Perceived Factors of Trustworthiness in Generative Artificial Intelligence (GenAI): Towards an Understanding of how to Assess and Build Trustworthiness

Max Evans[1], Godwin Akrong, [1] and Anthony Wensley[2]
[1]McGill University, Montreal, Canada
[2]University of Toronto, Canada

max.evans@mcgill.ca
godwin.akrong@mail.mcgill.ca
anthony.wensley@utoronto.ca

Abstract: Given the benefits and risks associated with GenAI adoption in organizations, many academics and practitioners have stressed the importance of understanding how humans come to trust these technologies and the information and knowledge (e.g., solutions/decisions) they produce. The objective of this paper is to further examine human trust in AI technologies through the lens of a widely accepted organizational trust theory and model developed by Mayer, Davis, and Schoorman. More specifically, this paper focuses on developing a better understanding of perceived factors of GenAI trustworthiness since assessing trustworthiness is a critical determinant of trust. Building on the existing theory and model, it is proposed that an individual's perception of one or more of the following dimensions of trustworthiness - ability, integrity, and benevolence - will determine how trustworthy they find GenAI to be. Ability (or competence) refers to the trustee's specific skills, knowledge, and expertise required in a specific domain. Integrity reflects the trustee's sound values or principles (e.g., fairness, consistency, justice). Benevolence is an altruistic loyalty that reflects the trustee's concern for the welfare, needs, desires, and interests of the individual over organizational or profit motives. Many researchers have proposed assessments related to GenAI ability, but integrity and benevolence are more difficult to assess, as technologies do not intrinsically embody human values or altruistic behaviors. Consequently, other parties within the organizations, such as AI designers and developers, strategic decision-makers, or the organization may be conflated into perceptions of these dimensions. The paper continues by briefly discussing how emotions and organizational culture may influence individuals' perceptions of trustworthiness and concludes by suggesting potential directions and strategies for building and representing each dimension of perceived trustworthiness in the context of GenAI.

Keywords: Trust, Generative artificial intelligence (GenAI), Trustworthiness, Model oftTrust, Organizational

1. Artificial Intelligence (AI) Technology Adoption

The adoption of Artificial Intelligence (AI) technologies in organizations has drastically increased over the last couple of years (Singla et al., 2024). Furthermore, McKinsey and Co. (Singla et al., 2025) recently found that AI adoption is happening across all sectors and organizational functions, however, each organization applies "the technology where it can generate the most value—for example, service operations for media and telecommunication companies, software engineering for technology companies, and knowledge management for professional-services organizations" (p.5). AI technologies can be understood as a "generation of technologies capable of interacting with the environment by (a) gathering information from outside (including from natural language) or from other computer systems; (b) interpreting this information, recognizing patterns, inducing rules, or predicting events; (c) generating results, answering questions, or giving instructions to other systems; and (d) evaluating the results of their actions and improving their decision systems to achieve specific objectives" (Glikson and Woolley, 2020, p. 628).

Given the vast and ever-growing volume of data and information available to organizations today, it is no surprise that organizations are exploring ways to leverage AI, and more specifically, Generative AI (GenAI), for strategic planning (Leoni et al., 2024), decision-making (Marvi et al., 2024), operational efficiencies (e.g., process or task improvements and automations) (Brynjolfsson et al., 2018; Ferrario et al., 2020; Arias-Pérez and Vélez-Jaramillo, 2022; Giovine et al., 2024), or innovation (i.e., creative tasks) (Giovine et al., 2024; Zhang et al., 2025). These implementations will not likely slow down, as organizations are seeing immediate benefits through increased revenues and reductions in costs (Singla et al., 2024).

AI also presents considerable risks for organizations and their employees. The most publicized risk relates to job loss and labour displacements from the automation of various organizational tasks and processes (Briggs and Kodnani, 2023; Zhang et al., 2025) but other known risks include algorithmic bias, misinformation, inaccuracy, hallucinations, technological dependence, as well as privacy and cybersecurity concerns (Singla et al., 2024; Zhang et al., 2025). Most importantly, AI lacks explainability (i.e., it is non-deterministic and non-transparent), making its behavior difficult (or impossible) for individuals to understand or predict (Giovine et al., 2024; Zhang

et al., 2025). Many of these risks open the possibility of using inaccurate or biased information in the decisions being made.

Considering the benefits and risks, many academics and practitioners have stressed the importance of understanding how humans come to trust these technologies and the information and knowledge (e.g., solutions/decisions) they produce (Ferrario et al., 2020; Glikson and Woolley, 2020). The objective of this paper is to further examine how AI trustworthiness can be assessed through the lens of a widely accepted organizational trust theory and model developed by Mayer, Davis, and Schoorman.

2. A Theoretical and Empirical Model of Trust

In the academic literature, trust is a widely defined and operationalized concept, with hundreds of different psychometric measures. However, there are only a handful of characteristics of trust (and associated models) that are based on well-grounded theory and have been empirically tested, especially in an organizational setting (McEvily and Tortoriello, 2011).

