Dec 12th, 12:00 AM

Review of Research on Human Trust in Artificial Intelligence

Yonggang Li
Information Systems and Analytics, National University of Singapore, e0530832@u.nus.edu

Jungpil Hahn
National University of Singapore, jungpil@nus.edu.sg

Review of Research on Human Trust in Artificial Intelligence

Completed Research Paper

Yonggang Li
National University of Singapore
11 Research Link, Singapore 119391
yonggang@comp.nus.edu.sg

Jungpil Hahn
National University of Singapore and
AI Singapore
11 Research Link, Singapore 119391
jungpil@nus.edu.sg

Abstract

Artificial Intelligence (AI) represents today's most advanced technologies that aim to imitate human intelligence. Whether AI can successfully be integrated into society depends on whether it can gain users' trust. We conduct a comprehensive review of recent research on human trust in AI and uncover the significant role of AI's transparency, reliability, performance, and anthropomorphism in developing trust. We also review how trust is diversely built and calibrated, and how human and environmental factors affect human trust in AI. Based on the review, the most promising future research directions are proposed.

Keywords: Artificial intelligence, review, human trust in AI, trust building, trust calibration, environment, user

Introduction

Artificial intelligence (AI), as an investigation into the true nature of intelligence, dates back to the 1950s (Nasirian 2017). AI is at the heart of what has been dubbed the Fourth Industrial Revolution (Xu et al. 2018), characterized by a shift in agency and control from humans to technology (Glikson and Woolley 2020).

AI is a set of technologies that enable machines to sense, understand, react and learn (Bowen and Morosan 2018), not only to automate but also to empower machines to exhibit mechanical, analytical, and intuitive intelligence (Huang and Rust 2018). Generally, AI uses computers as a carrier to simulate and display human intelligence processes that include learning (obtaining information and rules for utilizing the information), reasoning (using rules to arrive at approximate or unambiguous conclusions), and evolution (self-correction) (Gillath et al. 2021). Recent advances in computer and data sciences have led to a renewed interest in AI (Kerasidou 2021). AI is dramatically rebuilding the world as we know it – AI-based solutions are impacting the way organizations and jobs are designed, how decisions are made, and how knowledge is managed (Wirtz et al. 2018). In recent decades, more and more advanced AI technology has become increasingly common with an increasing impact on the lives of individuals and the functioning of our society (Makridakis 2017). Various AI technologies have been applied for different use cases, for instance, smart devices, medical diagnostic aids, autonomous vehicles, and personal assistants (e.g., Siri and Alexa) (Chi et al. 2020). AI is also becoming more and more important in business, where it has a growing range of applications from simple "chatbots" used in customer service to more complex analytical solutions based on deep learning (Okuda and Shoda 2018).

Although the prevalence of and the roles AI plays have increased exponentially, full adoption or delegation to AI models is not yet desired. This means that human-AI collaboration is necessary (Zhang et al. 2020). One important aspect of human-AI collaboration is that humans should trust in the AI systems in the same way humans typically trust other human partners (Rousseau et al. 1998).

Trust is critical for AI to continue to gain a social license. The European Commission's AI High-Level Expert Group (AI HLEG) stressed that without evidence that AI systems can be trusted, their widespread

acceptance will be hampered, and the potential for economic and social benefits will not be realized (Smuha 2019). Zhang et al. (2019) report that trust is the most significant factor in promoting positive user attitudes towards autonomous vehicles. Kamal et al. (2020) examined factors that influence patients' willingness to use telemedicine, and trust is one of the most important ones. Trust is also considered as an important prerequisite for users' acceptance of AI support services (Ostrom et al. 2019). However, trust is not that simple. Despite AI's exponential growth in popularity (e.g., Stormont 2008; You and Robert 2018), many people struggle to trust AI. For instance, although AI-based autonomous driving technology is available on cars such as Tesla, Audi, etc. However, the auto-driver function is actually rarely used due to a lack of trust.

There might be various reasons why people do not trust AI. The first possible reason is that they do not understand AI, such as how AI processes the collected data (biased training on data samples) and how it derives the outcomes (which is also called black-box issues), hence the field of research known as explainable AI (XAI), which attempts to build human-understandable and explainable systems that can explain their decisions (Doran et al. 2007), is growing popular. Another reason for distrusting AI is public fear – some people are simply afraid of AI (Winfield and Jirotka 2018). Some of these fears are fueled perhaps by the press and media hype, but some are motivated by genuine concerns about the technology's potential impact on jobs or privacy. For example, people are scared of losing their jobs by being displaced by AI, or even worse, losing their lives. Indeed, a recent survey conducted by Oxford University's Center for the Governance of AI, showed that many Americans are worried about AI (Zhang and Dafoe 2019). The third reason is that given their probabilistic nature, the correctness of AI can never be guaranteed for a particular decision. For example, even the most advanced AI-based image recognition algorithm cannot achieve 100% accuracy on all data sets. Furthermore, the accuracy of AI models depends heavily on the historical data used to train them, which can be affected by input errors, biases, and changes in the environment.

Regardless of the reason, a lack of trust may lead to a series of subsequent problems, such as reduced cooperation, efficiency, and productivity (Braynov and Sandholm 2002). A lack of trust can also reduce or prevent the integration of AI systems and agents into teams (Groom and Nass 2007), human collaboration, and the adoption of new technologies more broadly (Jeffries and Reed 2000). Currently, a lack of trust in AI systems is a major drawback in the acceptance of the technology in many fields including healthcare, finance, smart devices, etc. Similarly, Gillath et al. (2021) stated that the lack of trust is one of the main obstacles preventing people from taking full use and advantages that AI offers. According to a recent survey by IBM for Business Value, 82% of businesses (93% of high-performing companies) are considering or advancing the adoption of AI because of the technology's ability to increase revenue, improve customer service, reduce costs, and manage risk. However, according to the same survey, while these enterprises are aware of the huge benefits of AI technologies, 60% are concerned about liability issues.

While trust has been shown to be important for the acceptance and adoption of a broad range of technologies (Makridakis 2017), AI has created an array of trust challenges that are qualitatively different from traditional technologies. They proposed that the biggest challenge for society and businesses will be to harness the benefits of existing AI technologies to provide immense opportunities for new products/services and vast productivity gains while avoiding the dangers and disadvantages of rising unemployment and rising wealth inequality.

Given the critical role of trust, there is a strong practical need to understand what influences and promotes trust in AI (Lockey et al. 2021). Already, a substantial amount of research has studied the antecedents of human trust in AI and the factors impacting human trust in AI from different perspectives (Bao et al. 2021; Kaplan, 2021): 1) human perspective (e.g., human characteristics such as age, gender, etc.), 2) AI perspectives, such as performance, transparency, reliability, and uncertainty of AI, and 3) trust perspective, such as the trust-building process, trust calibration, human-AI interaction quality and so on. Table 1 shows examples of current research that studies the antecedents of trust in AI.

Despite existing literature that has investigated the factors that impact the adoption of AI devices in services (e.g., Gursoy et al. 2019), current studies still lack a deep understanding of trust in AI technologies. A systematic review of why people develop trust in AI technologies may offer additional contributions to society. Glikson and Woolley (2020) reviewed how AI differs from general technologies and presented existing research on the antecedents of human trust in AI that has been conducted over the past 20 years across multiple disciplines. The review grouped the form of AI into three clusters (i.e., robotic, virtual, and embedded) and identified the level of machine intelligence of AI (i.e., its capabilities) as important

prerequisites for the development of trust. The review further proposed a framework to address the factors that shape users' cognitive and emotional trust in AI. The review concluded by articulating the critical role of transparency, tangibility, reliability, and immediacy of AI in the development of cognitive trust, as well as the role of AI's humanness (i.e., anthropomorphism) for the development of emotional trust. Lockey et al. (2021) put forward that AI can benefit society, but it also entails great risks and societal acceptance and adoption of AI depend on stakeholders' trust in AI. However, the literature on AI trust is fragmented and it is still unclear with respect to the various vulnerabilities faced by different stakeholders. Lockey et al. (2021) conducted a review of the literature to assess the known causes and consequences of trust in AI and collated five trust issues that are specific to or exacerbated by AI. In addition, it developed a conceptual matrix that identifies key vulnerabilities to stakeholders for each challenge and proposes a multi-stakeholder approach for future research. Lotfalian and Bayrak (2021) proposed that the degree of trust of human-intelligent system interaction determines the degree of dependence of users on the system, which directly affects the contribution of intelligent systems to human decision-making. Therefore, the authors reviewed the existing literature on trust in human-AI interaction focusing on the key areas of system design to solve the problem of user trust in AI. They introduced how trust impacts the utilization of AI, described various situations in which users interact with AI systems, and classified the ways that trust is formed. They also divided the key influencers that are important for forming user trust into three types: human-related attributes, intelligent system design, and task characteristics. Finally, an analytical model to assess and predict trust was also presented. Sperrle, et al, (2020) focusing on explainable AI (XAI), distinguished between comparative and applied research, revealing methodological differences in machine learning, human-computer interaction, and visual analysis. They argued that each of these disciplines addresses a distinct part of the XAI process. The authors evaluated XAI as a whole and proposed a conceptual model describing the sources of bias and trust-building.

