_{(313)}$ = .68, p = .500), benevolence (ΔM = .00, $t_{(313)}$ = .02, p = .988), integrity (ΔM = .10, $t_{(313)}$ = .94, p = .346). These results support the alternative of Hypothesis 1, suggesting that information disclosure does not significantly relate to perceived trustworthiness.

Supporting Hypothesis 2a and 2c, participants in the High Accuracy condition perceived AI career advisor to possess higher ability (ΔM = .57, $t_{(313)}$ = 4.17, p < .001) and integrity (ΔM = .48, $t_{(313)}$ = 4.50, p < .001) than those in the Low Accuracy condition. However, the effect of Accuracy on benevolence (ΔM = .29, $t_{(313)}$ = 1.91, p = .056) did not reach statistical significance.

Hypothesis 3 did not receive support, as the interaction between AI information disclosure and accuracy was not significant in affecting perceptions of AI ability ($F_{(1, 311)}$ = 1.19, p = .277, partial η^2 = .004), benevolence ($F_{(1, 311)}$ = 0.55, p = .460, partial η^2 = .002), and integrity ($F_{(1, 311)}$ = 0.42, p = .518, partial η^2 = .001). Although the descriptive means suggested that disclosure slightly reduced the gap between High Accuracy and Low Accuracy conditions across all three trustworthiness dimensions, these differences were not statistically significant.

Results of linear regression analyses indicated that perceived AI trustworthiness in ability (B = .75, SE = .06, p < .001) and benevolence (B = .13, SE

= .05, p = .004) displayed significant and independent effects on trust in AI, whearas the effect of integrity did not reach statistical significance (B = .15, SE = .08, p = .080). The three trustworthiness beliefs together explained 56.4 percent of the variance in trust in AI. Pairwise comparisons of the regression coefficients further indicated that the effect of ability was significantly larger than the effect of both benevolence (t = 7.94, p < .001) and integrity (t = 6.00, p < .001). H4 is thus generally supported.

The mediating effect of AI trustworthiness was tested using the SPSS PROCESS macro with 5,000 iterations. Results revealed that, among the three AI trustworthiness dimensions, only perceived *ability* mediated the effect of AI information accuracy on trust in AI (indirect effect = .41, SE = .11, CI = [.22, .63]). No mediating effect was observed for the relationship between AI information disclosure and trust in AI. Thus, H5 is partially supported.

Binary logistic regressions showed that trust in AI had a significant and positive effect on predicting trust behavior (B = .19, Wald statistic = 5.29, Odds Ratio = 1.21) and trust appropriateness (B = .31, Wald statistic = 14.35, Odds Ratio = 1.37). With a one-unit increase in participants' trust in AI, the odds of participants choosing an AI-recommended job rather than a non-AI-recommended job will increase by 1.21 times, and the odds of participants displaying appropriate trust rather than over- or under-trust will increase by 1.37 times. These results provide support for H6 and H7.

Finally, H8 and H9 were supported, as trust in AI is significantly and positively related to participants' willingness to follow AI's job recommendations (B = .71, SE = .04, p < .001) as well as give information to the AI career advisor in a subsequent session (B = .51, SE = .05, p < .001).

The results are summarized in Figure 4.

Supplementary Analysis

In addition to the hypothesized relationships argued above, I conducted several exploratory analyses, which uncovered interesting insights.

Additional Metrics of Viewing Behavior. The use of a self-developed and self-hosted website enabled us to track participants' viewing behaviors when they were going through the career assessment session. Specifically, I tracked participants' time spent on viewing (a) the career assessment results, (b) disclosed information (for those in the Disclosure conditions), (c) the job recommendation page, and (d) each of the recommended jobs. On average, participants in the Disclosure conditions spent 1.17 minutes viewing the manipulation paragraph, and spent more time viewing AI-recommended jobs (ΔM = 35.68s, $t_{(313)}$ = 1.81, p = .071) than those in the No Disclosure conditions. Independent samples t-tests showed that participants who received accurate information from AI spent significantly more time viewing AI-recommended jobs (ΔM = 47.77s, $t_{(313)}$ = 2.44, p = .015), while spending less time viewing non-AI-recommended jobs (ΔM = 21.25s, $t_{(313)}$ = 1.95, p = .053) than their counterparts.

Mediation Effects of Perceived Trustworthiness and Trust in AI. Building upon H5, I conducted Structural Equation Modeling (SEM) analyses to explore whether perceived trustworthiness and trust in AI sequentially mediated the effect of disclosure and accuracy on key outcomes. The theorized model (see Figure 1) showed good fit to the data ($\chi^2_{(14)}$ = 32.58, *RMSEA* = .07, *CFI* = .98, *TLI* = .95, *SRMR* = .04). Results indicated that AI information accuracy had a significantly indirect effect on (a) willingness to follow AI's recommendations (β = 0.11, $p < .001$) and (b) willingness to give information to the AI (β = 0.08, $p < .001$), manifesting through

perceived ability and trust in AI sequentially. The summary of the SEM model is presented in Appendix 4.

Moderating Effect of Individual Characteristics. I explore the potential moderating role of four variables representing individual demographics (age, gender), traits (propensity to trust), and expertise (familiarity with AI), as these have been proposed by previous models (Venkatesh et al., 2003) to be potential moderators affecting technology acceptance and trust formation. I conducted the moderating analyses using two-way ANOVA and SPSS PROCESS Macro. Results suggested moderating effects of gender and familiarity with AI. The plots of estimated marginal means are provided in Appendix 5.

Gender. Interestingly, the effect of AI information disclosure was significantly positive for males and negative for females (yet non-significant) in forming ability beliefs ($F_{(1,310)}$ = 7.07, p =.008) and willingness to give information ($F_{(1,310)}$ = 8.26, p =.004). On the other hand, the effect of AI information accuracy was positive for both genders, but more positive for females than for males in forming ability beliefs ($F_{(1,310)}$ = 6.55, p =.011), integrity beliefs ($F_{(1,310)}$ = 4.06, p =.045), and trust in AI ($F_{(1,310)}$ = 6.91, p = .009). The conditional effects for females were all positive and significant, while positive yet non-significant for males.

Familiarity with AI. The effect of AI information accuracy on willingness to follow AI recommendations ($F_{(1,310)}$ = 2.94, p = .08) and willingness to give information ($F_{(1,310)}$ = 7.35, p = .007) were both positive when participants had average or higher familiarity with AI but not significant when participants had lower familiarity with AI.

Discussion

This study provides a primary test of two facets theorized in the framework of AI transparency – AI information disclosure and AI information accuracy, unpacking their effects in forming trustworthiness beliefs, trust attitude, and subsequent behavioral responses to the task at hand or to future possible interactions. In the context of career assessment and recommendation, the experimental results showed a significant and positive impact of AI information accuracy on trustworthiness perceptions, while a very limited impact on AI information disclosure.

It appeared that mere disclosure of information about the inner logic and operating rules of AI did not help generate more favorable trust-related responses in users. Although contrary to common thoughts, this finding is consistent with some of the prior research findings on AI transparency (Bayer et al., 2021; Leichtmann et al., 2023). One possible reason may be that participants were exposed to a large amount of information during the career assessment session, which makes it challenging for participants to distinguish the manipulated information and other information provided during the task. For example, in addition to the manipulated information about how the AI career advisor works that was displayed on a separate page, participants were also exposed to an explanation of each vocational interest dimension, a one-sentence description of each recommended job, and an introduction to the chosen job after they made the job choice. Such information, although not intentionally designed to be part of the manipulation, may still be processed by participants as information disclosed by AI. This issue will also appear in real-life scenarios, raising a need for future research to explore ways to distinguish the disclosed information from other information and make it salient to AI users.

Supporting our hypotheses, the accuracy of AI information was steadily helpful in cultivating trustworthiness perceptions and trust in AI, which in turn led to

more adoption of AI recommendations, higher choice accuracy, and higher intentions to interact with AI post-task. Such an effect did not vary depending on the Disclosure manipulation. In addition, ability belief was found to be the only mediator transferring the effect of accuracy to trust in AI. The important role of accuracy echoes with prior research that placed a higher emphasis on AI's capability (e.g., Mcknight et al., 2011). One of the hypotheses (H7) centers around the concepts of appropriate trust, overtrust, and undertrust. Our study results showed a moderately positive correlation between accuracy manipulations and accuracy of job choice ($r = .43$, $p < .001$). A further look at the crosstab analysis revealed that, descriptively, 85 out of the 315 participants fell into the overtrust category, and 69 out of the 315 participants fell into the undertrust category. This gives rise to an interesting future research direction regarding how overtrust and undertrust can be mitigated.

It is also worth noting that the results of supplementary analyses hinted at the role of individual characteristics as a contingent factor for AI information disclosure and accuracy to manifest their effects. Primary results suggested that AI information disclosure may work better for males, while AI information accuracy may work better with females. Interestingly, people with more experience and expertise with AI reacted more favorably to manipulations of information disclosure and information accuracy. One reason may be that their familiarity with AI enables them to identify and appreciate the functional value of the disclosed information and accurate information. It might also be because people more familiar with AI will weigh information quality more heavily than those who are less familiar with AI. These patterns indicate some research questions for further research to explore.

References

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211.

Alam, L., & Mueller, S. (2021). Examining the effect of explanation on satisfaction and trust in AI diagnostic systems. *BMC Medical Informatics & Decision Making*, *21*(1), 1–15. https://doi.org/10.1186/s12911-021-01542-6

Ashoori, M., & Weisz, J. D. (2019). In AI we trust? Factors that influence trustworthiness of AI-infused decision-making processes. *arXiv Preprint:1912.02675*.

Bayer, S., Gimpel, H., & Markgraf, M. (2021). The role of domain expertise in trusting and following explainable AI decision support systems. *Journal of Decision Systems*, *32*(1), 110–138. https://doi.org/10.1080/12460125.2021.1958505

Benbasat, I., & Wang, W. (2005). Trust in and adoption of online recommendation agents. *Journal of the Association for Information Systems*, *6*(3), 4. https://doi.org/10.17705/1jais.00065

Berger, C. R. (1986). Uncertain outcome values in predicted relationships: Uncertainty reduction theory then and now. *Human Communication Research*, *13*(1), 34–38. https://doi.org/10.1111/j.1468-2958.1986.tb00093.x

Berger, C. R., & Calabrese, R. J. (1975). Some explorations in initial interaction and beyond: Toward a developmental theory of interpersonal communication. *Human Communication Research*, *1*(2), 99–112. https://doi.org/10.1111/j.1468-2958.1975.tb00258.x

Bigras, E., Jutras, M. A., Sénécal, S., Léger, P. M., Black, C., Robitaille, N., ... & Hudon, C. (2018). In AI we trust: Characteristics influencing assortment

planners' perceptions of AI based recommendation agents. In *HCI in Business, Government, and Organizations: 5th International Conference, HCIBGO 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15-20, 2018, Proceedings 5* (pp. 3-16). Springer International Publishing.

Blackman, R., & Ammanath, B. (2022). Building transparency into AI projects. *Harvard Business Review, 22*(6). https://hbr.org/2022/06/building-transparency-into-ai-projects

Chanda, T., Hauser, K., Hobelsberger, S., Bucher, T. C., Garcia, C. N., Wies, C., ... & Brinker, T. J. (2024). Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma. *Nature Communications*, *15*(1), 524.

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

De Freitas, J., Agarwal, S., Schmitt, B., & Haslam, N. (2023). Psychological factors underlying attitudes toward AI tools. *Nature Human Behaviour*, *7*(11), 1845–1854. https://doi.org/10.1038/s41562-023-01734-2

DeLone, W. H., & McLean, E. R. (1992). Information systems success: The quest for the dependent variable. *Information Systems Research*, *3*(1), 60–95.

Delone, W. H., & McLean, E. R. (2003). The DeLone and McLean model of information systems success: A ten-year update. *Journal of Management Information Systems*, *19*(4), 9–30.

DeLone, W. H., & McLean, E. R. (2004). Measuring e-Commerce Success: Applying the DeLone & McLean Information Systems Success Model. *International Journal of Electronic Commerce*, *9*(1), 31–47.

https://doi.org/10.1080/10864415.2004.11044317

Edmonds, M., Gao, F., Liu, H., Xie, X., Qi, S., Rothrock, B., Zhu, Y., Wu, Y. N., Lu, H., & Zhu, S.-C. (2019). A tale of two explanations: Enhancing human trust by explaining robot behavior. *Science Robotics*, *4*(37), eaay4663. https://doi.org/10.1126/scirobotics.aay4663

Elofson, G. (2001). Developing trust with intelligent agents: An exploratory study. In C. Castelfranchi & YH. Tan (Eds.), *Trust and Deception in Virtual Societies* (pp. 125–138). Springer. https://doi.org/10.1007/978-94-017-3614-5_6

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660.

Gobel, K., Niessen, C., Seufert, S., & Schmid, U. (2022). Explanatory machine learning for justified trust in human-AI collaboration: Experiments on file deletion recommendations. *Frontiers in Artificial Intelligence*, *5*, 919534. https://doi.org/10.3389/frai.2022.919534

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517–527.

Hayes, A. F. (2012). *PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling*. University of Kansas, KS.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Kim, J., Giroux, M., & Lee, J. C. (2021). When do you trust AI? The effect of number presentation detail on consumer trust and acceptance of AI recommendations. *Psychology & Marketing*, *38*(7), 1140–1155. https://doi.org/10.1002/mar.21498

Kyung, N., & Kwon, H. E. (2022). Rationally trust, but emotionally? The roles of

cognitive and affective trust in laypeople's acceptance of AI for preventive care operations. *Production & Operations Management*, 10591478231225891. https://doi.org/10.1111/poms.13785

Langer, M., & König, C. J. (2023). Introducing a multi-stakeholder perspective on opacity, transparency and strategies to reduce opacity in algorithm-based human resource management. *Human Resource Management Review*, *33*(1), 100881. https://doi.org/10.1016/j.hrmr.2021.100881

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

Lee, M. K., & Turban, E. (2001). A trust model for consumer internet shopping. *International Journal of Electronic Commerce*, *6*(1), 75–91.

Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., & Mara, M. (2023). Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior*, *139*, 107539. https://doi.org/10.1016/j.chb.2022.107539

Lind, E. A., & Van den Bos, K. (2002). When fairness works: Toward a general theory of uncertainty management. *Research in Organizational Behavior*, *24*, 181–223.