Theorists generally agree that trust is dependent on a trustor's (i.e., the person doing the trusting) expectation of how the trustee (i.e., person/actor being trusted) will behave, or what the outcomes of the act of trusting (e.g., interaction) would be. When a trustee fulfills the trustor's expectations, this acts as a mechanism for creating and increasing trust. In many cases, especially those where no prior interactions exist, the trustor must be willing to take a risk (or make themselves vulnerable) to engage in a trusting behavior (Mayer et al., 1995). Of course, no risk-taking behavior is needed for there to be a willingness to be vulnerable (or trust), as all the risk is "inherent in the behavioral manifestation" or engaging in the trusting behavior and "assuming the risk" or making themselves vulnerable (Mayer et al., 1995, p. 724). Perceived risks moderate an individual's willingness to be vulnerable and their risk-taking behavior (Mayer et al., 1995). The level of trust needed to engage in a trusting behavior will directly relate to the level of perceived risk or vulnerability involved in the situation (relative to the known benefits). The higher the perceived risk, the more trust is needed. Thus, context is a particularly important factor to consider because trust is largely situational and should always be qualified with a consideration of what the trustee is being trusted to do (Mayer et al., 1995; O'Neill, 2002; 2020). Trust typically develops over time through interactions between the trustor and the trustee but often changes based on a particular situation or context. For example, one may trust their nanny to care for their kids more than to fix their computer.

A trustor needs to make themselves vulnerable (through their risk-taking behavior) to gather data about the trustee to see if they can be trusted in a particular situation. The trustee's response (or the outcomes of the trusting behavior) is then used to assess their trustworthiness. Interactions (i.e., outcomes) where expectations of the trustee's actions are fulfilled lead to a greater perception of their trustworthiness. If a trustor does not have enough interactions to accurately assess trustworthiness, they exhibit a "trust similar to blind faith" (Mayer et al., 1995, p. 730) or gauge trustworthiness through other individuals' (or the organization's) accounts of the trustee.

Numerous factors or dimensions have been proposed to explain trustworthiness. However, the most widely adopted factors are those identified by Mayer et al. (1995; 2007). According to the authors, a trustor's overall perception of the trustworthiness of the trustee will depend on their assessment of the trustee's ability, integrity, and/or benevolence. Although a high amount of all three dimensions will produce a high perception of trustworthiness, all three are not needed for the trustor to engage in a specific trusting (i.e., risk-taking) behavior. Moreover, each dimension contributes uniquely to the perception of trustworthiness, and one dimension may be more pertinent or meaningful in a particular trusting situation than the others (e.g., Chen et al., 2011; Evans, 2012; 2013). For this reason, Mayer et al. (1995) propose that trustworthiness should not be viewed as binary, but rather as a continuum where the relative importance and strength of each dimension may change over time, especially as the number of interactions increases.

Trust is further complicated by the proposition that every trustor has relatively stable internal characteristics that influence the amount of trust they have for any given trustee. Grounded in personality or social construction (culture, past experiences), Mayer et al. (1995) argue that individuals have a general willingness (or propensity) to trust others, especially prior to interacting with them, which transfers from one trusting situation to the next. In the context of AI, this trait could be understood as the overall willingness of the trustor to trust in AI, in the absence of a specific situation, context, or AI system (Hoff and Bashir, 2015; Schlicker et al., 2024).

Mayer et al. (1995) define trust 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" (p. 712). In addition, the authors (Mayer et al., 1995; Schoorman et al., 2007) present a multidisciplinary Integrative Model of Trust [MoT], which further explains the relationship between all the factors discussed above (see Figure 1).

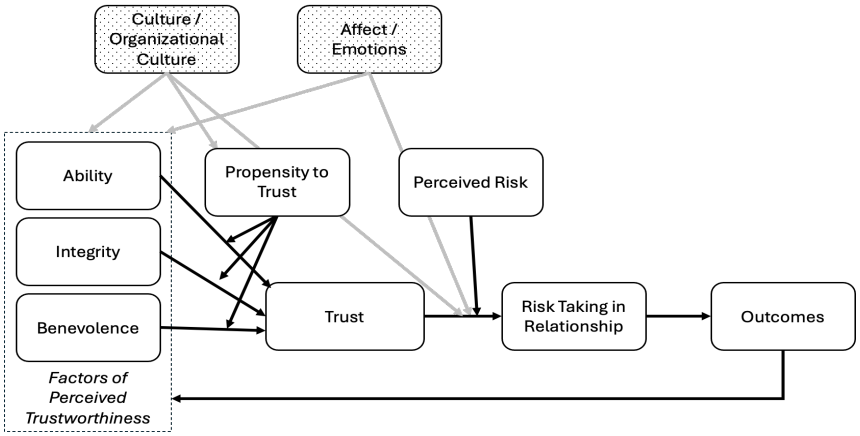

Figure 1: Mayer et al.'s MoT (Adapted to include emotions and culture)

This definition of trust, and the MoT, are among the most widely adopted and investigated conceptualizations of trust. Moreover, many MoT factors have been successfully operationalized and validated across numerous settings and disciplines (Schoorman et al., 2007; McEvily and Tortoriello, 2011). The MoT is also appropriate for studies looking at human trust in AI technologies because it is designed for organizational settings, applicable across different levels of analysis, and designed for unidirectional trust analysis (Mayer et al., 1995; Schoorman et al., 2007).