Category	**Context**	**Key Constructs**
Performance	Human-robot interaction	Machine intelligence, (AI's capabilities)
	Chatbots	Response quality, in time responding
	Autonomous vehicles	Usefulness
Transparency	Recommendation systems	Causability, interpretability
	Medical image vision	Interpretation
	Production management	Perceived comprehensibility
Representation	AI in service contexts	Anthropomorphism
	Social Robots	Facial characteristics, emotional expressions
Interaction	Voice assistant systems	Interaction quality
	AI-based chatbots	Consumer-AI relation type (assistantship vs. friendship)
	Conversational assistant	Reciprocal self-disclosure
Emotion	Medical diagnostic assistant	Attachment style
Individual features	Online trust game	Individual features (e.g., openness to experience), users' perceptions of social presence

Table 1. Research Studying the Antecedents of Human Trust in AI

These reviews make sense to some extent. However, these reviews ignore either the environmental and task-related factors (Glikson and Woolley 2020) or the development of trust itself (Lotfalian and Bayrak 2021). This gap motivates us to conduct a comprehensive review of research on human trust-AI. In this paper, we review the academic literature of the past ten years to examine the causes and consequences of trust in AI systems. In contrast to existing work, this review differs from them in the following ways: 1) rather than focus on a particular research discipline, we attempt to integrate research from different disciplines; 2) we attempt to further maximize inclusivity by focusing on trust in all forms of AI; and 3) we strive to be comprehensive by including every aspect of human trust in AI including human traits, environmental factors, the trust mechanism (trust-building, and trust calibration) and AI characteristics.

The remainder of this review is as follows. In the next section, we first briefly lay the foundations to establish common ground – i.e., define of what is AI, its representations in real life, and the notion of trust from a multidisciplinary perspective. Next, we outline the review methodology. We then discuss existing models of human trust AI from different aspects: human characteristics and the trust-building process, AI features along with discussions and critiques related to each aspect.

Foundations

Artificial Intelligence

Although humans have had a long period of trying to understand the functions of intelligence, AI was first proposed in 1956 (Karger 2020). In recent decades, managers and scholars have shown renewed interest in the field of artificial intelligence (Brock and Wangenheim 2019). At present, artificial intelligence is a broad and booming domain with various practical applications and research topics.

What is the definition of AI? Different papers have different definitions of AI. The term AI is widely used in many fields such as computer science, engineering, healthcare and management. In a broad sense, it can be defined as a computer program capable of making intelligent decisions (Asan et al. 2020). Such a definition includes computer programs operating using predefined rules and (usually) a model which derived from data. Asan et al. (2020) defined AI as automation, which can be utilized to make explicit and repetitive decisions. Glikson and Woolley (2020) mentioned that, in management studies, AI is defined as a new generation of technologies capable of interacting with the environment by collecting information from outside and environment or from other computer systems; interpreting the gathered information, identifying patterns, generalizing rules or predicting events; producing results, answering questions; and finally, assessing the results of their actions and improving their decision-making systems to achieve better results in the future tasks, which is also referred to as evolution (Ferras-Hernandez 2018). Rossi (2018) stated that AI is one of the scientific disciplines aiming at creating machines that can fulfill many tasks where human intelligence is needed. The futurist literature assumes that AI is a group of algorithms capable of performing nearly all tasks as well as or better than humans (Glikson and Woolley 2020). Russell and Norvig (2009) define AI as machines that can learn knowledge, make decisions and interact with the environment autonomously. Bedué and Fritzsche (2021) define AI as an overarching term for a variety of different design approaches that enable technical devices to perform operations similar to human cognition.

However, such a superintelligence called "strong" or "omnipotent" AI does not exist. Therefore, in this article, we focus on "weak" AI (Raj and Seamans, 2019). Weak AI is based on many technologies that can simulate fragments of human intelligence. To better understand how AI differs from traditional general technologies, it is worth introducing other common components of AI, namely machine learning (ML) and deep learning (DL). The relationships between AI, ML and DL are shown in Figure 1. ML-based technology is driving nearly every aspect of modern society; from economics to management, from daily life to industries (Karger 2020). It is increasingly appearing in products such as digital cameras and smart mobile phones (LeCun et al. 2015). DL is a subset of ML and has grown tremendously in the past decade due to the availability of large data sets and parallel computing power (LeCun et al. 2015). The architecture of DL consists of different modules at multiple levels. Each layer can transform the input data and learn (e.g., new features). DL has shown impressive performance in many fields including speech recognition, face recognition, visual target recognition, and target detection (LeCun et al. 2015).

AI should be differentiated from traditional, non-AI technologies. Even though AI programs apply pre-programmed rules, it *learns* to operate much better than their generator, making better decisions than the program creator could. In other words, AI can "interpret external data correctly, learn from such data, and exhibit flexible adaptation" (Vickers 2017, p. 17). This is an important distinction between AI and prior/traditional technologies whose operations is usually limited and constrained by the knowledge of their creators. Given the unique qualities of AI technology, it is important to differentiate AI from automation, as the two terms are frequently used interchangeably (e.g., Lee and See 2004). Automation is when computers follow pre-defined rules and procedures to repetitively perform tasks that can also be performed by humans (Parasuraman and Riley 1997). Traditional automation produces behaviors that are well-understood and predictable. And the outcomes of the behaviors are generally explainable as they are deterministic (Raj and Seamans 2019). However, automation can also be implemented using AI-based technology, meaning that ML algorithms can set up the rules that can be automated. But more importantly,

these rules can also learn and adapt based on feedback. Therefore, automation plays a role in performing actions determined by AI systems, so this review includes studies that explore trust in intelligent and AI-based automation.

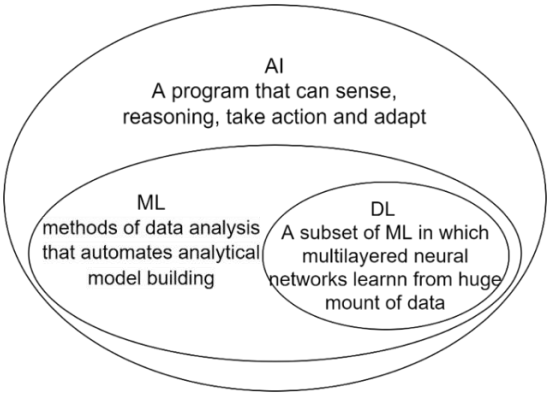

Figure 1. The Relationship between AI, ML, and DL

In this review, we focus on weak AI and use OECD's definition, as recently recommended by AI experts (Krafft et al. 2020, p. 77): "a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments … AI systems are designed to operate with varying levels of autonomy."

Trust

Definition	Author(s)
The conscious adjustment of one's dependence on another	Zand (1972)
A behavior that makes the trustor vulnerable based on the actions of the trustee	Fishbein and Ajzen (1977)
The degree to which a person is willing to attribute goodwill to others and has confidence in words and actions of others	Cook and Wall (1980)
A state involving confident positive expectations about another's motives with respect to oneself in risky situations	Boon and Holmes (1991)
The extent to which a person is confident in and willing to act on the basis of, the words, and actions, of another	McAllister (1995)
The specific expectation that another's actions will be beneficial rather than detrimental and the generalized ability to take for granted … a vast set of characteristics of the social order.	Creed et al. (1996)
The willingness of the trustor to assume a risk based on the cognitive assessment of the trustee' s past experience under uncertain circumstances	Cho et al. (2015)

Table 2. Common Definitions of Trust

What is trust? Again, there is no consensus on the definition of trust. Vereschak et al. (2021) proposed that trust is an attitude, which implies that trust does not systematically translate into behavior. For example, a doctor's level of trust might be sufficiently high to follow advice from an AI, but they may decide not to because none of their colleagues use the system. Social cognitive approaches to defining trust suggest that trust is not so much an attitude as a feeling about something (Castelfranchi and Falcone 2010). Trust cannot always be fully observed by a third party (unless communicated clearly and objectively in oral or written form), which has important implications for the choice of methods for studying trust. For example, this means that observational studies are not enough to draw conclusions about the level of trust among

individuals and should be combined with other methods. Supplementary qualitative methods for assessing trust can be retrospective (i.e., about the past experience) or non-retrospective (i.e., during interactions). Ajenaghughrure et al. (2019) define the relationship between the two entities, i.e., the trustor (e.g., users) and the trustee (e.g., artificial intelligence technology) as a composite cognitive process (psychological prudence, reasoning, and psychological processing which involves memory, learning, and gained knowledge) under the guidance of the trustor based on the following to evaluate the credibility of the trustee: 1) the trust intention of trustors (e.g., prior working experience, the attitudes towards risks) to trust the trustee (AI-based technology); 2) the beliefs of the trustors which is based on observable trustworthiness features (e.g., competence, risk) of the trustee which is in line with expectations of the trustor; and 3) the expected behavior of the trustor towards the trustee (dependency, cooperation, use decision, etc.). Several other definitions of trust are summarized in Table 2. In addition, different disciplines also interpret the meaning of trust differently (see Table 3).