Liu, B. J. (2021). In AI we trust? Effects of agency locus and transparency on uncertainty reduction in human-AI interaction. *Journal of Computer-Mediated Communication*, *26*(6), 384–402. https://doi.org/10.1093/jcmc/zmab013

Lundberg, H., Mowla, N. I., Abedin, S. F., Thar, K., Mahmood, A., Gidlund, M., & Raza, S. (2022). Experimental analysis of trustworthy in-vehicle intrusion detection system using eXplainable Artificial Intelligence (XAI). *IEEE Access*,

10, 102831–102841. https://doi.org/10.1109/ACCESS.2022.3208573

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.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709–734.

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

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

Min, F., Zou, F., He, Y., & Jiang, X. (2021). Research on users' trust of chatbots driven by AI: An empirical analysis based on system factors and user characteristics. In *2021 IEEE International Conference on Consumer Electronics and Computer Engineering* (pp. 55–58). IEEE. https://doi.org/10.1109/ICCECE51280.2021.9342098

Nakashima, H., Mantovani, D., & Machado, C. (2022). Users' trust in black-box machine learning algorithms. *Revista de Gestão*, *31*(2), 237–250. https://doi.org/10.1108/REGE-06-2022-0100

Nye, C. D. (2022). Assessing interests in the twenty-first-century workforce: Building on a century of interest measurement. *Annual Review of Organizational Psychology and Organizational Behavior*, *9*, 415–440.

Pitardi, V., & Marriott, H. R. (2021). Alexa, she's not human but... Unveiling the drivers of consumers' trust in voice-based artificial intelligence. *Psychology &*

Marketing, *38*(4), 626–642. https://doi.org/10.1002/mar.21457

Rounds, J., Su, R., Lewis, P., & Rivkin, D. (2010). *O*NET Interest Profiler Short Form Psychometric Characteristics: Summary*.

Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors*, *58*(3), 377–400.

Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*, *29*(4), 260–278.

Schnackenberg, A. K., & Tomlinson, E. C. (2016). Organizational transparency: A new perspective on managing trust in organization-stakeholder relationships. *Journal of Management*, *42*(7), 1784–1810.

Schnackenberg, A. K., Tomlinson, E., & Coen, C. (2021). The dimensional structure of transparency: A construct validation of transparency as disclosure, clarity, and accuracy in organizations. *Human Relations*, *74*(10), 1628–1660. https://doi.org/10.1177/0018726720933317

Selten, F., Robeer, M., & Grimmelikhuijsen, S. (2023). "Just like I thought": Street-level bureaucrats trust AI recommendations if they confirm their professional judgment. *Public Administration Review*, *83*(2), 263–278. https://doi.org/10.1111/puar.13602

Shi, S., Gong, Y., & Gursoy, D. (2021). Antecedents of trust and adoption intention toward artificially intelligent recommendation systems in travel planning: A heuristic–systematic model. *Journal of Travel Research*, *60*(8), 1714–1734. https://doi.org/10.1177/0047287520966395

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-*

Computer Studies, *146*, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Shin, D., Zaid, B., Biocca, F., & Rasul, A. (2022). In platforms we trust? Unlocking the black-box of news algorithms through interpretable AI. *Journal of Broadcasting & Electronic Media*, *66*(2), 235–256.

Suen, H.-Y., & Hung, K.-E. (2023). Building trust in automatic video interviews using various AI interfaces: Tangibility, immediacy, and transparency. *Computers in Human Behavior*, *143*, 107713. https://doi.org/10.1016/j.chb.2023.107713

Tomlinson, E. C., & Schnackenberg, A. (2022). The effects of transparency perceptions on trustworthiness perceptions and trust. *Journal of Trust Research*, *12*(1), 1–23. https://doi.org/10.1080/21515581.2022.2060245

Tuncer, S., & Ramirez, A. (2022). Exploring the role of trust during human-AI collaboration in managerial decision-making processes. *24th International Conference on Human-Computer Interaction, HCII 2022, 13518 LNCS*, 541–557. https://doi.org/10.1007/978-3-031-21707-4_39

Van Den Bos, K. (2009). Making sense of life: The existential self trying to deal with personal uncertainty. *Psychological Inquiry*, *20*(4), 197–217. https://doi.org/10.1080/10478400903333411

Venkatesh, V., Thong, J. Y. L., Chan, F. K. Y., & Hu, P. J. H. (2016). Managing citizens' uncertainty in e-government services: The mediating and moderating roles of transparency and trust. *Information Systems Research*, *27*(1), 87–111. https://doi.org/10.1287/isre.2015.0612

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425–478. https://doi.org/10.2307/30036540

Wang, B., Rau, P.-L., & Yuan, T. (2022). Measuring user competence in using artificial intelligence: Validity and reliability of artificial intelligence literacy scale. *Behaviour & Information Technology*, *42*, 1–14. https://doi.org/10.1080/0144929X.2022.2072768

Wu, J.-J., Khan, H. A., Chien, S.-H., & Wen, C.-H. (2022). Effect of customization, core self-evaluation, and information richness on trust in online insurance service: Intelligent agent as a moderating variable. *Asia Pacific Management Review*, *27*(1), 18–27. https://doi.org/10.1016/j.apmrv.2021.04.001

Xiang, H., Zhou, J., & Xie, B. (2022). AI tools for debunking online spam reviews? Trust of younger and older adults in AI detection criteria. *Behaviour & Information Technology*, 1–20. https://doi.org/10.1080/0144929x.2021.2024252

Xiao, B., & Benbasat, I. (2007). E-commerce product recommendation agents: Use, characteristics, and impact. *MIS Quarterly, 31*(1), 137–209.

Yokoi, R., Eguchi, Y., Fujita, T., & Nakayachi, K. (2021). Artificial Intelligence is trusted less than a doctor in medical treatment decisions: Influence of perceived care and value similarity. *International Journal of Human-Computer Interaction*, *37*(10), 981–990. https://doi.org/10.1080/10447318.2020.1861763

Yokoi, R., & Nakayachi, K. (2019). The effect of shared investing strategy on trust in artificial intelligence. *Japanese Journal of Experimental Social Psychology*, *59*(1), 46–50. https://doi.org/10.2130/jjesp.1819

Yu, K. Y. T., Goh, K. H., Yu, S., & Wu, T. (2024, April). Attitude towards AI interviewing: Scale development and validation. *The 2024 SIOP Annual Conference*, Chicago, IL, United States.

Yu, L., & Li, Y. (2022). Artificial intelligence decision-making transparency and employees' trust: The parallel multiple mediating effect of effectiveness and

discomfort. *Behavioral Sciences*, *12*(5), 127. https://doi.org/10.3390/bs12050127

Yu, S., Kawasaki, S., & Yu, K. Y. T. (2023). Trust in artificial intelligence (AI): An integrative and meta-analytic review. *Academy of Management Proceedings*, *2023*(1), 11136.

Zhang, Y., Vera Liao, Q., & Bellamy, R. K. E. (2020). *Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making*. 295–305. https://doi.org/10.1145/3351095.3372852

Zhou, J., Hu, H., Li, Z., Yu, K., & Chen, F. (2019, August). Physiological indicators for user trust in machine learning with influence enhanced fact-checking. In *3rd International Cross-Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE)* (pp. 94-113). Springer International Publishing. https://doi.org/10.1007/978-3-030-29726-8_7

Tables and Figures

Table 1. Summary of Essay 1 articles on AI transparency & trust in AI

Source	Variable	Description	Transparency-Related Facets					Type of AI
			Disclosure	Accuracy	Clarity	Personalization	Others	
Glikson & Woolley (2020)	Transparency	The level to which the underlying operating rules and inner logics of the technology are *apparent* to the users (p.631)	√		√			General
DeFreitas et al. (2023)	Transparency	The degree to which the internal mechanics of a system are *observable* and *understandable* by humans (p.1847)	√		√			General
Möhlmannn et al. (2023)	Algorithmic transparency	The *ease* of platform workers (as in their perceptions) to observe how the input data and inner workings of the algorithm affect algorithmic outputs.	√					Algorithm management
	Algorithmic opacity	The *difficulty* of platform workers (as in their perceptions) to observe how the input data and inner workings of the algorithm affect algorithmic outputs	√					
Alam & Mueller (2021)	Explanation	Explanation about AI's prediction and diagnosis (global: in general vs local: for specific cases)	√				Content	AI diagnosis chatbot
Angerschmid et al. (2022)	Explanation	Whether explanations are provided, and whether they are based on examples or important features					Content	Health insurance decision-making
Ashoori & Weisz (2019)	Interpretability	The extent to which the process by which the model arrived at a recommendation can be *examined* and *understood*	√		√			ADSS
	Confidence score	Whether the *confidence* a model has in its recommendation is *visible*	√					
Bayer et al. (2021)	Explanation	Provision of justifications for AI's suggestions	√				Content	ADSS
Bigras et al. (2018)	Information richness	The *amount* of information provided to assist decision-making	√				Richness	RA
Chanda et al. (2023)	XAI	Provision of explanations that close the interpretation gap and can be *easily interpreted*	√		√			AI medical diagnosis
	Display of XAI confidence	XAI's communicated *confidence*	√	√				

Source	Variable	Description	Transparency-Related Facets					Type of AI
			Disclosure	Accuracy	Clarity	Personalization	Others	
Elofson (2001)	Intelligent agent-generated decision heuristics	Provision of the decision heuristics of the intelligent agent	√					IA
Gobel et al, (2022)	Explanation	Provision of explanations about why a suggestion is made	√					ADSS
	Information uncertainty	The extent to which why a suggestion is made is *uncertain*	√					
	Credibility	The extent to which the information provided is *credible*		√				
Kim et al. (2021)	Accuracy of information	Provision of information about the accuracy rate of the AI system	√	√				RA
	Preciseness of information	Provision of the *preciseness* of the accuracy rate	√				Precise-ness	
Kyung & Kwon (2022)	AI transparency	Provision of explanations about how AI generates the recommendation	√					Health intervention
Leichtmann et al. (2023)	Explanation	Provision of visual explanations about AI's decisions to users	√					ADSS
	Educational intervention	Provision of information that educates users to *better comprehend* AI technology, such as how AI works	√		√			
Liu (2021)	Real transparency	Provision of the rationales for AI-made decisions	√					AI for fake news detection
	Placebic transparency	Provision of statements that are *tautological* to the AI-made decision without actual explanations	√				Placebo	
Lundberg et al. (2022)	Explanation	Explanations based on visuals or rules	√				Content	AI-based in-vehicle intrusion detection
	Interpretability	Whether the AI model's mapping from input to output is *hard to understand*			√			
Min et al. (2021)	Perceived Personalization	The extent to which the information delivered to the receivers is *personalized* according to their unique preferences				√		AI chatbot
Nakashima et al. (2022)	Explanation artifacts	Provision of explanations that clarify the mental models behind the analytical model	√					AI finance advisor
Nasirian et al. (2017)	Information quality	Perceived quality of the provided information					Quality	AI voice assistant system

Source	Variable	Description	Transparency-Related Facets					Type of AI
			Disclosure	Accuracy	Clarity	Personalization	Others	
Ribes et al. (2021)	Transparent AI/XAI	Provision of *understandable* justifications for algorithm outputs	√		√			AI news content aggregator
	Detail of explanation	Level of details provided in the explanation	√				Richness	
Sassmannshausen et al. (2021)	Perceived comprehensibility	Provision of explanations to help users *understand* the AI's decision	√		√			RA
Schmidt et al. (2020)	Explanation	Provision of explanation of AI's prediction (by highlighting decisive words in the texts)	√					ML-based DSS
	Confidence score	Provision of a score on the tool's classification *confidence*	√	√				
Selten et al. (2023)	XAI	Provision of *understanding* to non-technical audiences by answering the why-question	√		√			ADSS
Shi et al. (2021	Perceived personalization	The extent to which an AI-based recommendation system understands and represents users' *personal interests*				√		RA
Shin (2021)	XAI	Machine learning and AI technologies that can offer *human-understandable* justifications for their output or procedures (Gunning et al., 2019)	√		√			AI news recommendation
	Explainability	The ability to explain *how* an algorithm works in order to understand how and *why* it has delivered particular outcomes	√					
	Causality	The extent to which an explanation of a statement to a user achieves a specified level of *causal understanding* with effectiveness, efficiency, and satisfaction	√		√		Causality	
Suen & Hung (2023)	Transparency	Same as Glikson & Woolley's (2020) definition	√		√			AI-based video job interviews
Tuncer et al. (2022)	Interpretability	Whether it is possible for users to *examine* and *comprehend* the interaction by which the model achieves the suggestion	√		√			ADSS
	Confidence score	The degree of *confidence* of the model	√	√				
Woodcock et al. (2021)	Explanation	Provision of the causal history of an event; an explainer explains something to generate *understanding* in a recipient	√		√			AI medical chatbot
Wu et al. (2022)	Customization	The extent to which product recommendations are based on customer preferences and demands				√		IA

Source	Variable	Description	Transparency-Related Facets					Type of AI
			Disclosure	Accuracy	Clarity	Personalization	Others	
Xiang et al. (2022)	Explanation	AI's detection criteria (e.g., based on the textual features of reviews or the behavioral features of reviewers)	√					AI-based spam detection tool
Xu et al. (2022)	Explanation	Provision of explanations of the potential benefits and costs of accepting the system	√					AI in HRM
Yang et al. (2020)	Explanation	Provision of explanations of the machine learning classification based on examples	√				Content	ADSS
Yokoi et al. (2021)	Perceived uniqueness neglected	Neglection of one's unique characteristics and symptoms in AI's decision-making				√		AI medical diagnosis
Yu & Li (2022)	Transparency	The degree to which humans *understand* the inner workings or logic of a technology			√			AI task assignment in organizations
	AI decision-making Transparency	The degree to which an AI system releases *objective* information about its working mode	√					
	Perceived transparency	The availability of *subjectively* perceived information	√					
Zhang et al. (2020)	Explanation	Explanations of contributing attributes to the model's prediction	√					ADSS
	Confidence score	*Confidence* level of the model		√				ADSS
Zhou et al. (2019)	Explanation	Presentation of influence of training data points to the model	√					ML

Notes. XAI refers to explainable AI; ADSS refers to AI-based decision support systems; RA refers to recommendation agents; IA refers to intelligent agents.