3. Trust in Humans vs. Conversational GenAI

There are some important differences in discussing how humans place trust in other humans and GenAI technologies. The first distinction is that the perceived and actual risk associated with trusting AI-produced information is different than the perceived or actual risk associated with acting on AI-generated information, without further challenging or verifying it. The former involves negligible risk since the AI is being used simply as a decision-aid. In fact, it may even reduce perceived risk by giving users the ability to privately ask questions that may be perceived as ignorant or risky if asked of a work colleague. On the other hand, taking the AI solution/information, unchallenged, and acting on it (or allowing the AI to autonomously act) bears a higher risk since there are now perceived and actual risks to the organization or its operations. As AI autonomy increases, or it is expected to complete more important tasks, make riskier decisions, or implement complex solutions, perceived (and actual) risk increases, requiring more trustworthiness in the AI.

Some researchers argue that humans are naturally skeptical and risk-averse when it comes to relying on new technologies in general (Omrani et al., 2022) and AI systems specifically (Schmidt et al., 2020), especially when using them for important decision-making tasks (Solberg et al., 2022). Textor et al. (2022) suggest such concerns may be due to AI's inability to understand and act on ethical principles, seeing ethical decision-making as a uniquely human skill. Other noted concerns include fairness, accountability, transparency, misuse, and privacy (Omrani et al., 2022; Marwala, 2024). Research also suggests that trustworthiness may be influenced by third-party evaluations, culture (Mayer et al., 1995; Schoorman et al., 2007), and past experiences with AI (McKnight et al., 2017; Omrani et al., 2022; Choung et al., 2023a). Furthermore, humans are prone to cognitive biases (Solberg et al., 2022; Chi et al., 2023; Riley and Dixon, 2024) and affect (moods and emotions), which may distort assessments of risk and trustworthiness (Schoorman et al., 2007; Schmidt et al., 2020; Albahri et al., 2023). These same distortions have been argued to exist with trust in AI (e.g., Evans and Wensley, In press), with some researchers showing empirical evidence for it (Gillath et al., 2021; Chen et al., 2011; Frazier et al., 2016; Solberg et al., 2022; Choung et al., 2023a).

Conversational AI is also unique in how users interact with it. For instance, the iterative nature of conversations allows users to receive numerous complex answers or solutions in a brief period of interacting with the AI. This type of information and knowledge exchange is rare in human-to-human interactions, as humans need time to process solutions, only have expertise in a limited set of domains, or may not have access to relevant information. Each solution or answer the AI provides, like information/knowledge exchanges with humans, acts as a proxy for helping to determine trustworthiness in a particular domain. For AI, trustworthiness may be

gauged in a shorter timeframe because interactions are more efficient (i.e., individuals can have a greater number of information and knowledge exchanges than with humans). However, conversational AI may also pose a problem for gauging trustworthiness because of how AI solutions are 'negotiated' to fit the user's prompts. GenAI models use Natural Language Processing to analyze prompts, which are designed to infer and respond to the user's conveyed tone, as well as their wording or phraseology. Inadvertently, users may introduce biases into their prompts that alter the GenAI solution in a way that it better aligns with the user's expectations or affects information quality. From a trust perspective, the GenAI is crafting a solution (outcome) that the user will inherently perceive as more or less trustworthy based on how prompts are designed, or conversations are negotiated.

In a similar vein, it is important to note that the range of possible generated solutions fundamentally hinges on the data set used for training the GenAI (generally referred to as the corpus), as well as the statistical models and algorithms employed. The latter defines how the GenAI interprets and learns from the data, ultimately limiting or expanding the range of possible solutions. Users have no practical method to evaluate the corpus, the statistical models, or the algorithms used in decision-making, and even if they were able to access them, they would have no way to understand the nondeterministic logic used in generating any given solution. From a trust perspective, one could argue there is inherently higher risk in situations requiring the need for higher trustworthiness in lieu of sheer 'blind trust.'

4. Trustworthiness Dimensions and GenAI

Although every factor in the MoT is critically important for there to be an accurate assessment of trust, trust researchers (e.g., McKnight et al., 2000; Cho et al., 2015; Yilmaz and Liu, 2022) have stressed the specific importance of the trustworthiness factor. One of the most commonsensical arguments supporting this proposition is that individuals should want to put more trust in the trustworthy and less in the untrustworthy (O'Neill, 2002; 2020).

As previously mentioned, the MoT proposes (and operationalizes) three dimensions, which as a set "explain a major portion of trustworthiness while maintaining parsimony" (Mayer et al., 1995, p. 722): Ability, Integrity, and Benevolence. Before elaborating on how each dimension applies in the context of GenAI, it is important to note that trustworthiness is situational (Mayer et al., 1995; Hoff and Bashir, 2015; Schlicker et al., 2024), and assessments change across tasks, domains, and even over time. It is also worth acknowledging that, unlike humans, AI technologies do not inherently possess benevolence, integrity, or any other values, making these dimensions difficult to directly observe or attribute to AI. Perceptions of these dimensions likely reflect some strategic organizational directive or organizational culture/value, which is 'programmed' into the decision-making of the technology. This may cause users to conflate these trustworthy dimensions between the AI and the strategic planners who instituted the organizational cultures and directives, or the designers and developers of the technology. Moreover, designing and developing AI technologies to autonomously and programmatically follow a set of values (or act benevolently) may bias or modify the information/solution, influencing other trustworthy dimensions like ability (e.g., see Evans and Wensley, In Press).