Discipline	**Meaning**	**Author(s)**
Sociology	The subjective probability that, in the absence of uncertainty and ignorance, the other party will perform an act that will not harm my interests	Gambetta (1988)
Philosophy	Risky behavior stemming from a personal and moral relationship between two entities	Lahno and Lagerspetz (1999)
Psychology	Cognitive learning gained from social experience based on the consequences of trusting behaviors	Rotter (1980)
Organizational Management	Willingness to take risks and being fragile to the relationships based on competence, integrity, and kindness	Mayer et al. (1995)
Automation	The attitude that one agent will achieve another agent's goals in the case that incomplete knowledge is endowed with uncertainty	Dzindolet, et al. (2003)
Computing	The estimated subjective probability that an entity will exhibit reliable behavior for a particular operation in a potentially risky situation	Cho et al. (2011)

Table 3. Multidisciplinary Definitions of Trust

Trust In AI

Before introducing how trust works between humans and AI. We first explore how trust works between human beings, i.e., interpersonal trust. Interpersonal trust refers to confidence or belief in another person and a willingness to be vulnerable to him or her (or to each other). Various studies have examined interpersonal trust. Dietz and Den (2006) examined the extent to which measures and operationalization of intra-organizational trust reflect the basic elements of the current conceptualization of trust inside the workplace. Jacovi et al. (2021) proposed that to comprehend human trust in AI (i.e., Human-AI trust), a useful starting point is to examine research in philosophy, psychology, and organizational management on how humans trust other people (i.e., interpersonal trust). They present a basic definition of trust as a transaction between two communities: if A believes that B can fulfill the best interest of A, and accepts vulnerability to the actions of B, then A trusts B (Mayer et al.1995). They stated that the goal of trust is to "make social life predictable [by anticipating the impact of behavior] and make it easier to collaborate between people" (Misztal 2013).

Trust is especially relevant to the relationship between humans and AI because of the perceived risks embedded in this setting, as well as the complexity and uncertainty of AI behavior. AI is also considered a technology that will slowly replace different types of human work and radically change organizational structures (Davis 2019). It is unclear whether low-skilled and low-wage workers, such as front-line service representatives, are at higher risk of being displaced by AI (Huang and Rust 2018; Wirtz et al. 2018) compared to white-collar workers and senior managers who rely on analysis and rational knowledge processing, even though their high costs make their replacement economically attractive (Loebbecke and Pico 2015). Some tasks finished by humans are already being replaced by AI.

Literature Review Methodology

This paper reviews the ways in which AI trust is currently discussed in the literature in the fields of computer science (CS), human-computer interaction (HCI), information systems (IS), AI-based robotics, management, marketing, philosophy, and psychology. To narrow our focus on human trust in AI, we first searched Google Scholar with the following keywords: AI, human trust AI, human AI, robot-human interaction, intelligent automation, trust in robots, and trust in technology. We collected literature published in the last 10 years (from 2010 to 2020) to address related work that has accompanied recent developments in AI technology. The search retrieved around 400 peer-reviewed papers from the fields of organizational behavior, HCI, robot-human interaction, IS, information technology, and engineering. We then filtered articles based on content, including those related to human trust in AI, and excluding descriptions of algorithms/architectures (i.e., not involving trust). After exclusion, around 300 of the above articles remained. We also include review papers published recently that focus on trust in AI technology.

A Framework of Human Trust in AI

This review is organized by focusing on four foundational factors related to human trust in AI: 1) environmental factors, 2) human-related factors, 3) trust-building and calibration-related factors, and 4) AI characteristics. Specifically, for AI characteristics, we mainly focus on transparency and interpretability, 2) Performance: reliability and accuracy, 3) anthropomorphism and humanness, and 4) other factors (i.e., usefulness/helpfulness, privacy, and automation). In each subsection, we start by presenting prior findings and then review existing articles on each dimension, followed by how they impact stakeholders and possible future research directions. The review framework of this article is shown in Figure 2.

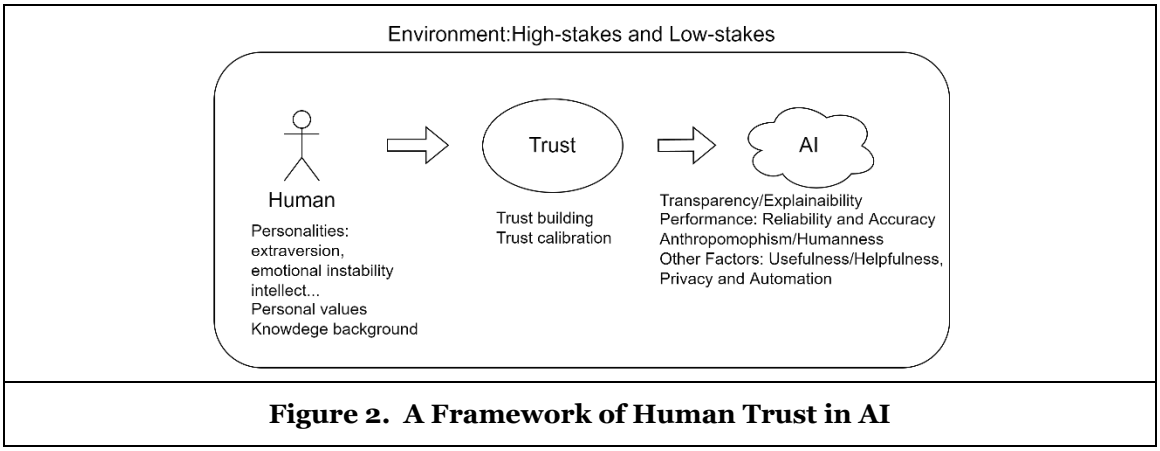

Figure 2. A Framework of Human Trust in AI

Environmental Factors

In this review, the environmental factors that impact human trust in AI mainly refer to the nature of the task, the culture, and institutional factors. The environmental factors have been characterized as high vs. low stakes. Akula (2021) categorized task factors from low-stakes to high-stakes such as AI-based chatbots, medical diagnosis and therapy, autonomous vehicles, and military applications (i.e., from low to high risk).

It has been proposed that physical and external environmental factors can influence the ability of users to leverage AI, comprehend the performance of AI systems, and finally build trust (Benda et al. 2022). The physical settings (e.g., hospital overall arrangement, facility siting, resource access) may not be as relevant to medical care-related AI systems as physical automation systems in other fields (e.g., - transportation, industrial environments) or other issues related to patient safety (e.g., wrong medication) (Benda et al. 2022). However, other external inducements (e.g., government policies) may have important implications on trust. For instance, the baseline level of trust may be influenced by the supporters and champions of AI systems (e.g., professional medical organizations).

Wang and Siau (2018) argued that the nature of healthcare tasks (e.g., importance and complexity) will influence people's willingness to comply with agent directives (Salem et al. 2015). Ingrams et al. (2021)

showed that highly complex tasks produce decisions that are more inefficient, regardless of whether people or AIs are in charge. They proposed that researchers and practitioners are more concerned with the balance between instrumentality and value-based quality when designing and implementing AI applications. They also found that people view AI-led decisions as less red tape and less trustworthy than human ones. Wang and Siau (2018) examined whether trust in suppliers and trust in technologies are preconditions for trust in the context of autonomous vehicles (AVs) which is treated as a high-risk application of AI. Wanner et al. (2020) stated that decisions towards adopting complex technologies, such as AI, are very complex and are subject to a range of cognitive biases, especially for applications that involve high-stakes decisions and therefore are of high criticality (Das and Teng 1999). They treat high-stakes decisions as those where a wrong or late decision can result in the loss of human lives. Ajenaghughrure et al. (2019) studied how risk affects users' trust before and after users interact with technologies such as self-drive cars. Through experimental analysis, they demonstrated that in different risk situations, users delegate control to AVs in different ways.