Table 2. Results of confirmatory factor analysis

Model	*χ2*	*df*	*p (χ2)*	*χ2/df*	*CFI*	*TLI*	*RMSEA*	*RMSEA 90%CI*	*p- RMSEA<0.05*	*SRMR*
Six-factor model	140.07	75	.00	1.87	.981	.974	.052	[.039, .066]	.365	.036
Three-factor model	930.55	87	.00	10.70	.755	.704	.175	[.165, .186]	.000	.113
Two-factor model	1163.60	89	.00	13.07	.688	.632	.196	[.186, .206]	.000	.118
One-factor model	1304.61	90	.00	14.50	.647	.588	.207	[.197, .217]	.000	.126

Note. In the six-factor model, three Trustworthiness variables, Trust in AI, Willingness to Follow, and Willingness to Give Information were each loaded on one factor. In the three-factor model, three Trustworthiness variables, Trust in AI, and the two Outcomes were each loaded on one factor. In the two-factor model, three Trustworthiness variables and Trust in AI were loaded on one factor, and the two Outcomes were loaded on another factor. In the one-factor model, all six variables were loaded on one factor.

Table 3. Means, standard deviations, reliabilities and correlations among variables

Variables		*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	Accuracy Dummy	0.51	0.50	/															
2	Disclosure Dummy	0.50	0.50	.03	/														
3	Perceived Accuracy	4.23	1.45	.38***	.04														
4	Perceived Disclosure	4.63	1.25	.17**	.11	.46***													
5	Ability Beliefs	4.68	1.25	.23***	.04	.60***	.45***	*(.92)*											
6	Benevolence Beliefs	3.97	1.34	.11	.00	.16**	.26***	.31***	*(.89)*										
7	Integrity Beliefs	5.00	0.98	.25***	.05	.46***	.37***	.70***	.43***	*(.93)*									
8	Trust in AI	4.12	1.48	.29***	.00	.69***	.42***	.73***	.37***	.59***	*(.77)*								
9	Choice of AI Recommendation	0.71	0.46	.34***	-.01	.13*	.02	.06	-.05	.07	.13*	/							
10	Choice Accuracy	0.66	0.48	.43***	.03	.31***	.071	.18***	.01	.15**	.22***	-.13*	/						
11	Willingness to Follow	3.34	1.44	.17**	.01	.61***	.36***	.58***	.35***	.49***	.73***	.07	.17**	*(.87)*					
12	Willingness to Give Information	4.02	1.42	.13*	.02	.42***	.27***	.44***	.22***	.39***	.53***	.10	.14*	.56***	*(.78))*				
13	Age	2.22	1.58	-.07	.00	-.02	-.07	-.02	-.06	-.04	-.05	-.03	-.01	.03	.01	/			
14	Gender	0.62	0.49	.03	.03	.03	.01	.07	-.07	-.03	.05	.11*	-.05	.01	-.01	-.51***	/		
15	Propensity to Trust	4.10	1.46	.01	.05	.15**	-.02	.10	.08	.04	.19***	-.08	.05	.19***	.19***	.06	-.09	*(.93)*	
16	Familiarity with AI	4.78	1.00	.07	-.02	.12*	.08	.10	.07	.11*	.10	.09	.03	.09	.25**	-.13*	.06	.15**	*(.81)*

Note. $N = 315$. $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$; two-tailed. Cronbach's alphas were presented in parentheses on the diagonal. Accuracy Dummy and Disclosure Dummy refer to dummies of accuracy and disclosure manipulations. For gender, male = 0, female = 1. AI refers to artificial intelligence; AICA refers to AI career assessment.

Figure 1. Theoretical framework of Essay 2

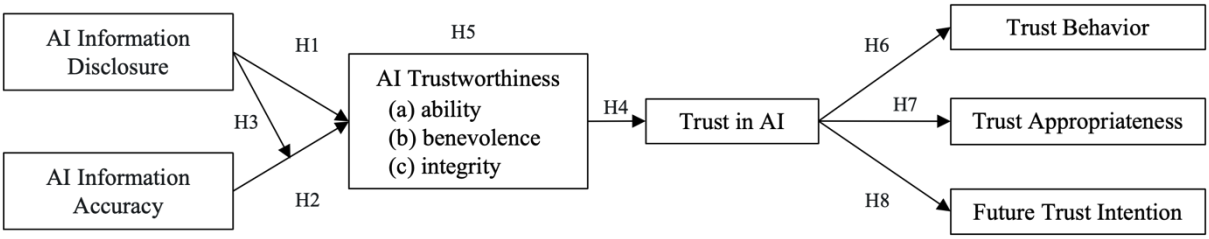


Figure 2. Study procedure

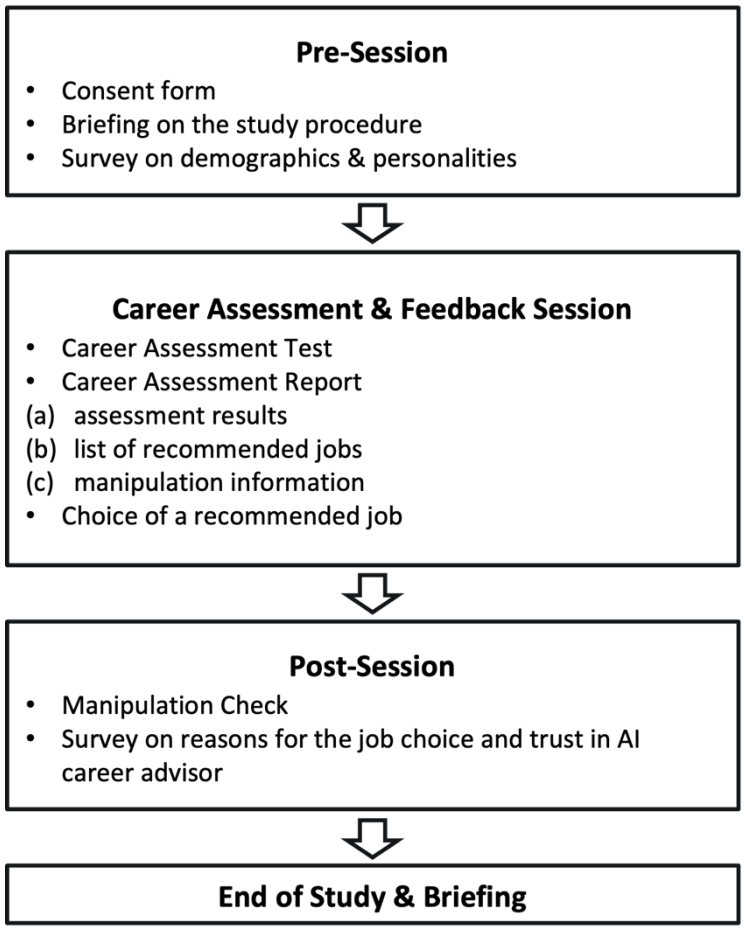
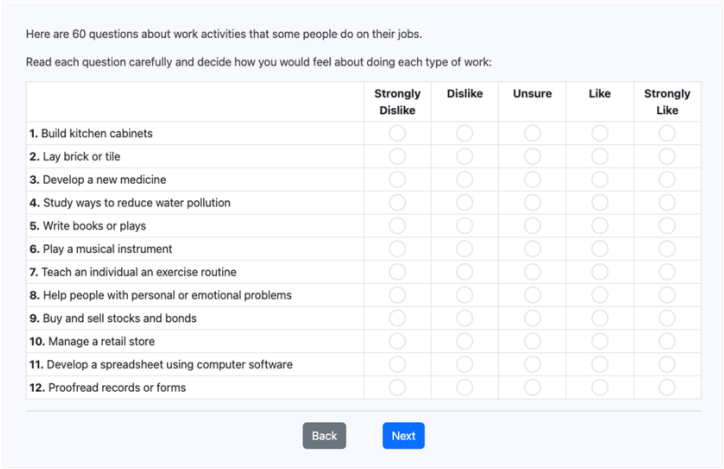
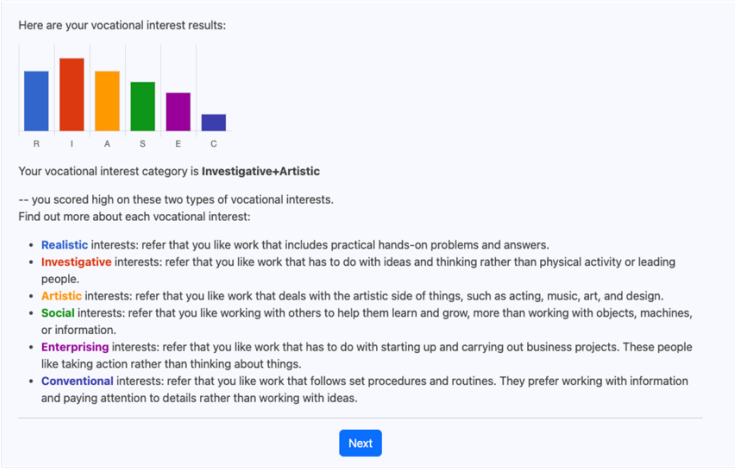


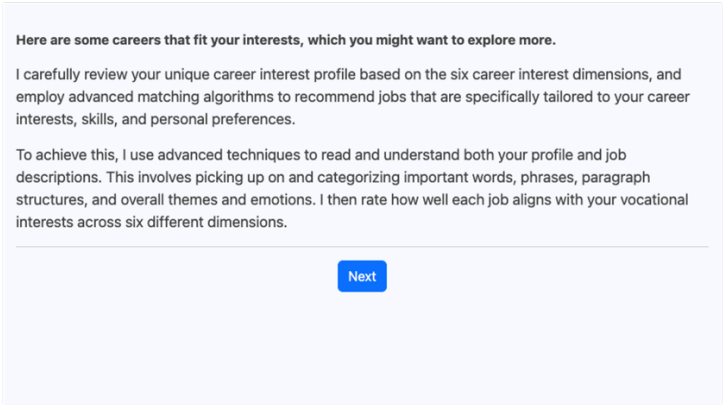
Figure 3. Samples of the career assessment interface



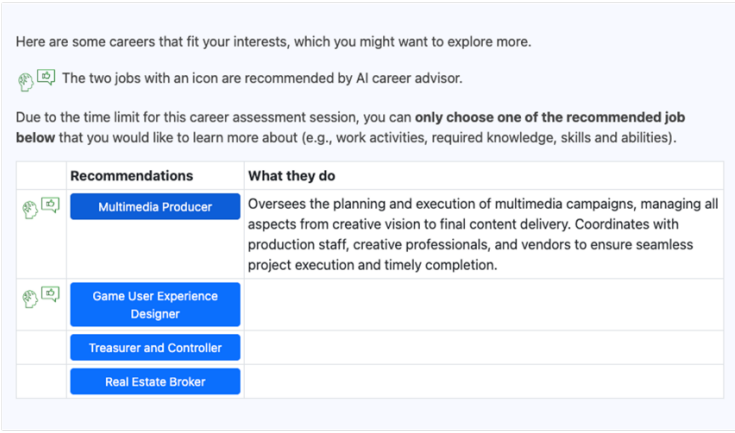
(a) Career Assessment Test



(b) Career Assessment Results

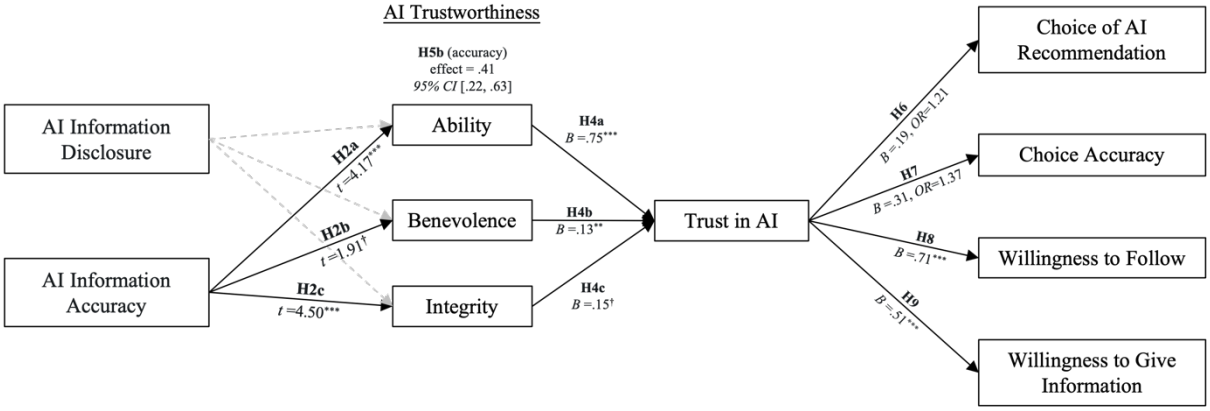


(c) Manipulation Information



(d) Job Recommendations & Choice

Figure 4. Summary of hypothesis test results



Note. $N = 315$. $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$, $^{\dagger}p < .10$; two-tailed. Only significant results were present.

Appendices

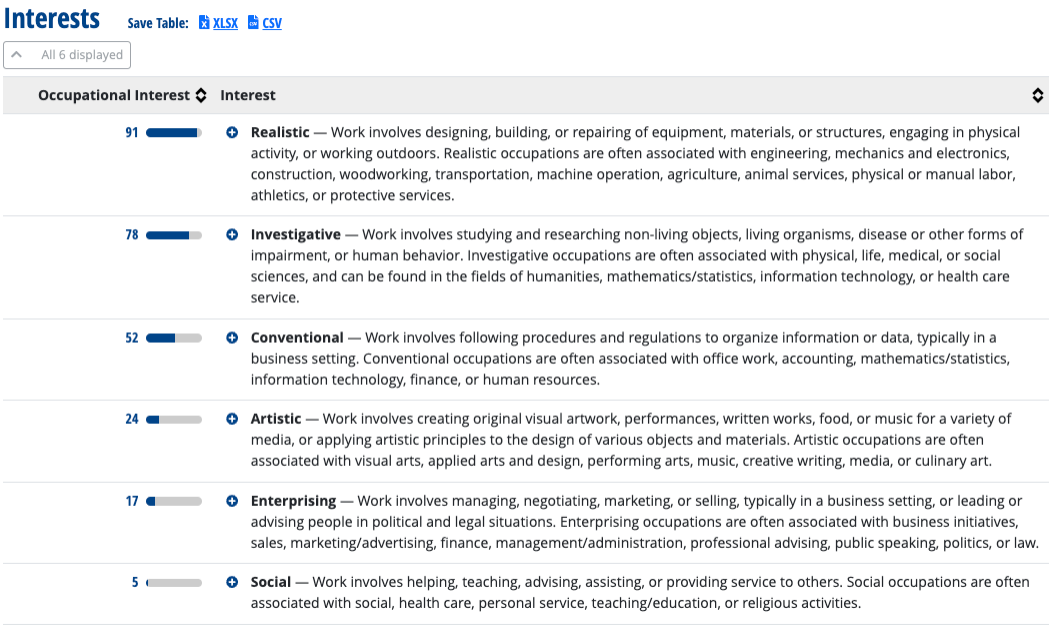
Appendix 1. Generation of job recommendations

Each participant will have their six-letter vocational interest profile (e.g., SEIRAC) upon completion of the career assessment test. The order of the six letters represents the relative ranking of six vocational interest dimensions such that the first letter represents the highest-scored interest dimension (e.g., S for social) and the last letter represents the lowest-scored interest dimension (e.g., C for conventional).