The next section explores each of the three dimensions of perceived trustworthiness. Each dimension is defined and explained in a human-human trust context and then in a human-AI trust context. It is important to note that although numerous studies exist exploring trust in AI, the AI or AI-related research papers cited below directly acknowledge or build on Mayer et al.'s MoT model and their Perceived Factors of Trustworthiness. This is meant to confirm the constructs are understood and applied uniformly, which is critical given the large number of trust measures and models available (McEvily and Tortoriello, 2011).

4.1 Ability/Competence

Ability-based trustworthiness refers to a "group of skills, competencies, and characteristics that enable a party to have influence within some specific domain" (Mayer et al., 1995, p. 717). It implies possession of domain-specific information or knowledge that has the potential to increase organizational performance, increase operational efficiency, or lead to greater innovation.

In humans, ability or competence is traditionally specific to a single, fixed domain (Zand, 1972), which supports why individuals may warrant higher trust in one domain than another (Schoorman et al., 2007). Since disparate knowledge may be required across different domains, ability-based trustworthiness in one domain does not guarantee it in another. Hoff and Bashir (2015) argue that a similar domain-specific trust will exist with AI technologies, suggesting researchers measure trustworthiness relative to the characteristics of the task being

performed by the AI. The context of the information or knowledge being requested is significant since GenAI will perform better in some contexts than others, based on the corpus and machine learning algorithms.

Characteristics often connected with ability include professionalism, dedication, preparedness, and carefulness. In the context of decision- or task-aiding GenAI, ability-based trustworthiness will largely be assessed based on how "helpful, competent, or useful" (Glikson and Woolley, 2020, p. 631) the information (or solutions) the system provides are. Lewis and Marsh (2022) suggest that evidence of competence comes from past performance and/or an understanding of the AI's mechanisms and capabilities.

Several researchers have built on, or extended, Mayer et al.'s (1995) ability-based trustworthiness dimension to apply it in an AI or AI-related context (e.g., automated-systems, machine learning technologies, multi-agent systems, automation, and robotics). From a theoretical perspective, this dimension is commonly adopted verbatim, referring to the technologies' ability (e.g., Kessler et al., 2017; Toreini et al., 2020; Schlicker et al., 2021) or competence (e.g., Lewis and Marsh, 2022; Matsuura and Shirasaka, 2024). These terms can be considered as being synonymous, both referring specifically to the technology's skills and domain knowledge. Other theorists suggest latent computer science-related dimensions, which relate closely to the construct of ability, like explainability (Giovine et al., 2024), accuracy (Ferrario et al., 2020), transparency (Hoff and Bashir, 2015; Glikson and Woolley, 2020), performance, robustness, latency, or accuracy (Schlicker et al., 2021; 2025).

Ability (e.g., Langer et al., 2023; Agnihotri and Bhattacharya, 2024) and competence-based trustworthiness (e.g., Gulati et al., 2018) have also been explored empirically, in the context of AI and AI-related technologies. These factors largely measure whether the technology performs tasks effectively and efficiently (i.e., meeting user expectations), considers important information, possesses technical know-how, and provides accurate results. Other empirical studies, building on the MoT, employ latent subdimension variables to represent ability- or competence-based trustworthiness, like performance (Park, 2020; Chita-Tegmark et al., 2021; Alarcon et al., 2024a), functionality (Choung et al., 2023b; Kleizen et al., 2023), capability and reliability (Chita-Tegmark et al., 2021). Most of these constructs accurately capture aspects of ability and competence; however, some also introduce the notion of consistency over time, which, as will be discussed below, is likely an aspect of integrity-based trustworthiness, not ability. It is important not to conflate these constructs. Moreover, Singh and Mayer (2024) warn that although a greater perception of ability may increase trustworthiness in the AI, if the AI is perceived as malevolent, higher ability could also lead to decreased trust. This highlights the importance for research studies to explore all the dimensions of trustworthiness and the interactions between them to better understand their impact on AI trustworthiness.

4.2 Integrity

Integrity-based trustworthiness reflects a perception that the "trustee adheres to a set of principles that the trustor finds acceptable" or that are "sound" (Mayer et al., 1995, p. 719). These principles can vary slightly; however, common themes include acting fairly, consistently, and with a strong "sense of justice" (p. 719). As with ability, several researchers have built on or extended Mayer et al.'s (1995) integrity-based trustworthiness dimension to apply it in an AI or AI-related context. From a theoretical perspective, this dimension is commonly adopted directly, referring to the technologies' integrity (Kessler et al., 2017; Toreini et al., 2020; Schlicker et al., 2021; Lewis and Marsh, 2022). Most of these authors propose definitions that generally suggest the trustee's adherence to a set of principles acceptable to the trustor. Schlicker et al. (2021) provide some further guidance on this measure in AI by associating it with related factors like transparency, adherence to ethical guidelines, and reliability. Several other theorists follow suit with factors that relate closely to the construct of integrity, like honesty (i.e., making and honoring agreements, avoiding deception, and adhering to promises or expectations) (Lewis and Marsh, 2022; Mehrotra et al., 2024), reliability (i.e., the technology exhibits behavior that is expected by the user, and that it acts consistently over time) (Hoff and Bashir, 2015; Glikson and Woolley, 2020; Matsuura and Shirasaka, 2024), predictability (i.e., the technology performs actions consistently and expectedly) (Toreini et al., 2020; Lewis and Marsh, 2022), transparency (i.e., the technology informs the user about the decision-making logic or operating rules) (Glikson and Woolley, 2020; Hoff and Bashir, 2015; Shin, 2021; Mehrotra et al., 2024), and fairness (Mehrotra et al., 2024).