Environmental factors are important. However, they are external factors. A more reasonable way may be to combine internal factors and external environmental factors together and study how external factors influence which perspective of trust in AI so as to reduce or improve people's trust in AI.

The Human Perspective

As the main subject of trust (i.e., the trustor), human features are very important in the analysis of human trust in AI. There are many studies examining and exploring how human features impact trust in AI. Bawack et al. (2021) explored how personality and prior experiences affect trust in AI-enabled voice assistants. They found that: 1) extraversion, emotional instability, and intellect had no significant effect on trust; 2) the effect of agreeableness on human trust in AI-enabled voice assistants is significant positive; 3) conscientiousness has a significant effect on trust; 4) people with rigorous nature would build trust based on specific actions of the companies. Mehrotra et al. (2017) pointed out the current research gap on the role of personal value in trust in AI. Therefore, this paper studied how human and agent value similarity (VS) affects human trust in that agent. They find that the more similar the values, the more trust. Crockett et al. (2020) evaluated whether the general public's perceived risk and trust in AI differed from those who studied computer science in tertiary education, by defining the general public as individuals with no expertise in AI but at high risk as potential users of AI systems. The results indicate that opinions on risk differed significantly between the two groups in particular applications. Both sides strongly believed that education on how the AI system works was vital in trust building. Hassenzahl and Tractinsky (2006) proposed that user experience (UX) which was defined as the intersection of pleasure, emotion, and experiential facets is a significant variable to be studied in the process of human-machine interactions. UX is one of the main drivers which strengthen technology interactions. This study builds off of previous technology trust research and seeks to understand how humans are predisposed to either trust or distrust AI via collecting 762 respondents from Amazon MTurk and San Francisco State University and analyzed self-report data for personality, risk experience, disposition to trust, technology trusting beliefs, perceptions, and intentions. The study mainly focused on two humanness factors embodied by conversational AI (i.e., speaking and listening) that affect how humans trust AI-based chatbots (Hu and Lu 2021). The authors explored users' heterogeneous perception patterns based on the two humanness factors and studied how this heterogeneity is relevant to trust in conversational AI. These studies examined the impact of different levels of education on trust in AI, and the impact of similarity between users and AI on trust in AI. It might make sense to study why the same types of people trust AI differently.

The Trust Perspective

In this section, we conduct an analysis of how the trust between humans and AI is developed. Specifically, the trust development process contains initial trust-building and continuous trust adjustment, which is also called trust calibration (Hoffman et al. 2018).

Trust-building

Trust is dynamic. It can easily change according to the behavior of trusted agents (Glikson and Woolley 2020). Hoff and Bashir (2015) hypothesized that people's trust in technology unfolds differently from the development of trust in other humans due to a general positive preference for new technologies. The trustor

always needs to decide whether to trust the trustee based on the trustee's disposition, subjective norms, and the trustee's reputation before they have first-hand information about the trustee (McKnight et al. 1998). Such trust is called initial trust (Siau et al. 2004). After acquiring first-hand information, interaction is needed to build trust in a gradual manner (Kim 2012) resulting in continuous trust (Siau and Shen 2003).

Several factors are at play during trust-building, including the representation of AI, reviews from other users and trialability. Objective representation can build trust. This notion is supported by several studies on AI. Rossi (2018) proposed that representation plays an important role in building initial trust. He stated that the more similar a robot is to humans, the more emotional connections will be created between humans and AI. For example, humans find it much easier to trust a robot dog, where the dog is conventionally a loyal friend of humans. The image and perception will impact how people build initial trust with AI and reviews from other people will also influence the initial trust level. Moreover, it was found that opportunities to try a new AI application (i.e., trialability) can promote initial trust to a higher level. Dzindolet et al. (2003) and De Visser et al. (2016) also stated that representations are important for initial trust. However, their findings seem to be in contrast with Rossi (2018). They find that users hold higher than average initial trust in automatic decision assistance. In other words, initial trust in a machine-like computer system was higher than that in a human-like computer system. Maadi et al. (2016) put forward that to build initial trust, one important thing to overcome is the perceptions of risk (McKnight et al. 2002), which will create a possible willingness to accept the technologies. They mentioned that reducing fear and perceptions of risk (by influencing emotional trust) may be more effective than providing information about potential errors, especially in the initial stages of trust. Inclination and personality are also important factors in building trust. Personality has a significant impact on initial trust-building when no first-hand knowledge has been gained (McKnight 1998). Aoki (2021) investigates the initial trust of the public in AI-based decision assistance via conducting online experiment testing in the context of the Japanese long-term nursing care industry. They verified that "humans are still in the decision loop" (HDL) makes a difference in the initial trust of the public.

Another path to building trust is trust transfer. In nature, trust transfer theory illustrates the relationship between known trusted sources and new unknown targets. Studies have shown that when the relationship between a target and a trusted source is relatively strong, users' trust in familiar trusted sources can be transferred to a relatively unfamiliar target. Therefore, trust transfer is a basic form of trust adjustment between two objects. Renner (2021) showed that if users recognize the functionalities or providers of an AI technology, they may also transfer their beliefs to new unknown AI-capable technologies once they put them in the same category. Wang and Singh (2007), on the other hand, proposed that trust cannot be trivially propagated such that A may trust B who trusts C, but A may not trust C. This seemingly contradictory view makes it worthwhile to investigate to what extent and under what conditions trust can be transferred from objects to AI.

Trust Calibration

Trust in AI is a long-term process. After trust-building, it is important to maintain proper trust in AI as the environment may change over time. Successful collaborations between end-users and AI-based agents require users to appropriately regulate their trust levels with the actual reliability of the agents (Okamura and Yamada 2020). This process is called trust calibration, i.e., avoid over-trust in AI when it is actually incorrect, or under-trust in AI when it is correct in fact (Naiseh et al. 2021; Lee and See 2004).

AI tools are often seen as closed and opaque to human decision-makers (Naiseh et al. (2021). A necessary requirement for their success is to be able to provide users with an understanding and meaningful explanation of themselves. While explanations often have positive connotations, research suggests that the assumptions behind user interactions and engagement with these explanations can introduce trust calibration biases, such as promoting unreasonable or less thoughtful agreement or disagreement with AI recommendations. Thus, Naiseh et al. (2021) explored how we can help with trust calibration by explaining interaction design. They suggest that the interpretation of trust calibration should focus primarily on eliminating undesirable behaviors, such as skipping interpretation. Future work should focus on striking a balance between explaining trust calibration in a sufficiently effective manner without potentially damaging the user experience and being seen as a tool of persuasion rather than an aid to critical thinking. Recent studies revealed that explainable AI (XAI) fails to support trust calibration because they assume that users would accept explanations cognitively and interpret them without bias. Naiseh (2021) hypothesized that

XAI interaction design can help users' cognitive engagement with XAI and then enhance trust calibration. They proposed friction as a boost-based method to help users to calibrate their trust in AI. Motivated by the fact that it is still not clear why explanations do not support trust calibration, Pieters (2011) explored the situations where trust calibration failed in the presence of explanations. Zhang et al. (2020) suggested that local, predictor-specific explanations may not have a perceptible effect on trust calibration, even if they are theoretically valid for such a task. It would be worthwhile to further investigate to what extent and what explanatory properties can improve trust calibration.

The AI Perspective

In this section, the studies on how different AI features including transparency-explainability, performance, humanness, fairness, and ethical issues affect human trust in AI are briefly reviewed.

Transparency/Interpretability

Despite AI's significant growth, AI is often thought of as a black box (Adadi and Berrada 2018; Castelvecchi 2016) and its "black box" nature poses challenges in generating sufficient trust. Advanced algorithms (such as DL) are inherently opaque and thus unexplainable. The solution to the black box issues is to create self-explanatory AI whose decisions and predictions are transparent. However, there is a tension between accuracy and explainability, as the best-performing models are often the least transparent and hardest to interpret, while the models that are capable of providing the clearest explanation are typically the least accurate. Thus, there are ongoing research efforts dedicated to making AI more transparent and easier to explain, with the ultimate goal of improving user trust.