To create a list of recommended jobs for participants, I select jobs that are representative of each vocational interest profile from the O*Net 28.0 database. The database provides information about 874 occupations and their corresponding occupation codes, interest codes, job zones, and ratings on each RIASEC dimension. I only focus on jobs located in Job Zone 4 or 5, as these job zones require education levels (i.e., bachelor's degree, graduate school or above) that fit the target participants of the study. This filter left us with a pool of 353 occupations.

Representative jobs for each vocational interest profile are identified based on its interest codes and ratings on RIASEC dimensions in the O*Net database (www.onetonline.org/find/descriptor/browse/1.B.1/). For example, "Automotive Engineer" (www.onetonline.org/link/summary/17-2141.02) is identified as the representative job for the [Realistic + Investigative (RI)] profile, as it scores 91/100 on the Realistic dimension and 71/100 on the Investigative dimension (see Figure A1); "Recreation Workers" (www.onetonline.org/link/summary/39-9032.00) is identified as the representative job for the [Social + Enterprising (SE)] profile, as it scores 65 out of 100 on the Social dimension and 59 out of 100 on the Enterprising dimension (see Figure A2).

Figure A1. Interests scores of automotive engineer



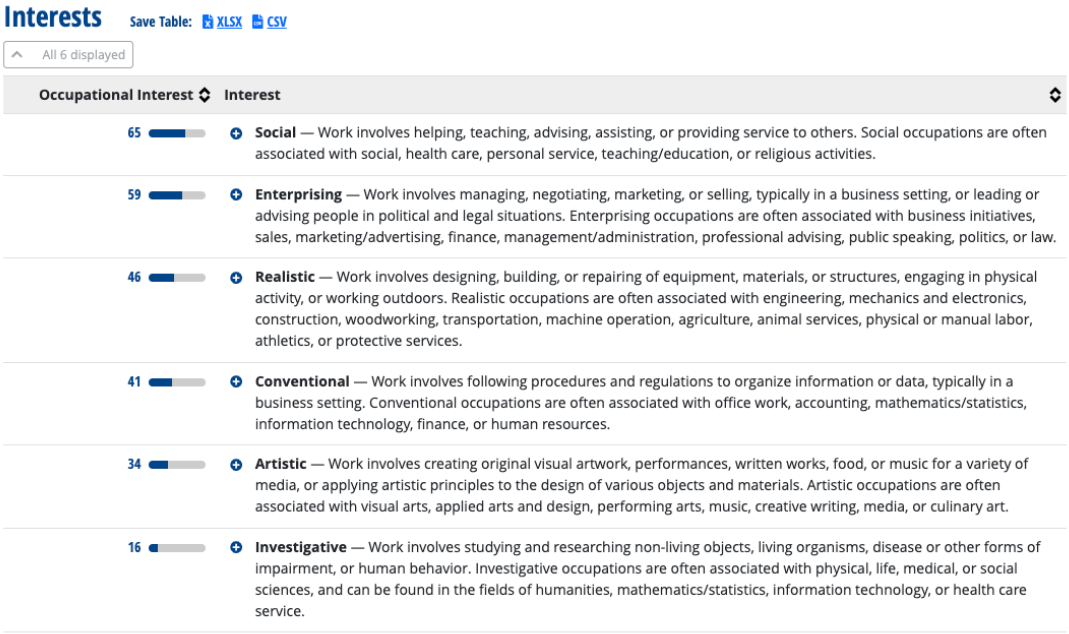
Interests

Save Table: XLSX CSV

All 6 displayed

Occupational Interest	Interest
91	**Realistic** — Work involves designing, building, or repairing of equipment, materials, or structures, engaging in physical activity, or working outdoors. Realistic occupations are often associated with engineering, mechanics and electronics, construction, woodworking, transportation, machine operation, agriculture, animal services, physical or manual labor, athletics, or protective services.
78	**Investigative** — Work involves studying and researching non-living objects, living organisms, disease or other forms of impairment, or human behavior. Investigative occupations are often associated with physical, life, medical, or social sciences, and can be found in the fields of humanities, mathematics/statistics, information technology, or health care service.
52	**Conventional** — Work involves following procedures and regulations to organize information or data, typically in a business setting. Conventional occupations are often associated with office work, accounting, mathematics/statistics, information technology, finance, or human resources.
24	**Artistic** — Work involves creating original visual artwork, performances, written works, food, or music for a variety of media, or applying artistic principles to the design of various objects and materials. Artistic occupations are often associated with visual arts, applied arts and design, performing arts, music, creative writing, media, or culinary art.
17	**Enterprising** — Work involves managing, negotiating, marketing, or selling, typically in a business setting, or leading or advising people in political and legal situations. Enterprising occupations are often associated with business initiatives, sales, marketing/advertising, finance, management/administration, professional advising, public speaking, politics, or law.
5	**Social** — Work involves helping, teaching, advising, assisting, or providing service to others. Social occupations are often associated with social, health care, personal service, teaching/education, or religious activities.

Figure A2. Interests scores of recreation worker



Interests

Save Table: XLSX CSV

All 6 displayed

Occupational Interest	Interest
65	**Social** — Work involves helping, teaching, advising, assisting, or providing service to others. Social occupations are often associated with social, health care, personal service, teaching/education, or religious activities.
59	**Enterprising** — Work involves managing, negotiating, marketing, or selling, typically in a business setting, or leading or advising people in political and legal situations. Enterprising occupations are often associated with business initiatives, sales, marketing/advertising, finance, management/administration, professional advising, public speaking, politics, or law.
46	**Realistic** — Work involves designing, building, or repairing of equipment, materials, or structures, engaging in physical activity, or working outdoors. Realistic occupations are often associated with engineering, mechanics and electronics, construction, woodworking, transportation, machine operation, agriculture, animal services, physical or manual labor, athletics, or protective services.
41	**Conventional** — Work involves following procedures and regulations to organize information or data, typically in a business setting. Conventional occupations are often associated with office work, accounting, mathematics/statistics, information technology, finance, or human resources.
34	**Artistic** — Work involves creating original visual artwork, performances, written works, food, or music for a variety of media, or applying artistic principles to the design of various objects and materials. Artistic occupations are often associated with visual arts, applied arts and design, performing arts, music, creative writing, media, or culinary art.
16	**Investigative** — Work involves studying and researching non-living objects, living organisms, disease or other forms of impairment, or human behavior. Investigative occupations are often associated with physical, life, medical, or social sciences, and can be found in the fields of humanities, mathematics/statistics, information technology, or health care service.

Appendix 2. Manipulation for AI information disclosure

I carefully examine your career assessment results and match them with the career interest profiles in my job database to find out the most suitable jobs for you. Specifically, advanced natural language processing techniques are employed to identify key elements (words, phrases, paragraph structures), classify them into distinct categories, and analyze the underlying themes and sentiments conveyed in your interests and job descriptions. The matching algorithm, driven by this sophisticated analysis, assigns weights to each of the six vocational interest dimensions.

Appendix 3. Detailed results of the CFA: Essay 2

Table A1. Standardized factor loadings: Essay 2

Factor	Item	Standardized loading
Perceived ability	A1	0.87
	A2	0.86
	A3	0.85
Perceived benevolence	B1	0.87
	B2	0.90
	B3	0.92
Perceived integrity	I1	0.76
	I2	0.73
	I3	0.70
Trust in AI	T1	0.93
	T2	0.95
Willingness to follow	WF1	0.84
	WF2	0.92
Willingness to give information	WG1	0.84
	WG2	0.76

Table A2. Inter-factor correlations: Essay 2

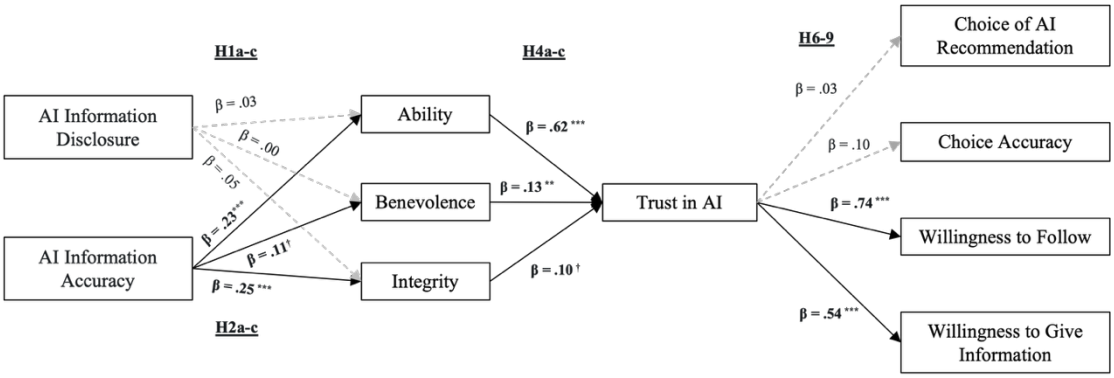
	1	2	3	4	5	6
1.Perceived ability	1.00					
2.Perceived benevolence	0.34	1.00				
3.Perceived integrity	0.84	0.50	1.00			
4.Trust in AI	0.80	0.39	0.69	1.00		
5. Willingness to follow	0.65	0.40	0.60	0.81	1.00	
6. Willingness to give information	0.53	0.23	0.50	0.62	0.67	1.00

Table A3. Squared correlations and average variance extracted: Essay 2

	AVE	1	2	3	4	5	6
1.Perceived ability	0.74	1.00					
2.Perceived benevolence	0.81	0.11	1.00				
3.Perceived integrity	0.53	0.70	0.25	1.00			
4.Trust in AI	0.89	0.64	0.15	0.48	1.00		
5. Willingness to follow	0.78	0.43	0.16	0.36	0.65	1.00	
6. Willingness to give information	0.65	0.28	0.05	0.25	0.38	0.45	1.00

Appendix 4. Summary of structural equation modeling results: Essay 2

Figure A3. Summary of SEM model: Essay 2



Note. $N = 315$. β = standardized estimated coefficient. $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$, $^{\dagger}p < .10$; two-tailed. $\chi^2_{(14)} = 32.58$, $RMSEA = .07$, $CFI = .98$, $TLI = .95$, $SRMR = .04$).

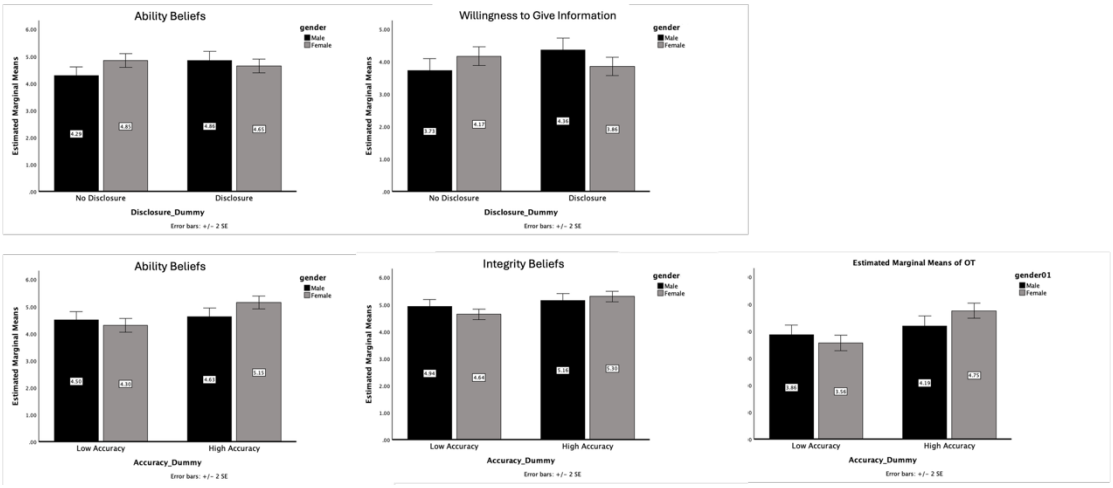
Table A3. Summary of indirect effects: Essay 2

Indirect effect	***β***	***SE***	***p***
Accuracy -> Ability -> Trust in AI -> Willingness to follow	0.11	0.08	0.000
Accuracy -> Ability -> Trust in AI -> Willingness to give information	0.13	0.09	0.000

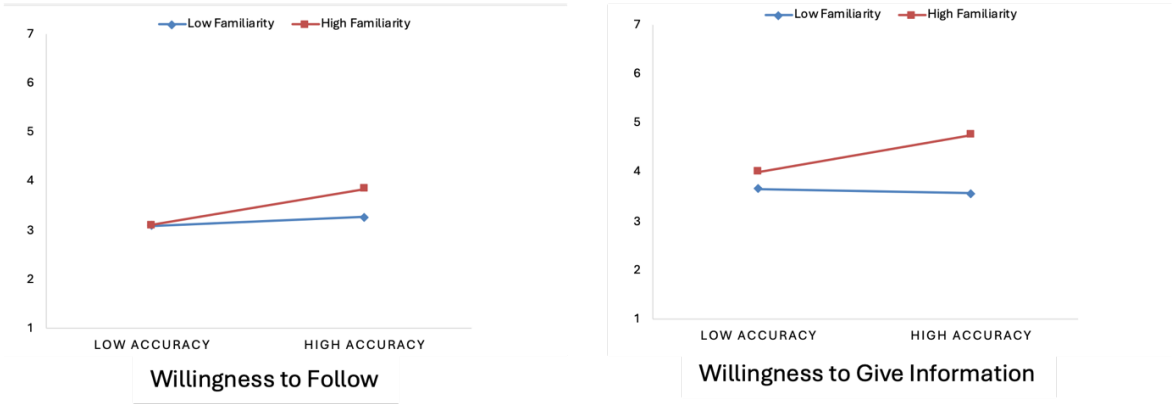
Note. $N = 315$. β= standardized estimated coefficient. Only significant paths are presented.