Integrity-based trustworthiness has also been explored empirically in the context of AI and AI-related technologies (e.g., Gulati et al., 2018; Choung et al., 2023b; Langer et al., 2023; Agnihotri and Bhattacharya, 2024). In these studies, integrity is generally used to describe technologies that provide honest, objective, consistent, unbiased, or reliable recommendations and those that adhere to ethical standards valued by the trustor. Other closely related factors include reliability (Chita-Tegmark et al., 2021), predictability (Gulati et al., 2018), accountability (Shin, 2021), fairness (Shin, 2021), and honesty (Gulati et al., 2018).

4.3 Benevolence-Based

Benevolence-based trustworthiness reflects a perception that the trustee has the trustor's welfare (e.g., needs, desires, and interests) in mind and intends on doing "good to the trustor" (Mayer et al., 1995, p. 718), "aside from an egocentric profit motive" or extrinsic reward (Schoorman et al., 2007, p. 345). Essentially, this is a perception that the trustee will not harm the trustor, knowingly, and will even go out of their way to help. Low benevolence-based trust would be assessed in cases where "information [is used] in a way that helps the company most, even at the possible expense of the employee" (Mayer et al., 1995, p. 721). Common characteristics associated with benevolence include altruism and loyalty.

Fewer researchers have proposed, adapted, or extended Mayer et al.'s (1995) benevolence-based trustworthiness dimension to an AI or AI-related context. From a theoretical perspective, this dimension is largely adopted directly, referring to the technologies' benevolence (Evans and Wensley, In press; Kessler et al., 2017; Lewis and Marsh, 2022; Schlicker et al., 2021; Toreini et al., 2020). Most of these authors propose definitions that generally suggest a belief that the technology will do good to the trustor, or act in the trustor's best interests. Benevolence has also been explored empirically, in the context of AI and AI-related technologies (Agnihotri and Bhattacharya, 2024; Choung et al., 2023b; Gulati et al., 2018; Langer et al., 2023). Operationalization in these studies reflects a similar definition. Other studies measure closely related factors like goodwill (Matsuura and Shirasaka, 2024), which is described as the trustee's intention to act in the trustor's best interest or purpose (Alarcon et al., 2024a; Alarcon et al., 2024b; Park, 2020), which is described as an alignment between the intended application of the system and the user's needs.

5. Building Trustworthiness in GenAI

Although research has been conducted identifying individual factors that influence trustworthiness dimensions, most of these findings offer limited practical application or guidance for GenAI use in organizations. This paper will, instead, conclude by briefly speaking to three potential avenues for building Gen-AI trustworthiness: subject matter expertise validation, explainable AI (XAI), and prompt engineering education.

Subject matter expertise validation builds perceived trustworthiness in AI by involving credible experts or trusted agencies to assess and endorse AI systems (Afroogh et al., 2024). This process can ensure transparency through comprehensive documentation, which outlines the system's lineage, safety, performance, and ethical considerations. Expert evaluations also reassure users about the system's accuracy, reliability, and ethical compliance, fostering trustworthiness without requiring technical expertise. In healthcare, subject matter experts like clinicians and biomedical scientists enhance trustworthiness by ensuring dataset quality, refining diagnostic criteria, interpreting outputs in real-world contexts, and advocating for AI adoption (Winter and Carusi, 2022). Their collaborative involvement throughout development integrates social, epistemic, and technical dimensions, making AI outputs relatable, actionable, and socially acceptable. This approach highlights that trustworthiness in AI is not just a technical achievement but a sociotechnical construct shaped by human expertise and interactions.

Explainable AI (XAI) is also a promising approach for building perceived trustworthiness in AI systems by making their decision-making processes transparent and understandable to users (Chamola et al., 2023). XAI addresses the black-box nature of AI models, which lack clarity in their operations, leading to concerns about correctness, bias, reliability, and accountability. By providing clear explanations of how AI systems arrive at their conclusions, XAI fosters trustworthiness and confidence among users, stakeholders, and regulators. Transparency and interpretability are key aspects of XAI, enabling users to comprehend the rationale behind AI decisions, particularly in high-stakes sectors like healthcare and finance. Additionally, XAI supports accountability by ensuring AI systems can be held responsible for their actions and aids regulatory compliance by adhering to legal frameworks like GDPR. XAI also enhances the reliability of AI-user interactions, fostering trust in the collaborative process between AI and its users (Ferrario and Loi, 2022). XAI research has already examined factors closely related to trustworthiness dimensions (e.g., accountability, fairness, transparency) and has found them to have a positive influence on trust (Shin, 2021). Recently, XAI researchers (Mehrotra et al., 2024) have even begun testing integrity-based AI explanations based on schemas for fairness, transparency, and honesty with similar positive results.