A general definition of transparency in the context of organizations is "the availability of information about an organization or actor that allows external actors to monitor the internal workings or performance of that organization" (Grimmelikhuijsen et al. 2013, p. 2). For AI systems, transparency reflects the extent to which users understand the basic operating rules and internal logic of technology and is considered to be key to developing trust in new technologies (Hoff and Bashir 2015). However, AI entails more problems concerning transparency compared to other technologies, especially when it comes to approaches such as DL. Transparency contains different types of interpretations of how AI works or why it makes particular decisions that users can understand, even if they have little technology background knowledge (Glikson and Woolley 2020). Many studies have theoretically or empirically verified that transparency and interpretability of AI applications are conducive to trust. In the context of healthcare, transparency and interpretability are critical. The outcomes of AI-based clinical systems should be transparent and interpretable by end-users. However, full transparency is extremely difficult to achieve in reality. Instead, varying degrees of transparency can be used based on factors such as the level of risk and clinicians' ability to evaluate decisions (Gretton 2018). Winfield and Jirotka (2018) proposed that transparency necessarily varies for different stakeholders – e.g., safety certification bodies or accident investigators clearly require different extents of transparency compared to system operators. Ideally, AI systems should be able to explain their own behaviors to non-experts and be transparent to experts.

Pavlin et al. (2021) proposed that there are many types of explanations, such as explainability of models, explainability of processes, explainability of data, explainability of sensors/sources, explainability of utility/cost, and explainability of actions. The authors explored how various types of explainabilities impact trust in AI systems which is based on the assumption that explainability promotes different types of evaluation that ultimately affect trust. However, Schmidt et al. (2020) argued that to what extent explanations actually help to promote trust in AI systems remains unknown. They conducted an experimental analysis with ML-based decision support tools for text classification. Kizilcec (2016) showed similar findings in the context of AI grading for students. They found that providing more transparency on the grading (i.e., how to calculate the raw grades) did not increase trust, which indicates that the type and amount of transparency play a role. Future studies should focus on analyzing what aspect of transparency can increase which types of users' (i.e., users with varying levels of knowledge of AI) trust in AI.

Performance: Reliability and Accuracy

Several research showed that user trust decreases as reliance, accuracy or performance decreases (De Visser et al. 2018; Yin et al. 2019). Reliability, or the performance of the same and expected behavior over time, is

also vital to technical credibility (Hoff and Bashir 2015). With AI, evaluating reliability is a tough task, especially in the context of high machine intelligence, because learning from data can result in different performance of the technology, even if the objective function never changes.

Several studies examined how the reliability of AI directly impacts human trust in AI. People will lose trust in the advice of an AI system that made mistakes in high-risk situations (Robinetee et al. 2017). Desai et al. (2013) proposed that the timing of a robot's failure is important and an early decrease in reliability reduced more real-time trust than later drops. Kumar (2020) stated that the impact of an error on trust depends both on the magnitude of the error and how unexpected it was. A small, unpredictable mistake can affect trust more than a large, frequent mistake.

The accuracy of AI systems is also critical for human trust in AI. Inaccurate results derived by AI systems can lead to inequality and harm. Kumar (2020) showed that higher accuracy models are preferred to lower accuracy models. Papenmeier et al. (2019) conducted a user study to investigate the effects of model accuracy and explanation fidelity on trust. They found that accuracy is more significant for user trust than explainability in affecting human trust in AI. However, AI systems can be configured to optimize various measurements of accuracy and may have high accuracy for specific predictions (such as white male results) but not for others (such as minority groups). Therefore, accuracy metrics alone may not be sufficient to gain trust in an AI application. Even though the outcomes of an AI system are accurate, users may still distrust it (Moran et al. 2013). However, it is necessary that users perceive that the results of the AI are accurate. For example, teams involved in large, street-based games often do not trust the (completely accurate) information provided by automated suggestions and may even ignore it, even though they are told that following this information is critical to their performance (Moran et al. 2013). That said, some researchers showed that users may also follow instructions derived by an inaccurate AI agent (Salem et al. 2015).

All in all, although most studies showed that accuracy has a positive effect on trust, this relationship is not universally maintained and deserves further study, especially for high-stakes tasks because the cost of inaccurate judgment and prediction is relatively high. As previously mentioned, accuracy alone may not be sufficient to induce human trust in AI. In other words, accuracy should work together with explainability and other factors to induce human trust in AI. However, the correlations between explainability and accuracy and other factors might need to be further explored empirically.

Anthropomorphism / Humanness

There is general agreement that anthropomorphism is a key feature that distinguishes AI from non-AI applications (Lockey et al. 2021). Anthropomorphism (human-likeness) involves the inclusion of human-like characteristics into the design of AI. Generally speaking, anthropomorphism has a positive effect on human perceptions and emotions. It has been theorized that the trust between humans and AI is positively related to the human-likeness of the AI agent (Waytz et al. 2014).

Even though some studies have anxiety that over-anthropomorphism may damage trust or lead to ethical and psychological issues (Baker et al. 2018), empirical research broadly supports the notion that anthropomorphism increases trust in AI, which has been shown in the context of AVs (Waytz et al. 2014). People show more trust in AVs with human characteristics than those without human features. Similarly, in the field of virtual agents, a study on the effect of virtual agent anthropomorphism on reliability reduction shows that although anthropomorphism decreases initial expectations, it increases trust elasticity (Kim et al. 2018). When performance deteriorates, trust declines more significantly for robotic agents compared with the agents with human-likeness. Studies have also shown that augmented reality and 3D agents are considered more trustworthy than traditional 2D interfaces (Huynh et al. 2018).

However, some studies provide evidence for the negative effect of anthropomorphism. In the context of robots, some researchers, based on the "Uncanny Valley" theory (Ho and MacDorman 2010), argue that encountering a robot with excessively human-like characteristics leads to unpleasant feelings (Ho and MacDorman 2010). For instance, there is a study that investigated the human-likeness of nursing robots and found that human-like robots are less trustworthy than more machine-like robots. Zlotowski et al. (2016) found similar results when examining human interactions with machine-like or human-like robots. It is the machine-like robot that achieved more empathy and more trustworthiness compared to a human-like robot. Another negative effect of anthropomorphism is that it may induce a higher level of anxiety in users (Appel et al. 2016).

However, most of the studies that showed a negative effect of human-likeness were concerned with initial trust (i.e., at the trust-building stage), which means these findings are derived from short-term interactions. The mismatch between robot appearance and machine intelligence is also likely to contribute significantly to these negative perceptions. In the long run, whether these negative effects remain significant is still unknown and calls for further research. Further research also needs to investigate when and how AI human-likeness enhances trust. AI is computer-based technology. And with human-like features of AI, some research applied computer as social actor (CASA) theory to AI. However, what challenges remain when applying CASA theory to AI is still not very clear.

Other Factors: Usefulness/Helpfulness, Privacy, and Automation.

According to Davis (1989), perceived usefulness is the extent to which individuals believe that using technology can improve the performance of their activities. The prior literature in the field of health informatics has consistently found that individuals who find health information systems useful are more likely to trust and accept the technology (Wang and Siau 2018).

It has been shown in a broad variety of contexts that the constructs of the technology acceptance model (TAM; Davis 1989), namely, perceived usefulness and perceived ease of use, play a key role in the development of trust in AI (Ejdys 2018). For example, evidence suggests that perceived usefulness is an important factor in consumer trust in autonomous vehicles (Zhang et al. 2019). In addition, if a system is visually attractive and easy to use, users may build a higher level of trust in it because such interface design reflects the competence and professionalism of the developer (Pengnate and Sarathy 2017). Liu and Tao (2022) also verified that the effects of perceived usefulness and ease to use on trust are positive. Generally, usefulness and helpfulness should be evaluated based on specific tasks. At the same time, whether usefulness and helpfulness are highly correlated with user knowledge level needs to be further examined.

Guo and White (2016) pointed out that personalization and loss of privacy are two important determinants of trust in AI in the healthcare context. Lockey (2021) proposed that adaptive automation, which means that automation is not fixed at the design stage but adapts to the situation, can improve the trust in robots in collaborative tasks more than without automation or static automation. Typically, for different stakeholders, privacy means different things. For instance, accountability for privacy and use of data are typically the primary concerns; for end users, privacy implies that personal data should be protected and not be inappropriately used; for society, surveillance should be implemented appropriately, and power should be balanced (Lockey et al. 2021). To what extent do different levels of privacy awareness of different stakeholders impact human trust AI might also need further exploration.