Appendix 5. Plots of moderating effects

Figure A4. Moderating effects of gender and familiarity with AI



(a) Moderating effect of gender



(b) Moderating effect of familiarity with AI

Essay 3: Role of AI Information Clarity and Personalization in Facilitating Trust in AI

Essay 2 illustrated the theoretical relationships between two content-related transparency facets (i.e., accuracy and disclosure) and trust-related outcomes, including trustworthiness, trust attitude, as well as adoption intentions and behaviors. Essay 3 extends the investigation of the multifaceted typology of AI transparency by examining whether clarity and personalization of AI-provided information play a role in shaping trust formation. I continue to draw on the framework of Organizational Transparency (Schnackenberg et al., 2021; Schnackenberg & Tomlinson, 2016) and Mayer et al.'s (1995) framework of trustworthiness to theorize how AI information clarity and AI information personalization relate to trust outcomes.

Hypothesis Development

AI Information Clarity and Perceived Trustworthiness

Early decision-making literature has pointed out that the choice of language in explaining a decision conveys important symbolic signals about the decision maker. Notably, the clarity and comprehensibility of the language used may serve as symbolic evidence of the decision maker's competence and motivation to communicate the decision-making process (Elsbach & Elofson, 2000). When AI presents information in a way that is easily understood by users, it demonstrates its capacity to carefully package and tailor content for effortless comprehension, even by users with limited technical expertise. This highlights the AI's ability to communicate effectively (Schnackenberg & Tomlinson, 2016). Besides, clear information helps the users to navigate through the information processing stages, such as processing, storing, and retrieving information (Xie & Derakhshan, 2021). This lifts up their understanding of AI's functionality and reinforces perceptions of AI's reliability

(Tuncer & Ramirez, 2022). On the contrary, hard-to-understand information, such as that phrased in professional jargon, was found to lower the perceived believability and logicality of the information giver.

Additionally, the voluntary provision of clear and lucid information by AI reduces the ambiguity regarding its intentions, signaling that it is considerate of users' perspectives when generating such information. The use of obscure or overly complex wordings may raise confusion and suspicion of whether AI deliberately confuses users to make it difficult to refute its decision (Elsbach & Elofson, 2000) or to achieve purposes that may be detrimental to users' interests. Therefore, information clarity contributes to the perception of benevolence. Similarly, it contributes to the perception of integrity by showcasing AI's honesty and potential adherence to ethical principles.

Given this, I propose the following hypothesis:

> ***Hypothesis 1.*** *AI information clarity increases perceived AI trustworthiness in terms of (a) ability, (b) benevolence, and (c) integrity.*

AI Information Personalization and Perceived Trustworthiness

Personalization refers to situations where AI-provided information takes users' personal interests, preferences, needs and demands, or unique characteristics into consideration, or whether such information is incorporated into the decision-making process (Komiak & Benbasat, 2006; Min et al., 2021; Shi et al., 2021; Wu et al., 2022; Yokoi et al., 2021). With the prevalence of AI technologies, especially in the form of recommendation agents (RAs), personalization has become an important quality for users to evaluate AI-provided information.

Highly personalized AI is more likely to have the capability to deliver more satisfactory output that aligns with users' needs and preferences. This also

underscores the system's sophistication and signals the use of complex, nuanced algorithms rather than oversimplified guidelines. As a result, users will attribute higher ability to AI that provides more personalized information.

Furthermore, personalization inherently demonstrates care and consideration of individual needs and preferences, fostering a sense of benevolence. Users may also perceive that the AI utilizes a similar approach or mindset for ranking and weighing relevant factors in the decision-making process. In doing so, AI is perceived as operating on principles that users deem acceptable or even favorable, enhancing perceptions of integrity. Thus, I propose the following hypothesis:

> ***Hypothesis 2.*** *AI information personalization increases perceived AI trustworthiness in terms of (a) ability, (b) benevolence, and (c) integrity.*

Interaction of AI Information Clarity and Personalization

In addition to the hypothesized main effects of AI information clarity and personalization on perceived trustworthiness, I propose that AI information clarity may help fully realize the benefit of personalization. By enhancing users' ability to interpret and understand how AI customizes its output for each user, it mitigates the risks of being misinterpreted by the users. Thus, I hypothesize the following:

> ***Hypothesis 3.*** *AI information personalization interacts with clarity to influence perceived AI trustworthiness in terms of (a) ability, (b) benevolence, and (c) integrity, such that the effect of personalization is more salient when AI information clarity is high rather than low.*

Perceived Trustworthiness, Trust in AI, and Consequences

Building upon the theoretical arguments in Essay 2, users' trustworthiness beliefs about AI will shape their trust attitudes, which eventually influence their

behavioral responses to both current and future tasks. This aligns with the Theory of Planned Behavior (Ajzen, 1991), which posits that attitudes toward a specific entity, which are formed based on certain beliefs, play a crucial role in shaping subsequent behavioral intentions and actions.

> ***Hypothesis 4.*** *Perceived AI trustworthiness in terms of (a) ability, (b) benevolence, and (c) integrity is positively related to trust in AI, with perceived ability expected to have the strongest influence.*
> ***Hypothesis 5.*** *Perceived AI trustworthiness mediated the effect of (a) AI information clarity and (b) AI information personalization on trust in AI.*

> ***Hypothesis 6.*** *Trust in AI is positively related to trust behavior.*
> ***Hypothesis 7.*** *Trust in AI is positively related to future trust intentions in terms of willingness to follow AI recommendations and willingness to give information.*

The overall theoretical framework of Essay 3 is presented in Figure 1.

Methods and Results

Sample and Procedure

A 2 (high vs. low clarity) by 2 (high vs. low personalization) between-subject online experiment was conducted to test the hypotheses. For this research design, G*Power suggested a minimum sample size of 128 to detect a medium effect size (r = .25) with 80% statistical power at a significance level of .05. I recruited 185 participants from Prolific, an online crowdsourcing platform, at an hourly rate of £9. After excluding 17 responses that failed the attention check, 168 valid responses (90.81%) were retained for data analysis.

To extend Essay 2's investigation of AI transparency, I targeted participants with similar demographic characteristics – undergraduate students majoring in

business-related disciplines (i.e., accounting, business, economics, finance, management, marketing) whose primary language was English. Compared to Essay 2, however, the Prolific sample was more ethnically diverse, enabling examination of the robustness of the findings across cultural contexts (40.5% Black or African American, 38.1% Caucasian or White, 14.9% Asian, 3% Hispanic or Latino, and 6% Others). Among the final sample of 168 participants, 45.8% of participants were female, and 66.7% were looking for a job at the time of the study. They had a mean age of 23.66 years (*SD* = 7.42) and an average of 3.45 years of internship or work experience (*SD* = 5.38). No participants had missing data on the focal variables.

The overall study procedure was the same as in Essay 2. Registered participants were randomly assigned to one of four experimental conditions and participated in an AI-mediated career assessment session, during which they viewed and chose among four recommended jobs. They also answered some questions about their personal information as well as their feelings, attitudes, and behaviors. As before, the platform had a built-in mouse-tracking function to track participants' viewing behaviors. More information about the career assessment test and report can be found in Essay 2.

AI information *clarity* and *personalization* were experimentally manipulated using two text paragraphs presented sequentially between the page showing assessment results and the page showing recommended jobs. First, participants read a paragraph about whether the job recommendations were intended to be highly personalized (i.e., specifically tailored to individual career interests, skills and preferences) or less personalized (i.e., commonly available to the general population). Next, they read a paragraph explaining how the AI career advisor generated the job recommendations, phrased either in layman's language (to increase perceived clarity)

or technical jargon (to decrease perceived clarity), depending on their assigned condition. Full examples of the manipulation texts are provided in Appendix 1.

Compared to Essay 2, where accuracy was manipulated more implicitly, I used more explicit manipulations in Essay 3 by displaying paragraphs with different wordings and styles. Given that clarity and personalization are less readily detected by users and may be more peripheral in information processing, this approach was intended to maximize the strength and salience of the manipulations.

Measures

Focal variables were measured with the same scales as in Essay 2. All items were rated on a seven-point Likert scale (1 = *strongly disagree*, 7 = *strongly agree*).

As before, I measured participants' *age*, *gender* (male = 0, female = 1), *familiarity with AI*, and their *propensity to trust*. In Essay 3, I included additional questions to capture participants' Big Five personalities, AI literacy, AI use frequency, as well as their familiarity with AI-mediated career assessment. Unless other stated, scales were rated on a seven-point Likert scale (1 = *strongly disagree*, 7 = *strongly agree*).

Big Five personalities were measured with a 10-item scale (Gosling et al., 2003), each trait measured with two items. Sample items included "sympathetic, warm" for agreeableness (α = .29), "disorganized, careless (reverse-coded)" for conscientiousness (α = .66), "extraverted, enthusiastic" for extraversion (α = .58), "anxious, easily upset" for neuroticism (α = .62), and "open to new experiences, complex" for openness to experience (α = .34). Due to the low reliability of agreeableness and openness to experience measures, I only included them in the correlational analysis but no other analyses.

AI literacy was measured with three items selected from the AI Literacy Scale

developed by Wang et al. (2022). A sample item is "I can evaluate the capabilities and limitations of an AI application or product after using it for a while" (α = .79).

AI use frequency was measured with a single item "How often do you use AI-based tools in your daily life?" on a seven-point Likert scale (1 = *never*, 7 = *always*) (Reis et al., 2024).

Finally, *familiarity with AI career assessment* was measured with three items adapted from Collins (2007). A sample item is "I am familiar with AI as a career advisor" (1 = *strongly disagree*, 7 = *strongly agree*; α = .83).

Results

Descriptive Statistics

Table 2 presents the descriptive statistics, reliabilities, and correlations of focal variables. Manipulations of clarity and personalization were not significantly related to perceived AI trustworthiness in terms of *ability*, *benevolence*, and *integrity*. Nevertheless, perceived clarity and perceived personalization both had significant and positive relationships with *ability* ($r_{clarity}$ = .59, $r_{personal}$ = .67), *benevolence* ($r_{clarity}$ = .39, $r_{personal}$ = .40,), and *integrity* ($r_{clarity}$ = .57, $r_{personal}$ = .62, all $p < .001$).

Consistent with my prediction, three AI trustworthiness dimensions were all positively related to trust in AI ($r_{ability}$ = .87, $r_{benevolence}$ = .58, $r_{integrity}$ = .80; all $p < .001$), willingness to follow ($r_{ability}$ = .72, $r_{benevolence}$ = .41, $r_{integrity}$ = .65; all $p < .001$), and willingness to give information ($r_{ability}$ = .46, $r_{benevolence}$ = .32, $r_{integrity}$ = .41, all $p < .001$). Perceived AI trustworthiness was not significantly associated with the choice of AI recommendation.

Trust attitude was positively correlated with willingness to follow (r = .78, $p < .001$) and willingness to give information (r = .50, $p < .001$), but not with the choice

of AI recommendation (r = -.13, p > .05). Choice of AI recommendation also had no significant relationship with the two future intentions.

Confirmatory Factor Analysis (CFA)

I conducted CFA in *R* (version 4.0.3, R Core Team, 2020) using the *lavaan* package (v.0.6.7) to verify our measurement model. Chi-square difference tests indicated that the six-factor model (i.e., three trustworthiness dimensions, trust in AI, willingness to follow, and willingness to give information) demonstrated a better fit to the data ($\chi^2_{(75)}$ = 171.94, *RMSEA* = .09, *CFI* = .96, *TLI* = .94, SRMR = .04) than several alternative models, showing the discriminant validity of the focal variables (see Table 1). All items loaded significantly and strongly on their intended factors (standardized loadings ranges from .73 to 1.00; see Appendix 2 Table A1). However, examination of the inter-factor correlations revealed potential overlap between perceived ability, perceived integrity and trust in AI[10], suggesting potential issues in discriminant validity despite good model fit.

Discriminant validity was evaluated using the Fornell–Larcker criterion, as Average Variance Extracted (AVE) values for each factor were compared to the squared inter-factor correlations. The AVEs for perceived ability (.70), perceived integrity (.59) and trust in AI (.86) were lower than their squared correlation with each other, indicating insufficient discriminant validity among these constructs. Discriminant validity was supported for perceived benevolence, willingness to follow, and willingness to give information. The full matrix of inter-factor correlations and AVE is presented in Appendix 2.

Manipulation Check

[10] Inter-factor correlation is .963 between perceived ability and integrity; .969 between perceived ability and trust in AI; and .932 between perceived integrity and trust in AI.

After the career assessment session, participants were instructed to recall the information provided by the AI career advisor regarding its job recommendations. To assess the effectiveness of the manipulations, participants responded to four items measuring perceived information clarity and four items measuring perceived information personalization on a 7-point Likert scale (1 = *strongly disagree*, 7 = *strongly agree*). Sample items included "The explanations are easily understandable for me, even if I have little technical knowledge" (perceived clarity, $\alpha = .81$) and "The AI career advisor understood and represented my personal interests" (perceived personalization, $\alpha = .93$).

Independent samples t-tests indicated that the manipulations did not produce the expected effects. Participants perceived a higher level of information clarity when the information was presented in technical jargon ($M = 6.08$, $SD = 0.76$) compared to layman's language ($M = 6.01$, $SD = 0.81$), although this difference did not reach statistical significance ($\Delta M = .07$, $t_{(166)} = -.57$, $p = .572$). Participants in the High Personalization condition perceived a significantly higher level of AI information personalization ($M = 5.64$, $SD = 1.11$) than those in the Low Personalization condition ($M = 5.35$, $SD = 1.45$), but this difference was also not significant ($\Delta M = .29$, $t_{(166)} = 1.49$, $p = .139$).

Given that participants' perceptions might be influenced by their AI-related knowledge and expertise, I conducted supplementary analyses controlling for AI use frequency, AI literacy, familiarity with AI, and familiarity with AI career assessment. When including these covariates, the effect of the two manipulations remained non-significant (clarity: $F_{(1,162)} = 2.51$, $p = .115$; personalization: $F_{(1,162)} = 2.97$, $p = .087$).

These results suggest that, although the manipulations were intended to induce differences in perceived clarity and personalization, participants' subjective

perceptions did not differ substantially across conditions. I discuss the implications of these findings in the Discussion section.

Analytic Strategy

All hypotheses were tested using the same analytic procedures as in Essay 2. Because the experimental manipulations were not effective, I conducted exploratory hypothesis tests using perceived clarity and personalization as predictors. These exploratory analyses are reported in the Supplementary Analysis section.

Hypothesis Tests

Consistent with the manipulation check results, our manipulations of AI information clarity and personalization did not manifest, having no significant effects on AI trustworthiness, failing to support H1 and H2. Patterns of means indicated higher perceived ability (ΔM = .10, $t_{(166)}$ = .53, p = .597), benevolence (ΔM = .05, $t_{(166)}$ = .22, p = .828), and integrity (ΔM = .10, $t_{(166)}$ = .63, p = .533) in the High Clarity (layman language) than Low Clarity (technical jargon) condition.