Finally, prompt engineering education is identified as a key method for building perceived trustworthiness in AI systems (Kraljic and Lahav, 2024). By teaching users how to craft well-structured prompts, it enhances the quality of AI-generated outputs, making interactions more reliable and goal-directed (Knoth et al., 2024). This education helps users understand AI's capabilities and limitations, reducing misconceptions and fostering confidence in its

outputs. It also enables users to critically evaluate results, address issues like hallucinations, and delegate tasks appropriately, which are essential for trust-building. Additionally, prompt engineering complements AI literacy by empowering users with knowledge about AI processes, roles, and limitations, allowing them to make informed decisions and maintain agency in AI-assisted tasks. Together, these skills create a foundation for responsible and confident use of AI in professional contexts. Still, critics suggest that prompt engineering may limit accessibility and currently does not align with natural human language use, advocating for collaborative AI interfaces that evolve intents with users to share effort (Kraljic and Lahav, 2024). Despite this, prompt engineering remains a valuable tool for fostering trustworthiness by empowering users to better understand and control AI outputs.

Ethics declaration: Ethical clearance was not required for this research.

Artificial intelligence (AI) declaration: No AI tools were used in the idea or content generation, writing, or editing of this paper.

References

Agnihotri, A., and Bhattacharya, S. (2024). Chatbots' effectiveness in service recovery. *International Journal of Information Management*, *76*, p. 102679. https://doi.org/10.1016/j.ijinfomgt.2023.102679

Alarcon, G. M., Lyons, J. B., Mayer, R. C., Barelka, A. J., and Bobko, P. (2024a). Why does one trust? A 360-degree perspective on the role of position power in weighting trustworthiness factors. *Military Psychology*, pp. 1-14. https://doi.org/10.1080/08995605.2024.2373576

Alarcon, G. M., Capiola, A., Lee, M. A., Willis, S., Hamdan, I. A., Jessup, S. A., and Harris, K. N. (2024b). Development and Validation of the system trustworthiness scale. *Human Factors*, *66*(7), pp. 1893-1913. https://doi.org/10.1177/00187208231189000

Albahri, A.S., Duhaim, A.M., Fadhel, M.A., Alnoor, A., Baqer, N.S., Alzubaidi, L., Albahri, O.S., Alamoodi, A.H., Bai, J., Salhi, A. and Santamaría, J., 2023. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion*, *96*, pp.156-191.

Afroogh, S., Akbari, A., Malone, E., Kargar, M., and Alambeigi, H. (2024). Trust in AI: progress, challenges, and future directions. *Humanities and Social Sciences Communications*, *11*(1), pp. 1-30. https://doi.org/10.1057/s41599-024-04044-8

Arias-Pérez, J., and Vélez-Jaramillo, J. (2022). Understanding knowledge hiding under technological turbulence caused by artificial intelligence and robotics. *Journal of Knowledge Management*, *26*(6), pp. 1476-1491. https://doi.org/10.1108/jkm-01-2021-0058

Brynjolfsson, E., Mitchell, T., and Rock, D. (2018). What can machines learn, and what does it mean for occupations and the economy? *AEA Papers and Proceedings, 108*, pp. 43–47. https://doi.org/10.1257/pandp.20181019

Briggs, J., and Kodnani, D. (2023). The potentially large effects of artificial intelligence on economic growth. *Global Economics Analyst*. Goldman Sachs Economics Research. https://www.gspublishing.com/content/research/en/reports/2023/03/27/

Chamola, V., Hassija, V., Sulthana, A. R., Ghosh, D., Dhingra, D., and Sikdar, B. (2023). A review of trustworthy and explainable artificial intelligence (xai). *IEEE Access*, *11*, pp. 78994-79015. https://doi.org/10.1109/access.2023.3294569

Chen, C. C., Saparito, P., and Belkin, L. (2011). Responding to trust breaches: The domain specificity of trust and the role of affect. *Journal of Trust Research, 1*(1), pp. 85-106. https://doi.org/10.1080/21515581.2011.552438

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

Chita-Tegmark, M., Law, T., Rabb, N., and Scheutz, M. (2021, March). Can you trust your trust measure?. In *Proceedings of the 2021 ACM/IEEE international conference on human-robot interaction*. (pp. 92-100). https://doi.org/10.1145/3434073.3444677

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

Choung, H., David, P., and Ross, A. (2023b). Trust and ethics in AI. *AI and Society*, *38*(2), pp. 733-745. https://doi.org/10.1007/s00146-022-01473-4

Cho, J. H., Chan, K., and Adali, S. (2015). A survey on trust modeling. *ACM Computing Surveys (CSUR)*, *48*(2), pp. 1-40.

Evans, M. M. (2012). *Knowledge sharing: An empirical study of the role of trust and other social-cognitive factors in an organizational setting* (Doctoral dissertation, University of Toronto).