Conclusion

AI and AI applications are increasingly becoming prevalent. Trust is important in the development and acceptance of AI. In this review, the antecedents of trust in AI are roughly grouped into four perspectives: environmental factors, the human-user perspective, the trust perspective, and the AI perspective.

To enhance trust, practitioners need to pay attention to maximizing the technological features in AI systems, such as transparency, accuracy, and anthropomorphism. At the same time, environmental factors should also be taken into account. Trust is dynamic, it is not only critical to build initial trust but to adjust the continuous trust.

This review is not without limitations. First, we do not take into consideration the interactions between the different factors from different perspectives. Another limitation is that it does not differentiate developing trust in human beings and in technologies. These are meaningful areas for further study.

Acknowledgements

The authors acknowledge the generous support from the AI Governance Pillar at AI Singapore. The authors also thank the members of the Garbage Can Lab at NUS School of Computing for constructive comments and valuable suggestions on the research.

References

Adadi, A., and Berrada, M. 2018. "Peeking inside the Black-box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access* (6), 52138-52160.

Ajenaghughrure, I.B., Sousa, S.C., Kosunen, I. J., and Lamas, D. 2019. "Predictive Model to Assess User Trust: A Psycho-physiological Approach," In *Proceedings of the 10th Indian Conference on Human-Computer Interaction*, Hyderabad, India, Nov 1-3, pp. 1-10.

Akula, A.R. 2021. "Gaining Justified Human Trust by Improving Explainability in Vision and Language Reasoning Models," University of California, Los Angeles.

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.

Appel, H., Gerlach, A.L., and Crusius, J. 2016. "The Interplay between Facebook Use, Social Comparison, Envy, and Depression," *Current Opinion in Psychology* (9), pp. 44-49.

Asan, O., Bayrak, A.E., and Choudhury, A. 2020. "Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians," *Journal of Medical Internet Research* (22:6), e15154.

Baker, A.L., Phillips, E K., Ullman, D., and Keebler, J.R. 2018. "Toward An Understanding of Trust Repair in Human-Robot Interaction: Current Research and Future Directions," *ACM Transactions on Interactive Intelligent Systems* (8:4), pp. 1-30.

Bao, Y., Cheng, X., De Vreede, T., and De Vreede, G.J. 2021. "Investigating the Relationship between AI and Trust in Human-AI Collaboration," *Hawaii International Conference on System Science*. Kauai, HI. Jan 5-8.

Bawack, R.E., Wamba, S.F., and Carillo, K.D.A. 2021. "Exploring the Tole of Personality, Trust, and Privacy in Customer Experience Performance during Voice Shopping: Evidence from SEM and Fuzzy Set Qualitative Comparative Analysis," *International Journal of Information Management* (58), 102309.

Bedué, P., and Fritzsche, A. 2021. "Can We Trust AI? An Empirical Investigation of Trust Requirements and Guide to Successful AI Adoption," *Journal of Enterprise Information Management* (35:2), pp. 530-549.

Benda, N.C., Novak, L.L., Reale, C., and Ancker, J.S. 2022. "Trust in AI: Why We Should be Designing for Appropriate Reliance," *Journal of the American Medical Informatics Association*, (29:1), pp. 207-212.

Boon, S. D., and Holmes, J. G. 1991. "The Dynamics of Interpersonal Trust: Resolving Uncertainty in the Face of Risk," In R. Hinde and J. Gorebel (eds.), *Cooperation and Prosocial Behaviour*, Cambridge University Press, pp. 190-211.

Bowen, J., and Morosan, C. 2018. "Beware Hospitality Industry: The Robots are Coming," *Worldwide Hospitality and Tourism Themes* (10:6), pp. 726–733.

Braynov, S., and Sandholm, T. 2002. "Contracting with Uncertain Level of Trust," *Computational Intelligence* (*18*:4), pp. 501-514.

Brock, J.K.U., and Von Wangenheim, F. 2019. "Demystifying AI: What Digital Transformation Leaders can Teach You about Realistic Artificial Intelligence," *California Management Review* (61:4), pp. 110-134.

Castelfranchi, C., and Falcone, R. 2010. "Trust Theory: A Socio-cognitive and Computational Model," John *Wiley*.

Castelvecchi, D. 2016. "Can We open the Black Box of AI?" *Nature News*, (538:7623), pp. 20.

Chi, O.H., Denton, G., and Gursoy, D. 2020. "Artificially Intelligent Device Use in Service Delivery: A Systematic Review, Synthesis, and Research Agenda," *Journal of Hospitality Marketing and Management* (29:7), pp. 757-786.

Cho, E., Myers, S.A., and Leskovec, J. 2011. "Friendship and Mobility: User Movement in Location-based Social Networks," In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA, Aug 21-24, pp. 1082-1090.

Cho, J.H., Chan, K., and Adali, S. 2015. "A Survey on Trust Modeling," *ACM Computing Surveys* (48:2), pp. 1-40.

Cook, J., and Wall, T. 1980. "New Work Attitude Measures of Trust, Organizational Commitment and Personal Need Non–fulfilment," *Journal of Occupational Psychology* (53:1), pp. 39-52.

Creed, W.D., Miles, R.E., Kramer, R.M., and Tyler, T.R. 1996. "Trust in Organizations," *Trust in Organizations: Frontiers of Theory and Research*, pp. 16-38.

Crockett, K., Garratt, M., Latham, A., Colyer, E., and Goltz, S. 2020. "Risk and Trust Perceptions of the Public of Artificial Intelligence Applications," In *2020 International Joint Conference on Neural*

Networks, Glasgow, Scotland, Jul 19-24, pp. 1-8.
Das, T.K., and Teng, B.S. 1999. "Managing Risks in Strategic Alliances," *Academy of Management Perspectives* (13: 4), pp. 50-62.
Davis, F.D. 1989. "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Quarterly* (13:3), pp. 319-340.
Davis, G.F. 2019. "How to Communicate Large-scale Social Challenges: The Problem of the Disappearing American Corporation," *Proceedings of the National Academy of Sciences* (116:16), pp. 7698-7702.
Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., and Yanco, H. 2013. "Impact of Robot Failures and Feedback on Real-time Trust," In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction,* Tokyo, Japan, Mar 3-6, pp. 251-258.
De Visser, E.J., Beatty, P.J., Estepp, J.R., Kohn, S., Abubshait, A., Fedota, J.R., and McDonald, C.G. 2018. "Learning from the Slips of Others: Neural Correlates of Trust in Automated Agents. *Frontiers in Human Neuroscience* (12), pp. 309.
Dietz, G., and Den Hartog, D.N. 2006. "Measuring Trust Inside Organisations," *Personnel Review* (35:5), pp. 557-588.
Doran, P., Tamma, V., and Iannone, L. 2007. "Ontology Module Extraction for Ontology Reuse: An Ontology Engineering Perspective. In *Proceedings of the Sixteenth ACM Conference on information and Knowledge Management,* Lisbon, Portugal, Nov 6-10, pp. 61-70.
Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., and Beck, H. P. 2003. "The Role of Trust in Automation Reliance," *International Journal of Human-Computer Studies* (58:6), pp. 697-718.
Ejdys, J. 2018. "Building Technology Trust in ICT Application at a University," *International Journal of Emerging Markets* (13), pp. 980-997.
Ferràs-Hernández, X. 2018. "The Future of Management in A World of Electronic Brains," *Journal of Management Inquiry* (27:2), pp. 260-263.
Fishbein, M., and Ajzen, I. 1977. "Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research," *Philosophy and Rhetoric* (10:2), pp. 130-132.
Gambetta, D. 1988. *Trust: Making and Breaking Cooperative Relations*, Blackwell
Gillath, O., Ai, T., Branicky, M.S., Keshmiri, S., Davison, R.B., and Spaulding, R. 2021. "Attachment and Trust in Artificial Intelligence," *Computers in Human Behavior* (115), 106607.
Glikson, E., and Woolley, A.W. 2020. "Human Trust in Artificial Intelligence: Review of Empirical Research," *Academy of Management Annals*, (14:2), pp. 627-660.
Gretton, C. 2018. "Trust and Transparency in Machine Learning-Based Clinical Decision Support," In J. Zhou and F. Chen (eds.), *Human and Machine Learning*, pp. 279-292.
Grimmelikhuijsen, S., Porumbescu, G., Hong, B., and Im, T. 2013. "The Effect of Transparency on Trust in Government: A Cross-National Comparative Experiment," *Public Administration Review* (73:4), pp. 575-586.
Groom, V., and Nass, C. 2007. "Can Robots be Teammates?: Benchmarks in Human–robot Teams," *Interaction Studies* (*8:*3), pp. 483-500.
Guo, J.Y., and White, E. 2016. "Autophagy, Metabolism, and Cancer," *In Cold Spring Harbor Symposia on Quantitative Biology* (81), pp. 73-78.
Gursoy, D., Chi, O.H., Lu, L., and Nunkoo, R. 2019. "Consumers Acceptance of Artificially Intelligent (AI) Device Use in Service Delivery," *International Journal of Information Management* (49), pp. 157-169.
Hassenzahl, M., and Tractinsky, N. 2006. "User Experience: A Research Agenda," *Behaviour and Information Technology* (25:2), pp. 91-97.
Ho, C.C., and MacDorman, K.F. 2010. "Revisiting The Uncanny Valley Theory: Developing and Validating an Alternative to the Godspeed Indices," *Computers in Human Behavior* (26:6), pp. 1508-1518.
Hoff, K.A., and Bashir, M. 2015. "Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust," *Human Factors* (57:3), pp. 407-434.
Hoffman, R.R., Mueller, S.T., Klein, G., and Litman, J. 2018. "Metrics for Explainable AI: Challenges and Prospects," *arXiv preprint arXiv*:1812.04608.
Hu, P., and Lu, Y. 2021. "Dual Humanness and Trust in Conversational AI: A Person-Centered Approach," *Computers in Human Behavior* (119), 106727.
Huang, M., and Rust, R.T. 2018. "Artificial Intelligence in Service," *Journal of Service Research* (21:2), pp. 155–172.
Huynh, B., Ibrahim, A., Chang, Y.S., Höllerer, T., and O'Donovan, J. 2018. "A Study of Situated Product Recommendations in Augmented Reality," *In 2018 IEEE International Conference on Artificial*