H3 also did not receive support, as the interaction between AI information clarity and personalization was not significant in influencing perceived AI ability ($F_{(1, 164)}$ = 1.15, p = .286, partial η^2 = .007), benevolence ($F_{(1, 164)}$ = 2.78, p = .097, partial η^2 = .017), or integrity ($F_{(1, 164)}$ = 1.39, p = .241, partial η^2 = .008).

Linear regression analyses revealed that only perceived AI ability (B = .77, SE = .08, $p < .001$) and integrity (B = .37, SE = .10, $p < .001$) significantly influenced trust in AI, supporting H4a and H4c. Surprisingly, perceived benevolence did not have a significant impact on trust in AI (B = .07, SE = .05, p = .173), providing no support for H4b. Pairwise comparisons of the regression coefficients further indicated that the effect of ability was significantly larger than the effect of both benevolence (t = 7.42, $p < .001$) and integrity (t = 3.12, $p < .01$). In addition, the effect of integrity is

also significantly larger than that of benevolence ($t = 2.68$, $p < .01$). This provides general support for H4.

H5 did not receive any support, as there were no significant mediating effects for the effects of AI information clarity and personalization on trust in AI.

Regarding the consequences, trust in AI did not significantly influence choices of AI recommendations ($B = -.28$, Wald statistic = 2.71, $p = .100$), but trust in AI is significantly and positively related to participants' willingness to follow AI's job recommendations ($B = .84$, $SE = .05$, $p < .001$) and give information to the AI career advisor in future sessions ($B = .55$, $SE = .07$, $p < .001$). Thus, H7 was supported, while H6 was not.

The results of the hypothesis tests are summarized in Figure 2.

Supplementary Analysis

Effects of Perceived Clarity and Personalization. As the manipulations did not work as expected, I conducted a set of exploratory hypothesis tests with perceived clarity and personalization as the independent variables. Linear regression analyses showed that *perceived clarity* had positive and significant relationships with perceived AI ability ($B = .92$, $SE = .10$), benevolence ($B = .70$, $SE = .13$), and integrity ($B = .75$, $SE = .08$, all $p < .001$). Similarly, *perceived personalization* was positively and significantly related to perceived AI ability ($B = .63$, $SE = .05$), benevolence ($B = .44$, $SE = .08$), and integrity ($B = .49$, $SE = .05$, all $p < .001$). The interaction effect of perceived clarity and perceived personalization on perceived AI trustworthiness was not significant. Yet, perceived AI *ability* and *integrity* significantly mediated the relationship between *perceived clarity* (ability: indirect effect = .68, $se = .11$, 95%CI [.48, .92]; integrity: indirect effect = .26, $se = .09$, 95%CI [.07, .44]) and trust in AI. Similar patterns were observed for perceived personalization (ability: indirect effect

= .33, *se* = .05, 95%*CI* [.24, .43]; integrity: indirect effect = .12, *se* = .05, 95%*CI* [.02, .21]).

Mediation Effects of Perceived Trustworthiness and Trust in AI. Similar to in Essay 2, I conducted SEM analyses to test two overall models, one with *manipulated* clarity and personality as the independent variables (model 1), and another model with *perceived* clarity and personality as the independent variables (model 2). The two theorized model both showed good fit to the data (model 1: $\chi^2_{(17)}$ = 22.12, *RMSEA* = .04, *CFI* = .99, *TLI* = .98, *SRMR* = .03; model 2: $\chi^2_{(11)}$ = 44.80, *RMSEA* = .13, *CFI* = .96, *TLI* = .89, *SRMR* = .04). Results revealed a sequential mediation effect of *perceived* clarity on (a) willingness to follow AI's recommendations (β = 0.11, p = .007) and (b) willingness to give information to the AI (β = 0.09, p = .013) through perceived ability. Similar paths work for perceived personalization on (a) willingness to follow AI's recommendations (β = 0.20, p < .001) and (b) willingness to give information to the AI (β = 0.17, p = .002) through perceived ability. Additionally, perceived personalization also influences willingness to follow AI's recommendations through perceived integrity (β = 0.06, p = .050). The summary of the SEM model is presented in Appendix 3.

Additional Metrics of Viewing Behavior. As in Essay 2, I tracked participants' time spent on viewing (a) the career assessment results, (b) manipulation information, (c) the job recommendation page, and (d) each of the recommended jobs. Descriptively, participants in the Low Clarity condition on average spent longer time than those in the High Clarity condition in viewing the manipulation paragraph (M_{jargon} = 18.84 min, M_{layman} = 8.35 min) and viewing AI-recommended jobs (M_{jargon} = 2.42 min, M_{layman} = 2.11 min), meanwhile spending less time in viewing non-AI-recommended jobs (M_{jargon} = 0.79 min, M_{layman} = 1.38 min). Participants in the High

Personalization condition on average spent longer time than those in the Low Personalization condition in viewing the manipulation paragraph (M_{high_p} = 19.13 min, M_{low_p} = 7.79 min) and viewing non-AI-recommended jobs (M_{high_p} = 1.15 min, M_{low_p} = 1.02 min), while spending less time in viewing AI-recommended jobs (M_{high_p} = 2.20 min, M_{low_p} = 2.32 min). These mean differences were not statistically significant.

Moderating Effect of Individual Characteristics. I explored the potential moderating effects of various constructs representing individual demographics (age, gender), traits (propensity to trust, personalities), experience (AI use frequency), and expertise (AI literacy, familiarity with AI, familiarity with AICA). The relationships were explored using two-way ANOVA and SPSS PROCES Macro. Results revealed the moderating effects of age, AI use frequency, familiarity with AI and AICA, conscientiousness, and extraversion. The plots of estimated marginal means are provided in Appendix 4.

AI-related experience & expertise. AI use frequency and familiarity with AI career assessment moderated the effect of AI information clarity (AIfreq: $F_{(1,164)}$ = 3.96, p = .05; famAICA: $F_{(1,164)}$ = 3.92, p = .05), such that AI information presented layman's language was associated with higher *benevolence* perceptions for those who used AI more often or were more familiar with AI, but with lower *benevolence* perceptions for those with fewer usage and lower familiarity. Also, the effect of *clarity* on *trust in AI* was positive for participants who were more familiar with AI-mediated career assessment, but negative and significant for those who were less familiar.

Age. Age moderates the effect of AI information personalization on trust in AI ($F_{(1,164)}$ = 4.25, p = .04), willingness to follow ($F_{(1,164)}$ = 4.25, p = .04) and give

information ($F_{(1,164)}$ = 4.25, p = .04), such that the relationship was positive for younger participants (i.e., 18-year-old) but negative (yet non-significant) for the elder (i.e., 31.08-year-old).

Big five personalities. *Conscientiousness* was found to moderate the effect of personalization on perceived AI benevolence ($F_{(1,164)}$ = 4.52, p = .04) and willingness to give information ($F_{con\ 1,164)}$ = 4.01, p = .05). The relationship between personalization and the two outcomes was positive and significant for less conscientious participants and negative for more conscientious participants. Finally, the effect of personalization was contingent on the participant's *extraversion* level – higher AI information personalization was related to a lower willingness to give information for those more extraverted but higher for those less extraverted ($F_{(1,164)}$ = 4.06, p =.05).

Discussion

This study investigates two other facets of AI transparency – AI information clarity and AI information personalization. A between-subjects experiment featuring an AI-mediated career assessment session revealed no significant main effects of the two on trustworthiness perceptions and no significant interactions either.

The results of manipulation checks indicated that participants rated the clarity level as very high (both means exceeded 6.0 out of 7.0) regardless of whether the manipulation information was written with layman's language or technical jargon, and rated the level of personalization as moderately high (both means exceeded 5.0 out of 7.0) regardless of their assigned conditions. A closer look at the sample revealed a possible reason for this – our participants, on average, rated themselves as having relatively rich AI experience ($M_{AIfrequency}$ = 4.98, SD = 1.35) and expertise ($M_{AIliteracy}$ = 5.72, SD = 0.81; $M_{familiarityAI}$ = 5.55, SD = 0.91; $M_{familiarityAICA}$ = 4.56, SD =

1.50)[11]. This relatively small variance in participants' AI experience and expertise to some extent hindered our ability to investigate the effect of clarity and personalization for those who are less knowledgeable with AI.

In the Supplementary Analysis section, I explored whether findings would vary if we replaced manipulation with subjective measures for AI information clarity and personalization. Interestingly, I found that the main effects of *perceived clarity* and *perceived personalization* on three AI trustworthiness dimensions were significant and positive, and that perceived AI ability and integrity mediated their effects on trust in AI. These findings provide some support for our hypothesized relationship. It also highlights the limitations of our manipulations in effectively eliciting participants' perceptions of clarity and personalization. Future studies should explore more sophisticated approaches to operationalizing the two facets of AI transparency, particularly clarity.

It is worth noting that the results of supplementary analyses indicated that the effects of AI information clarity and personalization did manifest in some situations, depending on certain individual characteristics. This is further elaborated in the next section.

[11] ANOVA results further revealed that participants differed significantly in their self-rated AI use frequency ($p = .034$) and familiarity with AI career assessment ($p = .091$). Results for H1 and H2 did not change if these four variables were included as covariates.

General Discussion

The purpose of Essay 2 and Essay 3 was to unpack the nuanced effects of different AI transparency facets in developing users' trust in AI. Drawing on the framework of Organizational Transparency (Schnackenberg et al., 2021; Schnackenberg & Tomlinson, 2016) and Mayer et al.'s (1995) framework of trustworthiness, I proposed a framework of AI transparency that views transparency as *information quality* and outlines four facets – AI information disclosure, accuracy, clarity, and personalization. Across two between-subject experiments in a career assessment context, I examined the effect of each AI transparency facet on users' trustworthiness beliefs, trust attitudes, behavioral responses (i.e., choice of AI recommendation, choice accuracy), and subsequent behavioral intentions (i.e., willingness to follow, willingness to give information).

Findings consistently demonstrated strong support for the importance of AI information accuracy in cultivating perceptions of ability and integrity, but *not* benevolence. In contrast, the effects of the other facets received very limited support, as they did not independently enhance trustworthiness perceptions in this context.

Consistent with Mayer et al.'s (1995) framework, significant relationships were found among trustworthiness perceptions, trust in AI, and future behavioral intentions. Among the trustworthiness dimensions, only perceived ability mediated the effect of AI information accuracy on trust in AI. However, trustworthiness and trust attitudes did not always translate into actual adoption of AI recommendations, particularly in Essay 2.

Overall, this research presented preliminary efforts in examining how AI transparency can be conceptualized as a multifaceted construct capturing the quality of AI information output, with some facets exerting stronger and more consistent

influence than others. Below, I discuss the common themes of findings arising from these experiments, followed by the theoretical and practical implications, limitations, and directions for future research.

Emerging Findings from Two Essays

In addition to the key findings illustrated in the Discussion sections in Essay 2 and Essay 3, three common themes of findings arose regarding (a) users' processing process of AI-provided information, (b) users' trustworthiness perceptions in the AI context, and (c) users' AI-following behavior.

Processing AI-Provided Information

Although only the manipulations of AI information accuracy manifested as expected, our embedded time metrics on the self-hosted platform enable us to have a glimpse at whether and for how long participants viewed and processed the given information. In Essay 2's study, people (in the Disclosure conditions) spent an average of 1.73 minutes on the page of manipulated information, 2.15 minutes on descriptions of AI-recommended jobs, and 1.17 minutes on descriptions of non-AI-recommended jobs. In Essay 3, people spent an average of 13.6 minutes on the page of manipulated information, 2.26 minutes on AI-recommended jobs, and 1.09 minutes on non-AI-recommended jobs. These statistics indicated that participants did pay attention to the experimental information, just that the nuances of the information were not obvious or impactful enough to affect the perceptions and decision-making process of the information recipients.

Relevant to these findings are the information processing theories that illustrate how individuals process information via various approaches. One of the most utilized theories in studies of human-technology interaction is the heuristic-systematic model (HSM) (Chaiken, 1980), which posits two parallel routes of

persuasion and information processing. In a systematic view, information recipients invest considerable cognitive effort in evaluating the given information in aspects such as argument validity. In comparison, a heuristic processing view proposes that individuals rely on comparatively little cognitive effort in evaluating information, and typically rely on simpler rules or heuristics to form their judgments (Chaiken, 1980; Liu, 2021; Shi et al., 2021; Shin, 2021). In our research context, the findings indicated a possibility that participants relied more on the heuristic route to process the manipulation information, treating it as a single informational clue without spending too much cognitive effort on the content and validity of the piece of information. This gives rise to a direction for future research to further investigate which information-processing route will be utilized more by AI users to form AI-related evaluations and influence subsequent decision-making processes.

Trustworthiness Perceptions in the AI Context

There have been discussions of whether Mayer et al.'s (1995) model of trustworthiness, primarily focusing on interpersonal trust, still applies in the context of AI. In a recent study, Lalot and Bertram (2025) provided empirical evidence for applying this model in studying trust in virtual AI, with perceived ability and integrity being more effective in forming trust. Focusing on an embedded form of AI, our two experiments in Essay 2 and Essay 3 revealed a similar effect of perceived ability and less robust effects of perceived benevolence and integrity. However, another study by Li and Bitterly (2024) found that benevolence played an important role when forming trust in AI management systems. These inconsistent findings suggest potential contingencies of how people attribute weights to the importance of these three aspects when evaluating AI's trustworthiness and forming trust in AI. Possible contingencies include AI's embodiment (robotic, virtual, embedded), personality structures of the

user (e.g., AI-related experience), or characteristics of the performed task and context (e.g., demand for empathy, Li & Bitterly, 2024; need for personalization, Qin et al., 2025).

Predicting Choice of AI Recommendation

In both studies, whether participants chose AI-recommended jobs was not significantly related to experimental conditions and trustworthiness beliefs. It was positively and significantly associated with trust in AI in Essay 2, but not in Essay 3. Such findings revealed some discrepancies from what we hypothesized according to the Theory of Planned Behavior.

One possible reason for this is the existence of additional factors that affect participants' choices. As an exploratory approach, I incorporate an open-ended question asking participants to elaborate on why they chose the particular job. While most of the participants emphasized "alignment with interest" as the critical factor influencing their choices, some participants made their choice based on the prospects of the job (e.g., monetary income), or even purely out of curiosity when the recommended job was new to them. Thus, there is a need for future research to employ a more fine-grained measure of trust behavior or refine the research design to tease out potential confounding factors.