Evans, M. M. (2013). Is trust the most important factor influencing knowledge sharing in organisations? *Journal of Information and Knowledge Management, 12*(4), p.1350038. https://doi.org/10.1142/s021964921350038x

Ferrario, A., Loi, M. and Viganò, E. (2020). In AI we trust incrementally: A multi-layer model of trust to analyze human-artificial intelligence interactions. *Philosophy & Technology*, *33*(3), pp.523-539. https://doi.org/10.1007/s13347-019-00378-3

Ferrario, A., and Loi, M. (2022, June). How explainability contributes to trust in AI. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency* (pp. 1457-1466). https://doi.org/10.1145/3531146.3533202

Frazier, M. L., Tupper, C., and Fainshmidt, S. (2016). The path (s) to employee trust in direct supervisor in nascent and established relationships: A fuzzy set analysis. *Journal of Organizational Behavior*, *37*(7), pp. 1023-1043. https://doi.org/10.1002/job.2091

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

Giovine, C., Bankhwal, M., Pometti, M., and Roberts, R. (2024, November 26). Building AI trust: The key role of explainability. *McKinsey and Company*. https://www.mckinsey.com/capabilities/quantumblack/our-insights/building-ai-trust-the-key-role-of-explainability

Gulati, S., Sousa, S., and Lamas, D. (2018, December). Modelling trust in human-like technologies. In *Proceedings of the 9th Indian conference on human-computer interaction* (pp. 1-10). https://doi.org/10.1145/3297121.3297124

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. https://doi.org/10.5465/annals.2018.0057

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. https://doi.org/10.1177/0018720814547570

Johnson, A. (2023, March 30). Which jobs will AI replace? These 4 industries will be heavily impacted. *Forbes*. https://www.forbes.com/sites/ariannajohnson/2023/03/30/which-jobs-will-ai-replace-these-4-industries-will-be-heavily-impacted/

Kessler, T. T., Larios, C., Walker, T., Yerdon, V., and Hancock, P. A. (2017). A comparison of trust measures in human–robot interaction scenarios. In *Advances in Human Factors in Robots and Unmanned Systems: Proceedings of the AHFE 2016 International Conference on Human Factors in Robots and Unmanned Systems, July 27-31, 2016, Walt Disney World®, Florida, USA* (pp. 353-364). Springer International Publishing. https://doi.org/10.1007/978-3-319-41959-6_29

Knoth, N., Tolzin, A., Janson, A., and Leimeister, J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence*, *6*, p. 100225. https://doi.org/10.1016/j.caeai.2024.100225

Kraljic, T., and Lahav, M. (2024). From prompt engineering to collaborating: A human-centered approach to AI interfaces. *Interactions*, *31*(3), pp. 30-35. https://doi.org/10.1145/3652622

Langer, M., König, C. J., Back, C., and Hemsing, V. (2023). Trust in Artificial Intelligence: Comparing trust processes between human and automated trustees in light of unfair bias. *Journal of Business and Psychology*, *38*(3), pp. 493-508. https://doi.org/10.31234/osf.io/r9y3t

Leoni, L., Gueli, G., Ardolino, M., Panizzon, M., and Gupta, S. (2024). AI-empowered KM processes for decision-making: empirical evidence from worldwide organisations. *Journal of Knowledge Management*, *28*(11), pp. 320-347. https://doi.org/10.1108/jkm-03-2024-0262

Levin, D. Z., and Cross, R. (2004). The strength of weak ties you can trust: The mediating role of trust in effective knowledge transfer. *Management science*, *50*(11), pp. 1477-1490. https://doi.org/10.1287/mnsc.1030.0136

Lewis, P. R., and Marsh, S. (2022). What is it like to trust a rock? A functionalist perspective on trust and trustworthiness in artificial intelligence. *Cognitive Systems Research*, *72*, pp. 33-49. https://doi.org/10.1016/j.cogsys.2021.11.001

Marvi, R., Foroudi, P., and Cuomo, M. T. (2024). Past, present and future of AI in marketing and knowledge management. *Journal of Knowledge Management*, *29*(11), pp. 1-31. https://doi.org/10.1108/jkm-07-2023-0634

Marwala, T. (2024). Opportunity Seeking Versus Risk Aversion. In *The Balancing Problem in the Governance of Artificial Intelligence* (pp. 51-67). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-97-9251-1_4

Matsuura, T., and Shirasaka, S. (2024, July). Model for Trust and Reliance on Technology: Exploring the Relationship Between Trustor, Trustee, and Context. In *2024 16th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)* (pp. 432-439). IEEE. https://doi.org/10.1109/iiai-aai63651.2024.00085

Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. *The Academy of Management Review, 20*(3), pp. 709–734. https://doi.org/10.2307/258792

Mehrotra, S., Jorge, C. C., Jonker, C. M., and Tielman, M. L. (2024). Integrity-based explanations for fostering appropriate trust in AI agents. *ACM Transactions on Interactive Intelligent Systems*, *14*(1), pp.1-36. https://doi.org/10.1145/3610578

McKnight, D. Harrison and Chervany, Norman L., "What is Trust? A Conceptual Analysis and an Interdisciplinary Model" (2000). *AMCIS 2000 Proceedings*. pp. 382. https://aisel.aisnet.org/amcis2000/382