Intelligence and Virtual Reality, Taichung, Taiwan, Dec 10-12, pp. 35-43.
Ingrams, A., Kaufmann, W., and Jacobs, D. 2021. "In AI We Trust? Citizen Perceptions of AI In Government Decision Making," *Policy and Internet*.
Jacovi, A., Marasović, A., Miller, T., and Goldberg, Y. 2021. "Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI.," In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event, Canada, Mar 3-10, pp. 624-635.
Jeffries, F.L., and Reed, R. 2000. "Trust and Adaptation in Relational Contracting," *Academy of Management Review* (*25:*4), pp. 873-882.
Kamal, S.A., Shafiq, M., and Kakria, P. 2020. "Investigating Acceptance of Telemedicine Services Through an Extended Technology Acceptance Model (TAM)," *Technology in Society* (60), 101212.
Kaplan, A.D., Kessler, T.T., Brill, J.C., and Hancock, P.A. 2021. "Trust in Artificial Intelligence: Meta-Analytic Findings," *Human Factors*, in press.
Karger, E. 2020. "Combining Blockchain and Artificial Intelligence-Literature Review and State of the Art," In *Proceedings of the 41st International Conference on Information Systems*, Hyderabad, India, Dec 13-16.
Kerasidou, A. 2021. "Ethics Of Artificial Intelligence in Global Health: Explainability, Algorithmic Bias and Trust," *Journal of Oral Biology and Craniofacial Research* (11:4), pp. 612-614.
Kim, J.B. 2012. "An Empirical Study on Consumer First Purchase Intention in Online Shopping: Integrating Initial Trust and TAM," *Electronic Commerce Research* (12:2), pp. 125-150.
Kim, K., Boelling, L., Haesler, S., Bailenson, J., Bruder, G., and Welch, G.F. 2018. "Does A Digital Assistant Need a Body? The Influence of Visual Embodiment and Social Behavior on the Perception of Intelligent Virtual Agents in AR," *In 2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Munich, Germany, Oct 16-20, pp. 105-114.
Kizilcec, R.F. 2016. "How Much Information? Effects Of Transparency on Trust in An Algorithmic Interface," *In Proceedings of the 2016 Conference on Human Factors in Computing Systems,* San Jose, CA, May 7-12, pp. 2390-2395.
Krafft, P.M., Young, M., Katell, M., Huang, K., and Bugingo, G. 2020. "Defining AI in Policy Versus Practice," *In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Oxford. Aug 1-3, pp. 72-78.
Kumar, A. 2020. "Trust between Humans and AI: A Short Review"
Lahno, B., and Lagerspetz, O. 1999. "Trust. The Tacit Demand," *Ethical Theory and Moral Practice* (2), pp. 433-435.
LeCun, Y., Yoshua B., and Geoffrey H. 2015. "Deep learning," *Nature, (*521:7553), pp. 436-444.
Lee, J.D., and See, K.A. 2004. "Trust In Automation: Designing for Appropriate Reliance," *Human factors* (46:1), pp. 50-80.
Liu, K., and Tao, D. 2022. "The Roles of Trust, Personalization, Loss of Privacy, and Anthropomorphism in Public Acceptance of Smart Healthcare Services," *Computers in Human Behavior* (127), 107026.
Lockey, S., Gillespie, N., Holm, D., and Someh, I.A. 2021. "A Review of Trust in Artificial Intelligence: Challenges, Vulnerabilities and Future Directions," *Hawaii International Conference on System Science*. Kauai, HI, Jan 5-8,
Loebbecke, C., and Picot, A. 2015. "Reflections on Societal and Business Model Transformation Arising from Digitization and Big Data Analytics: A Research Agenda," *Journal of Strategic Information Systems* (24:3), pp. 149–157.
Lotfalian S.M., and Bayrak, A.E. 2021. "A Survey of Important Factors in Human-Artificial Intelligence Trust for Engineering System Design," In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*.
Maadi, M., Marjan M., and Mohammad J. 2016. "Identification of Factors Influencing Building Initial Trust in E-Commerce," *Iranian Journal of Management Studies* (9:3), pp. 483-503.
MacDorman, K.F., Coram, J.A., Ho, C.C., and Patel, H. 2010. "Gender Differences in The Impact of Presentational Factors in Human Character Animation on Decisions in Ethical Dilemmas," *Presence* (19:3), pp. 213-229.
Makridakis, S. 2017, "The Forthcoming Artificial Intelligence (AI) Revolution: Its Impact on Society and Firms," *Futures* (90), pp. 46–60.
Mayer, R.C., Davis, J.H., and Schoorman, F.D. 1995. "An Integrative Model of Organizational Trust," *Academy of Management Review* (20:3), pp. 709-734.
McAllister, D.J. 1995. "Affect-And Cognition-Based Trust as Foundations for Interpersonal Cooperation in Organizations," *Academy of Management Journal* (38:1), pp. 24-59.