Theoretical Implications

This research work attempts to contribute to the literature on trust and human-AI interaction in several aspects. First, it is among the earliest efforts to apply the Organizational Transparency framework (Schnackenberg et al., 2021; Schnackenberg & Tomlinson, 2016) to the AI context and develop a systematic, multifaceted typology of AI transparency. By reviewing prior studies on AI transparency and trust in AI, this research identifies four critical facets of AI transparency – information

disclosure, accuracy, clarity, and personalization – which collectively capture the quality of AI-generated information. The findings support Schnackenberg and Tomlinson's (2016) proposition that information accuracy positively influences perceived trustworthiness in terms of ability, benevolence and integrity. Additionally, this research provides preliminary explanations for the inconsistency in the effect of AI transparency on trust in AI, as observed by previous research (L. Yu & Li, 2022). Although our findings did not replicate previous research findings regarding how the packaging of decision explanations (e.g., language choice) affects users' evaluations of trustworthiness, information processing research suggested a possible explanation – AI users may process the four different facets hierarchically and sequentially, with the effects of some facets potentially being overshadowed by the effect of the most prominent facet, such as AI information accuracy.

Second, our findings offer cumulative knowledge in the trust literature regarding the distinctiveness of trustworthiness beliefs, trust attitudes, intentions, and behaviors. Participants formed beliefs of AI trustworthiness that shaped their trust in AI, which in turn influenced their adoption of AI recommendations and intentions to engage with AI in the future. These findings align with Mayer et al.'s (1995) and McKnight et al.'s (2011) models, as well as with the Theory of Planned Behavior (Ajzen, 1991). While results in Essay 2 and Essay 3 generally support this conceptual distinction, some trustworthiness dimensions – particularly ability and integrity – showed high intercorrelations, suggesting that users' evaluations may partially overlap in practice. We highlight the importance of both conceptual clarity and careful measurement when examining trustworthiness and trust. Moreover, the evidence that ability, benevolence, and integrity exhibited differential relationships with trust indicates the value of examining these dimensions separately, rather than

aggregating them into a single indicator (Lalot & Bertram, 2025).

Finally, this research explores the boundary conditions of the effects of AI transparency by examining the role of individual characteristics. Factors such as age, gender, AI-related experience and expertise, and personality traits (e.g., propensity to trust and Big Five personalities) were all found to moderate the effects of AI transparency on various outcomes. While these findings are exploratory, they highlight the theoretical value of understanding *who* may benefit more from interactions with AI.

Practical Implications

Our research findings emphasized the complexity of designing AI transparency with the aim of improving users' interaction with AI. The multifaceted nature of AI transparency suggests that simply disclosing AI's functioning logic and process may be insufficient to build user trust –accuracy, clarity, and personalization must also be considered when crafting the piece of information to be communicated to users.

AI career coaching is emerging and increasingly adopted (Timis & Alurralde, 2023). Designed in the particular context of AI-mediated career assessment and recommendations, our research provides actionable insights for practitioners who aim to optimize the AI career coaching process for enhanced user trust and adoption outcomes. Specifically, the *accuracy* of AI-generated information significantly affects users' perceptions of AI trustworthiness. Among the three dimensions of AI trustworthiness, perceived *ability* explained the largest variance in user trust and subsequent usage behaviors, including adoption of AI recommendation, willingness to apply to the recommended jobs, and willingness to share additional personal information (e.g., name, contact number, or resume) for the AI career advisor to

conduct further analysis and refine their career advice.

Limitations and Future Directions

Several limitations inherent in this research may be addressed in future studies.

First, the manipulations for disclosure, clarity, and personalization of AI information require further refinement to fully capture their theoretical underpinnings. In this study, several measures were implemented to ensure participants noticed and paid attention to the experimental information according to their assigned conditions. For instance, our self-hosted platform simulated real-life career assessment interfaces and the AI coaching process to enhance realism and participant engagement. The manipulated information was presented on a separate page from the assessment results and job recommendations to ensure clear exposure. However, the Disclosure manipulation may have been confounded with the amount of information provided to participants, as participants in this condition naturally received more information than those in the No Disclosure condition. Additionally, the level of Disclosure implemented in this study was a relatively basic statement about the decision-making principle of the AI career advisor, which may have constrained the strength of the manipulation. Future research should consider designing control conditions that involve an equivalent amount of irrelevant information (e.g., history of career assessment). The disclosed information can be further strengthened by incorporating more in-depth explanations of AI decision-making processes, or even a partial view of the underlying algorithms.

For manipulations of clarity and personalization, the textual framing of clarity and personalization did not consistently translate into differences in participants' subjective perceptions. This pattern reveals the challenge of ensuring that the

experimental context as generally understood (canonical situations) is indeed effectively recognized and experienced by participants (functional situations; Block & Block, 1981). In our research context, even when the canonical situation was identical (e.g., exposure to technical jargon), participants' functional situations varied depending on their personality structure (i.e., "developmentally achieved perceptualizing schemata"; Block & Block, 1981, p. 87), such as their prior experience and familiarity with AI. Consistent with this perspective, I conducted exploratory analyses examining whether AI-related experience and expertise (e.g., AI use frequency, AI literacy) interacted with the manipulations to shape participants' perceptions of all four facets of AI transparency, but did not find any significant effects. This suggests that other unmeasured factors may have influenced participants' subjective construal of the manipulations. To address this gap, future research might develop stronger manipulations (e.g., incorporating visual aids) to enhance the salience of manipulation cues, or employ alternative methodologies to more accurately capture participants' perceptions of AI transparency. Additionally, it would be theoretically valuable to explore whether participants' AI literacy directly affects their ability to understand and process AI-provided information, ultimately influencing the effectiveness of such manipulations.

Second, although we examined the interaction effects between AI information disclosure and accuracy as well as between clarity and personalization, we did not systematically discuss the relationships among the four AI transparency facets. Although these facets each convey certain information cues to facilitate AI users' trustworthiness evaluations, it is possible that these cues are not processed at the same time, or with similar effort. For instance, information recipients may initially evaluate a message based on its ease of comprehension and personal relevance and proceed to

examine more peripheral cues (e.g., expertise, legitimacy) when they find the message difficult to comprehend (Elsbach & Elofson, 2000). Moreover, the present findings suggest that users may attach greater weight to cues that directly signal accuracy of the information or quality of AI itself, while reacting less to communication-related cues such as disclosure, clarity and personalization. Research in related domains (e.g., organizational justice) indicated the possibility that process-related cues (e.g., procedural justice) serve not only as independent predictors but also as potential moderators of the impact of more output-related cues (e.g., distributive justice). Therefore, future research can investigate not only whether individuals process these transparency cues hierarchically or sequentially, but also how different combinations of AI transparency facets jointly shape trust in AI.

Third, the discriminant validity of trust in AI and the trustworthiness perceptions – particularly ability and integrity – warrants caution when interpreting the mediation relationships. Although ability, benevolence, and integrity are theoretically distinct dimensions, their empirical intercorrelations exceeded recommended thresholds for discriminant validity. This pattern introduces ambiguity about whether the observed mediation effects reflect unique contributions of each trustworthiness dimension or more general trustworthiness perceptions. Our ability to empirically distinguish these constructs is to some extent constrained by the cross-sectional experiment design, in which trustworthiness, trust attitude, and future intentions were measured within a very short interval. Future research would benefit from employing more fine-grained experimental designs and measurement models to more clearly disentangle these related constructs over time.

Fourth, the generalizability of our findings is limited by the demographic similarity of the samples recruited in the two experiments. This design was adopted as

it allowed us to examine the effects of different transparency facets on similar participant groups, thereby minimizing potential confounding factors. As our vignettes were highly tailored to fit our participants to maximize the realism of the career assessment session, having similar participant groups enabled us to maintain the usage of the self-hosted platforms without having to make substantial changes to the system to tailor our materials for other groups of participants. Future research should investigate the impact of AI transparency across varied demographic groups to enhance the generalizability of findings.

Furthermore, this study did not hypothesize a priori on theoretically grounded moderators. However, our exploratory analyses suggested promising avenues for future research, including potential moderating effects of users' gender, age, experience and expertise related to AI technologies, and personality traits. As mentioned above, such personality structures may influence participants' subjective construal of the AI-provided information (Block & Block, 1981). Investigating these factors may yield valuable insights into the contingencies of the effects of different AI transparency facets.

Lastly, future studies can explore additional factors influencing users' decision-making processes beyond trust in AI. This would enrich our current understanding of AI reliance and the appropriateness of trust in AI. Future research should also consider alternative approaches to capture AI reliance as well as explore additional behavioral consequences influenced by AI transparency and user trust, depending on the specific research contexts.

References

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211.

Block, J., & Block, J. H. (1981). Studying situational dimensions: A grand perspective and some limited empiricism. In D. M. Magnusson (Ed.), *Toward a psychology of situations: An interactional perspective* (pp. 85–103). Hillsdale, NJ: Erlbaum.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, *39*(5), 752–766. https://doi.org/10.1037/0022-3514.39.5.752

Collins, C. J. (2007). The interactive effects of recruitment practices and product awareness on job seekers' employer knowledge and application behaviors. *Journal of Applied Psychology*, *92*(1), 180–190.

Elsbach, K. D., & Elofson, G. (2000). How the packaging of decision explanations affects perceptions of trustworthiness. *Academy of Management Journal*, *43*(1), 80–89. https://doi.org/10.2307/1556387

Gosling, S. D., Rentfrow, P. J., & Swann Jr, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528.

Komiak, S. Y., & Benbasat, I. (2006). The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS Quarterly*, 941–960. https://doi.org/10.2307/25148760

Lalot, F., & Bertram, A.-M. (2025). When the bot walks the talk: Investigating the foundations of trust in an artificial intelligence (AI) chatbot. *Journal of Experimental Psychology: General, 154*(2), 533–

551. https://doi.org/10.1037/xge0001696

Li, M., & Bitterly, T. B. (2024). How perceived lack of benevolence harms trust of artificial intelligence management. *Journal of Applied Psychology, 109*(11), 1794–1816. https://doi.org/10.1037/apl0001200

Liu, B. J. (2021). In AI we trust? Effects of agency locus and transparency on uncertainty reduction in human-AI interaction. *Journal of Computer-Mediated Communication*, *26*(6), 384–402. https://doi.org/10.1093/jcmc/zmab013

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709–734.

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

Min, F., Zou, F., He, Y., & Jiang, X. (2021). Research on users' trust of chatbots driven by AI: An empirical analysis based on system factors and user characteristics. *2021 IEEE International Conference on Consumer Electronics and Computer Engineering*, 55–58. https://doi.org/10.1109/ICCECE51280.2021.9342098

Qin, X., Zhou, X., Chen, C., Wu, D., Zhou, H., Dong, X., ... & Lu, J. G. (2025). AI aversion or appreciation? A capability–personalization framework and a meta-analytic review. *Psychological Bulletin*, *151*(5), 580–599.

Reis, M., Reis, F., & Kunde, W. (2024). Influence of believed AI involvement on the perception of digital medical advice. *Nature Medicine*, *30*(11), 3098–3100. https://doi.org/10.1038/s41591-024-03180-7

Schnackenberg, A. K., & Tomlinson, E. C. (2016). Organizational transparency: A new perspective on managing trust in organization-stakeholder relationships.

Journal of Management, *42*(7), 1784–1810.

Schnackenberg, A. K., Tomlinson, E., & Coen, C. (2021). The dimensional structure of transparency: A construct validation of transparency as disclosure, clarity, and accuracy in organizations. *Human Relations*, *74*(10), 1628–1660. https://doi.org/10.1177/0018726720933317

Shi, S., Gong, Y., & Gursoy, D. (2021). Antecedents of trust and adoption intention toward artificially intelligent recommendation systems in travel planning: A heuristic–systematic model. *Journal of Travel Research*, *60*(8), 1714–1734. https://doi.org/10.1177/0047287520966395

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, *146*, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Timis, D. A., & Alurralde, M. (2023, September 27). How could AI shape the future of career coaching? *World Economic Forum*. https://www.weforum.org/stories/2023/09/how-could-ai-shape-the-future-of-career-coaching/

Tuncer, S., & Ramirez, A. (2022). Exploring the role of trust during human-AI collaboration in managerial decision-making processes. *24th International Conference on Human-Computer Interaction, HCII 2022, 13518 LNCS*, 541–557. https://doi.org/10.1007/978-3-031-21707-4_39

Wang, B., Rau, P.-L., & Yuan, T. (2022). Measuring user competence in using artificial intelligence: Validity and reliability of artificial intelligence literacy scale. *Behaviour & Information Technology*, *42*, 1–14. https://doi.org/10.1080/0144929X.2022.2072768

Wu, J.-J., Khan, H. A., Chien, S.-H., & Wen, C.-H. (2022). Effect of customization,

core self-evaluation, and information richness on trust in online insurance service: Intelligent agent as a moderating variable. *Asia Pacific Management Review*, *27*(1), 18–27. https://doi.org/10.1016/j.apmrv.2021.04.001

Xie, F., & Derakhshan, A. (2021). A conceptual review of positive teacher interpersonal communication behaviors in the instructional context. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.708490

Yokoi, R., Eguchi, Y., Fujita, T., & Nakayachi, K. (2021). Artificial Intelligence is trusted less than a doctor in medical treatment decisions: Influence of perceived care and value similarity. *International Journal of Human-Computer Interaction*, *37*(10), 981–990. https://doi.org/10.1080/10447318.2020.1861763

Yu, L., & Li, Y. (2022). Artificial intelligence decision-making transparency and employees' trust: The parallel multiple mediating effect of effectiveness and discomfort. *Behavioral Sciences*, *12*(5), 127. https://doi.org/10.3390/bs12050127

Tables and Figures

Table 1. Results of confirmatory factor analysis

Model	*χ2*	*df*	*p (χ2)*	*χ2/df*	*CFI*	*TLI*	*RMSEA*	*RMSEA 90%CI*	*p- RMSEA<0.05*	*SRMR*
Six-factor model	171.94	75	.00	2.29	.956	.939	.088	[.070, .105]	.000	.042
Three-factor model	548.75	87	.00	6.31	.792	.750	.178	[.164, .192]	.000	.091
Two-factor model	574.97	89	.00	6.46	.782	.742	.180	[.166, .194]	.000	.095
One-factor model	651.87	90	.00	7.24	.747	.705	.193	[.179, .207]	.000	.099

Note. In the six-factor model, three Trustworthiness variables, Trust in AI, Willingness to Follow, and Willingness to Give Information were each loaded on one factor. In the three-factor model, three Trustworthiness variables, Trust in AI, and the two Outcomes were each loaded on one factor. In the two-factor model, three Trustworthiness variables and Trust in AI were loaded on one factor, and the two Outcomes were loaded on another factor. In the one-factor model, all six variables were loaded on one factor.