McKnight, D. H., Lankton, N. K., Nicolaou, A., and Price, J. (2017). Distinguishing the effects of B2B information quality, system quality, and service outcome quality on trust and distrust. *The Journal of Strategic Information Systems*, *26*(2), pp. 118-141. https://doi.org/10.1016/j.jsis.2017.01.001

McEvily, B., and Tortoriello, M. (2011). Measuring trust in organisational research: Review and recommendations. *Journal of Trust Research, 1*(1), pp. 23–63. https://doi.org/10.1080/21515581.2011.552424

Omrani, N., Rivieccio, G., Fiore, U., Schiavone, F., and Agreda, S. G. (2022). To trust or not to trust? An assessment of trust in AI-based systems: Concerns, ethics and contexts. *Technological Forecasting and Social Change*, *181*, p.121763. https://doi.org/10.1016/j.techfore.2022.121763

O'Neill, O. (2002). *A question of trust*. Cambridge University Press.

O'Neill, O. (2020). Trust and accountability in a digital age. *Philosophy, 95*(1), pp. 3–17.

Park, S. (2020). Multifaceted trust in tourism service robots. *Annals of Tourism Research*, *81*, p. 102888. https://doi.org/10.1016/j.annals.2020.102888

Riley, B. K., and Dixon, A. (2024). Emotional and Cognitive Trust in Artificial Intelligence: A Framework for Identifying Research Opportunities. *Current Opinion in Psychology*, p. 101833. https://doi.org/10.1016/j.copsyc.2024.101833

Schmidt, P., Biessmann, F., and Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*, *29*(4), pp. 260-278. https://doi.org/10.1080/12460125.2020.1819094

Schlicker, N., and Langer, M. (2021). Towards warranted trust: A model on the relation between actual and perceived system trustworthiness. In *Proceedings of Mensch und Computer 2021* (pp. 325-329). https://doi.org/10.1145/3473856.3474018

Schlicker, N., Baum, K., Uhde, A., Sterz, S., Hirsch, M. C., and Langer, M. (2024). How do we assess the trustworthiness of AI? Introducing the trustworthiness assessment model (TrAM). *PsyArXiv*. https://doi.org/10.31234/osf.io/qhwvx

Schlicker, N., Baum, K., Uhde, A., Sterz, S., Hirsch, M. C., and Langer, M. (2025). How Do We Assess the Trustworthiness of AI? Introducing the Trustworthiness Assessment Model (TrAM). *Computers in Human Behavior*, p. 108671. https://doi.org/10.31234/osf.io/qhwvx

Schoorman, D. F., Mayer, R. C., and Davis, J. H. (2007). An integrative model of organizational trust: Past, present, and future. *The Academy of Management Review, 32*(2), pp. 344–354. https://doi.org/10.5465/amr.2007.24348410

Selbst, A.D. and Barocas, S., 2018. The intuitive appeal of explainable machines. *Fordham L. Rev.*, *87*, p.1085. https://doi.org/10.2139/ssrn.3126971

Solberg, E., Kaarstad, M., Eitrheim, M. H. R., Bisio, R., Reegård, K., and Bloch, M. (2022). A conceptual model of trust, perceived risk, and reliance on AI decision aids. *Group and Organization Management*, *47*(2), pp. 187-222. https://doi.org/10.1177/10596011221081238

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*, p. 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Singh, M. P., and Mayer, R. C. (2024). Toward a scorecard (and roadmap) for trustworthy AI implementation in organizations. In *A Research Agenda for Trust* (pp. 91-104). Edward Elgar Publishing. https://doi.org/10.4337/9781802200942.00015

Singla, A., Sukharevsky, A., Yee, L., Chui, M., and Hall, B. (2004, May 30). The state of AI in early 2024: Gen AI adoption spikes and starts to generate value. *McKinsey and Company*. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai#/

Singla, A., Sukharevsky, A., Yee, L., Chui, M., and Hall, B. (2025). The state of AI: How organizations are rewiring to capture value. *McKinsey and Company*, *12*.

Textor, C., Zhang, R., Lopez, J., Schelble, B. G., McNeese, N. J., Freeman, G., ... and 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), pp. 252-281. https://doi.org/10.1177/15553434221113964

Yilmaz, L., and Liu, B. (2022). Model credibility revisited: Concepts and considerations for appropriate trust. *Journal of Simulation*, *16*(3), pp. 312-325. https://doi.org/10.1080/17477778.2020.1821587

Winter, P., and Carusi, A. (2022). 'If You're Going to Trust the Machine, Then That Trust Has Got to Be Based on Something': Validation and the Co-Constitution of Trust in Developing Artificial Intelligence (AI) for the Early Diagnosis of Pulmonary Hypertension (PH). *Science and Technology Studies*, *35*(4), pp. 58-77. https://doi.org/10.23987/sts.102198

Zand, D. E. (1972). Trust and managerial problem solving. *Administrative Science Quarterly, 17*, pp. 229–239. https://doi.org/10.2307/2393957

Zhang, Q., Zuo, J., and Yang, S. (2025). Research on the impact of generative artificial intelligence (GenAI) on enterprise innovation performance: a knowledge management perspective. *Journal of Knowledge Management*. https://doi.org/10.1108/jkm-10-2024-1198.