McKnight, D.H., Choudhury, V., and Kacmar, C. 2002. "The Impact of Initial Consumer Trust on Intentions to Transact with a Web Site: A Trust Building Model," *Journal of Strategic Information Systems* (11:3-4), pp. 297-323.
McKnight, D.H., Cummings, L.L., and Chervany, N.L. 1998. "Initial Trust Formation in New Organizational Relationships," *Academy of Management Review* (23:3), pp. 473-490.
Mehrotra, S., Jonker, C.M., and Tielman, M.L. 2021. "More Similar Values, More Trust? -The Effect of Value Similarity on Trust in Human-Agent Interaction," In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society,* May 19-21, pp. 777-783.
Misztal, B. 2013. "Trust in Modern Societies: The Search for the Bases of Social Order," John Wiley.
Moran, S., Pantidi, N., Bachour, K., Fischer, J.E., Flintham, M., Rodden, T., and Johnson, S. 2013. "Team Reactions to Voiced Agent Instructions in a Pervasive Game," *In Proceedings of the 2013 International Conference on Intelligent User Interfaces*, Santa Monica, CA, Mar 19-22, pp. 371-382.
Naiseh, M., Al-Mansoori, R.S., Al-Thani, D., Jiang, N., and Ali, R. 2021. "Nudging through Friction: An Approach for Calibrating Trust in Explainable AI," *In 2021 8th International Conference on Behavioral and Social Computing,* Doha, Qatar, Oct 29-31, pp. 1-5.
Naiseh, M., Al-Thani, D., Jiang, N., and Ali, R. 2021. "Explainable Recommendation: When Design Meets Trust Calibration," *World Wide Web* (24:5), pp. 1857-1884.
Naiseh, M., Cemiloglu, D., Al Thani, D., Jiang, N., and Ali, R. 2021. "Explainable Recommendations and Calibrated Trust: Two Systematic User Errors," *IEEE Computer* (54:10), pp. 28-37.
Nasirian, F., Ahmadian, M., and Lee, O.K.D. 2017. "AI-Based Voice Assistant Systems: Evaluating from the Interaction and Trust Perspectives," In *Proceedings of the 2017 Americas Conference on Information Systems.*
OECD. 2016. "Artificial Intelligence on Society," OECD Publishing, Paris, 2019. Retrieved from https://www.oecd-ilibrary.org/.ors, 58(3), pp. 377-400.
Okamura, K., and Yamada, S. 2020. "Adaptive Trust Calibration for Human-AI Collaboration," *PLOS One* (15:2), e0229132.
Okuda, T., and Shoda, S. 2018 "AI-based Chatbot Service for Financial Industry," *Fujitsu Scientific and Technical Journal* (54:2), pp. 4-8.
Ostrom, A.L., Fotheringham, D., and Bitner, M.J. 2019. "Customer Acceptance of AI in Service Encounters: Understanding Antecedents and Consequences," In *Handbook of Service Science,* pp. 77-103.
Papenmeier, A., Englebienne, G., and Seifert, C. 2019. "How Model Accuracy and Explanation Fidelity Influence User Trust," *arXiv preprint arXiv:*1907.12652.
Parasuraman, R., and Riley, V. 1997. "Humans and Automation: Use, Misuse, Disuse, Abuse," *Human Factors* (39:2), pp. 230-253.
Pavlin, G., de Villiers, J. P., Ziegler, J., Jousselme, A.L., Costa, P., Laskey, K., and Jansen, L. 2021. "Relations Between Explainability, Evaluation and Trust in AI-Based Information Fusion Systems," *In 24th International Conference on Information Fusion*, Sun City, South Africa, Nov 1-4.
Pengnate, S.F., and Sarathy, R. 2017. "An Experimental Investigation of the Influence of Website Emotional Design Features on Trust in Unfamiliar Online Vendors," *Computers in Human Behavior* (67), pp. 49-60.
Pieters, W. 2011. "Explanation and Trust: What to Tell the User in Security and AI?" *Ethics and Information Technology* (13:1), pp. 53-64.
Raj, M., and Seamans, R. 2019. "Primer on Artificial Intelligence and Robotics," *Journal of Organization Design* (8:1), pp. 1-14.
Renner, M., Lins, S., Söllner, M., Thiebes, S., and Sunyaev, A. 2021. "Achieving Trustworthy Artificial Intelligence: Multi-Source Trust Transfer in Artificial Intelligence-capable Technology," In *42nd International Conference on Information Systems*, Austin, TX.
Robinette, P., Howard, A.M., and Wagner, A.R. 2017. "Effect of Robot Performance on Human–Robot Trust in Time-Critical Situations," *IEEE Transactions on Human-Machine Systems* (47:4), pp. 425-436.
Rossi, F. 2018. "Building Trust in Artificial Intelligence," *Journal of International Affairs* (72:1), pp. 127-134.
Rotter, J.B. 1980. "Interpersonal Trust, Trustworthiness, and Gullibility," *American Psychologist* (35:1).
Rousseau, D.M., Sitkin, S.B., Burt, R.S., and Camerer, C. 1998. "Not So Different After All: A Cross-Discipline View of Trust," *Academy of Management Review* (23:3), pp. 393-404.
Russell Stuart, J., and Norvig, P. 2009. *Artificial Intelligence: A Modern Approach*, Pearson.
Salem, M., Lakatos, G., Amirabdollahian, F., and Dautenhahn, K. 2015. "Would You Trust a (Faulty) Robot? Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust," In *10th ACM/IEEE International Conference on Human-Robot Interaction*, Online Conference, March 2-5.

Schmidt, P., Biessmann, F., and Teubner, T. 2020. "Transparency and Trust in Artificial Intelligence Systems," *Journal of Decision Systems* (29:4), pp. 260-278.

Siau, K., and Shen, Z. 2003. "Building Customer Trust in Mobile Commerce," *Communications of the ACM* (46:4), pp. 91-94.

Siau, K., Sheng, H., Nah, F., and Davis, S. 2004. "A Qualitative Investigation on Consumer Trust in Mobile Commerce," *International Journal of Electronic Business* (2:3), 283-300.

Smuha, N. 2019. "Ethics Guidelines for Trustworthy AI," in *AI and Ethics*, Brussels, Belgium.

Sperrle, F., El-Assady, M., Guo, G., Chau, D.H., Endert, A., and Keim, D. 2020. "Should We Trust (X) AI? Design Dimensions for Structured Experimental Evaluations," *arXiv preprint arXiv:2009.06433*.

Stormont, D.P. 2008. "Analyzing Human Trust of Autonomous Systems in Hazardous Environments," In *Proceedings of the AAAI Human Implications of Human-Robot Interaction Workshop*, pp. 27-32.

Vereschak, O., Bailly, G., and Caramiaux, B. 2021. "On The Way to Improving Experimental Protocols to Evaluate Users' Trust in AI-Assisted Clinical Decision Making," In *CHI'21 Workshop: Realizing AI in Healthcare: Challenges Appearing in the Wild*.

Vickers, N.J. 2017. "Animal Communication: When I'm Calling You, Will You Answer Too?" *Current Biology* (27:14), pp. R713-R715.

Wang, W., and Siau, K. 2018. "Living with Artificial Intelligence–Developing a Theory on Trust in Health Chatbots," In *Proceedings of the 16th Annual Pre-ICIS Workshop on HCI Research in MIS*. San Francisco, CA, Dec 13.

Wang, Y., and Singh, M.P. 2007. "Formal Trust Model for Multiagent Systems," In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*. Hyderabad, India. Jan 6-12, pp. 1551-1556.

Wanner, J., Heinrich, K., Janiesch, C., and Zschech, P. 2020. "How Much AI Do You Require? Decision Factors for Adopting AI Technology," In *Proceedings of 41st International Conference on Information Systems*, Hyderabad, India, Dec 13-16.

Waytz, A., Heafner, J., and Epley, N. 2014. "The Mind in The Machine: Anthropomorphism Increases Trust in an Autonomous Vehicle," *Journal of Experimental Social Psychology* (52), pp. 113-117.

Winfield, A.F., and Jirotka, M. 2018. "Ethical Governance is Essential to Building Trust in Robotics and Artificial Intelligence Systems," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* (376:2133), 20180085.

Wirtz, J., Patterson, P.G., Kunz, W.H., Gruber, T., Lu, V.N., Paluch, S., and Martins, A. 2018. "Brave New World: Service Robots in the Frontline," *Journal of Service Management* (29:5).

Xu, M., David, J.M., and Kim, S.H. 2018. "The Fourth Industrial Revolution: Opportunities and Challenges," *International Journal of Financial Research* (9:2), pp. 90-95.

Yin, M., Wortman V.J., and Wallach, H. 2019. "Understanding The Effect of Accuracy on Trust in Machine Learning Models," *In Proceedings of the 2019 Conference on Human Factors in Computing Systems,* Glasgow, Scotland, May 4-9, pp. 1-12.

You, S., and Robert, L.P. 2018. "Human–Robot Similarity and Willingness to Work with A Robotic Co-Worker," In *13th ACM/IEEE International Conference on Human-Robot Interaction*, Chicago, IL, Mar 5-8, pp. 251-260.

Zand, D.E. 1972. "Trust and Managerial Problem Solving," *Administrative Science Quarterly* (17:2), pp. 229-239.

Zhang, B., and Dafoe, A. 2019. "Artificial Intelligence: American Attitudes and Trends," *Available at SSRN 3312874*.

Zhang, S., Yao, L., Sun, A., and Tay, Y. 2019. "Deep Learning Based Recommender System: A Survey and New Perspectives," *ACM Computing Surveys* (52:1), pp. 1-38.

Zhang, T., Tao, D., Qu, X., Zhang, X., Lin, R., and Zhang, W. 2019. "The Roles of Initial Trust and Perceived Risk in Public's Acceptance of Automated Vehicles," *Transportation Research Part C: Emerging Technologies* (98), pp. 207–220.

Zhang, Y., Liao, Q.V., and Bellamy, R.K. 2020. "Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making," In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Barcelona Spain, Jan 27-30, pp. 295-305.

Złotowski, J., Sumioka, H., Nishio, S., Glas, D.F., Bartneck, C., and Ishiguro, H. 2016. "Appearance of A Robot Affects the Impact of Its Behaviour on Perceived Trustworthiness and Empathy," *Journal of Behavioral Robotics* (7:1), pp. 55–66.