Table 2. Means, standard deviations, reliabilities and correlations among variables

	Variables	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11
1	Clarity Manipulation	0.50	0.50	/										
2	Personalization Manipulation	0.51	0.50	-.05	/									
3	Perceived Clarity	6.04	0.78	-.04	.04									
4	Perceived Personalization	5.50	1.29	-.04	.11	.63***								
5	Ability Beliefs	5.54	1.21	.04	.02	.59***	.67***	*(.87)*						
6	Benevolence Beliefs	4.69	1.42	.02	.04	.39***	.40***	.55***	*(.93)*					
7	Integrity Beliefs	5.61	1.03	.05	.08	.57***	.62***	.81***	.66***	*(.81)*				
8	Trust in AI	5.29	1.49	.03	.06	.59**	.74***	.87***	.58***	.80***	*(.92)*			
9	Choice of AI Recommendation	0.83	0.37	-.06	-.09	-.02	-.04	-.10	-.12	-.05	-.13	/		
10	Willingness to Follow	4.73	1.61	.04	.10	.49***	.66***	.72***	.41***	.65***	.78***	-.06	*(.89)*	
11	Willingness to Give Information	4.83	1.65	.07	.05	.31***	.34***	.46***	.32***	.41***	.50***	-.08	.50***	*(.82)*
12	Gender	1.48	0.53	.06	-.06	.20*	.14	.09	.12	.12	.13	-.08	.15	.01
13	Age	23.66	7.42	.08	.08	-.09	-.05	-.17*	-.19*	-.23**	-.13	.01	.02	.18*
14	Propensity to Trust	3.75	1.58	.07	.10	.13	.16*	.15	.30***	.17*	.24**	.01	.18*	.35***
15	AI Use Frequency	4.98	1.35	.16*	-.01	.29***	.36***	.45***	.41***	.41***	.45***	-.16*	.48***	.34***
16	AI Literacy	5.72	0.81	.08	-.09	.38***	.26***	.29***	.25**	.29***	.30***	-.01	.37***	.36***
17	Familiarity with AI	5.55	0.91	.11	-.13	.34***	.31***	.34***	.32***	.30***	.33***	-.09	.38***	.28***
18	Familiarity with AI Career Assessment	4.56	1.50	.13	.02	.43***	.43***	.55***	.49***	.52***	.54***	-.15*	.58***	.45***
19	Extraversion	3.93	1.51	.08	.02	.19*	.22**	.19*	.21**	.15*	.16*	-.11	.13	.18*
20	Agreeableness	4.99	1.21	.05	-.08	.11	.22**	.20**	.21**	.18*	.23**	-.01	.23**	.19*
21	Conscientiousness	5.48	1.26	.12	-.04	.14	.25***	.32***	.21**	.31***	.28***	-.01	.35***	.19*
22	Neuroticism	2.94	1.32	-.01	.04	-.06	-.11	-.16*	-.32***	-.24**	-.15*	-.05	-.21**	-.17*
23	Openness to Experience	5.13	1.25	.05	-.01	.24**	.15	.25**	.28***	.21**	.22**	-.12	.26***	.13

(continued)

Variables		12	13	14	15	16	17	18	19	20	21	22	23
12	Gender	/											
13	Age	-.01	/										
14	Propensity to Trust	-.21**	.11	*(.92)*									
15	AI Use Frequency	.08	-.07	.16*	/								
16	AI Literacy	.14	-.01	.13	.45**	*(.79)*							
17	Familiarity with AI	.11	.02	.16*	.57***	.59***	*(.84)*						
18	Familiarity with AI Career Assessment	.16*	-.06	.16*	.54***	.46***	.54***	*(.83)*					
19	Extraversion	.13	-.02	.09	.21**	.17*	.20*	.23**	*(.58)*				
20	Agreeableness	.22**	.25***	.12	.22**	.09	.24**	.23**	.11	*(.29)*			
21	Conscientiousness	.06	.13	.04	.21**	.23**	.16*	.29***	.05	.35***	*(.66)*		
22	Neuroticism	.18*	-.03	-.16*	-.21**	-.26***	-.19*	-.22**	-.15	-.15	-.46***	*(.62)*	
23	Openness to Experience	.26***	-.06	-.09	.28***	.28***	.21**	.40***	.35***	.24**	.28***	-.31***	*(.34)*

Note. $N = 168$. $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$; two-tailed. For gender, male = 0, female = 1.

Figure 1. Theoretical framework of Essay 3

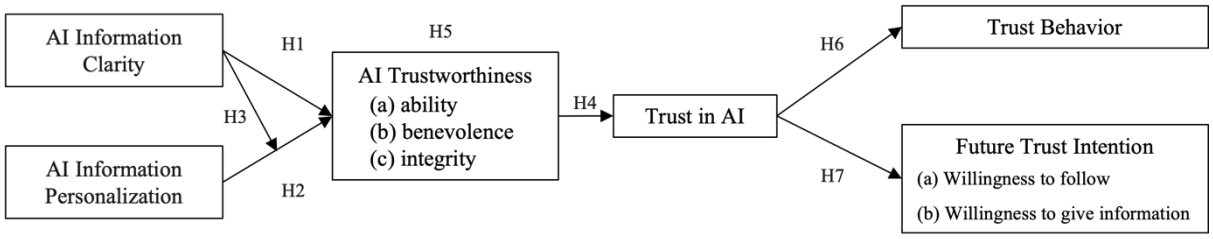
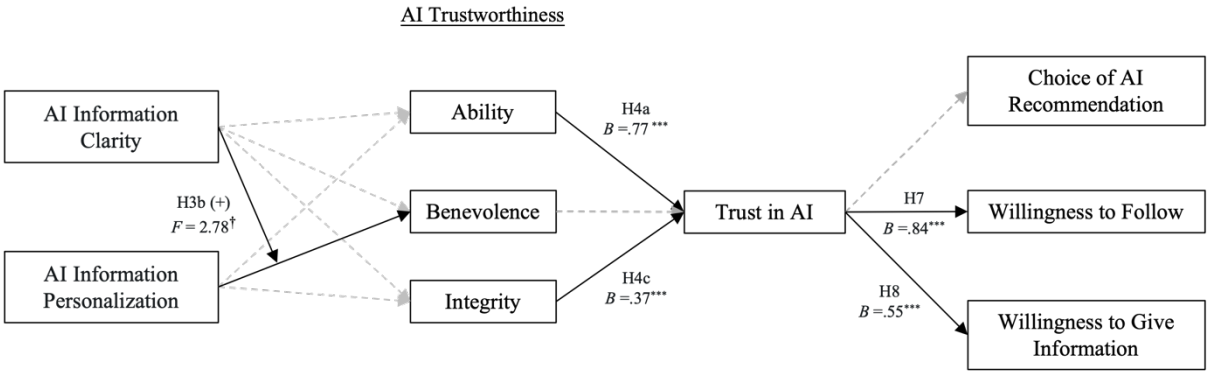


Figure 2. Summary of hypothesis test results



Note. $N = 168$. $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$, $^{\dagger}p < .10$; two-tailed. Only significant results were present.

Appendices

Appendix 1. Manipulations of AI information clarity and personalization

AI Information Personalization

- High Personalization: I carefully review your unique career interest profile based on the six career interest dimensions, and employ advanced matching algorithms to recommend jobs that are specifically tailored to your career interests, skills, and personal preferences.
- Low Personalization: I review your career interest profile based on the six career interest dimensions, and recommend jobs that are commonly available to the general population.

AI Information Clarity

- High Clarity (layman language): To achieve this, I use advanced techniques to read and understand both your profile and job descriptions. This involves picking up on and categorizing important words, phrases, paragraph structures, and overall themes and emotions. I then rate how well each job aligns with your vocational interests across six different dimensions.
- Low Clarity (technical jargon): To achieve this, I leverage advanced natural language processing techniques to perform entity extraction, hierarchical label classification, topic modeling (e.g., latent Dirichlet allocation), syntax analysis (e.g., dependency parsing), and sentiment analysis on your profile and job descriptions. The matching algorithm, fueled by this sophisticated analysis, assigns weights to each of the six vocational interest dimensions.

Appendix 2. Detailed results of the CFA: Essay 3

Table A1. Standardized factor loadings: Essay 3

Factor	Item	Standardized loading
Perceived ability	A1	0.87
	A2	0.78
	A3	0.86
Perceived benevolence	B1	0.90
	B2	0.89
	B3	0.94
Perceived integrity	I1	0.79
	I2	0.77
	I3	0.73
Trust in AI	T1	0.93
	T2	0.92
Willingness to follow	WF1	0.81
	WF2	1.00
Willingness to give information	WG1	0.90
	WG2	0.78

Table A2. Inter-factor correlations: Essay 3

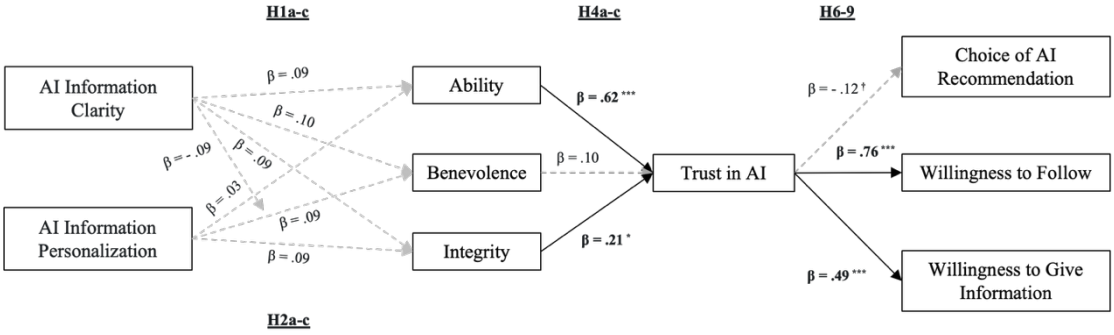
	1	2	3	4	5	6
1.Perceived ability	1.00					
2.Perceived benevolence	0.60	1.00				
3.Perceived integrity	0.96	0.76	1.00			
4.Trust in AI	0.97	0.63	0.93	1.00		
5. Willingness to follow	0.79	0.47	0.78	0.85	1.00	
6. Willingness to give information	0.55	0.36	0.52	0.58	0.55	1.00

Table A3. Squared correlations and average variance extracted: Essay 3

	AVE	1	2	3	4	5	6
1.Perceived ability	0.70	1.00					
2.Perceived benevolence	0.83	0.36	1.00				
3.Perceived integrity	0.59	0.93	0.58	1.00			
4.Trust in AI	0.86	0.94	0.40	0.87	1.00		
5. Willingness to follow	0.82	0.62	0.22	0.61	0.71	1.00	
6. Willingness to give information	0.71	0.30	0.13	0.27	0.34	0.30	1.00

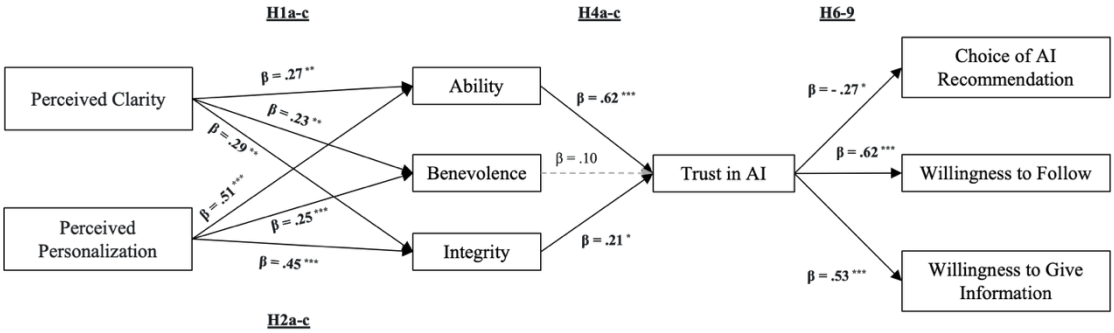
Appendix 3. Summary of structural equation modeling results: Essay 3

Figure A1. Effects of manipulated clarity and personalization



Note. N = 168. β = standardized estimated coefficient. $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$, $^{\dagger}p < .10$; two-tailed. $X^2_{(17)}$ = 22.12, *RMSEA* = .04, *CFI* = .99, *TLI* = .98, *SRMR* = .03).

Figure A2. Effects of perceived clarity and personalization



Note. N = 168. β= standardized estimated coefficient. $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$, $^{\dagger}p < .10$; two-tailed. $X^2_{(11)}$ = 44.80, *RMSEA* = .13, *CFI* = .96, *TLI* = .89, *SRMR* = .04).

Table A1. Summary of indirect effects: Essay 3

Indirect effect	***β***	***SE***	***p***
Perceived clarity -> Ability -> Trust in AI -> Willingness to follow	0.11	0.08	0.007
Perceived clarity -> Ability -> Trust in AI -> Willingness to give	0.09	0.08	0.013
Perceived personalization -> Ability -> Trust in AI -> Willingness to follow	0.20	0.07	0.000
Perceived personalization -> Ability -> Trust in AI -> Willingness to give	0.17	0.07	0.002
Perceived personalization -> Integrity -> Trust in AI -> Willingness to follow	0.06	0.04	0.050

Note. N = 168. β= standardized estimated coefficient. Only significant paths are presented.

Appendix 4. Plots of moderation effects

Figure A3. Moderating effects of AI-related experience and expertise

Figure A4. Moderating effects of age

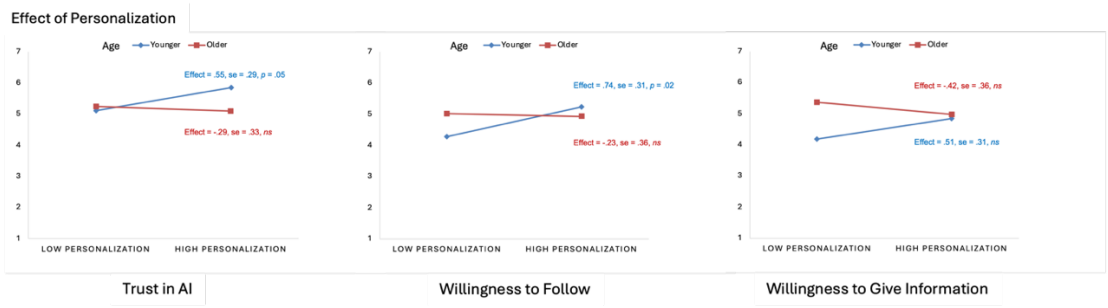


Figure A5. Moderating effects of conscientiousness and extraversion