Group & Organization Management
2022, Vol. 47(2) 187–222

DOI: 10.1177/10596011221081238
journals.sagepub.com/home/gom
SAGE

A Conceptual Model of Trust, Perceived Risk, and Reliance on AI Decision Aids

Elizabeth Solberg[*,1], Magnhild Kaarstad[1], Maren H. Rø Eitrheim[1], Rossella Bisio[2], Kine Reegård[1], and Marten Bloch[2]

Abstract

There is increasing interest in the use of artificial intelligence (AI) to improve organizational decision-making. However, research indicates that people's trust in and choice to rely on "AI decision aids" can be tenuous. In the present paper, we connect research on trust in AI with Mayer, Davis, and Schoorman's (1995) model of organizational trust to elaborate a conceptual model of trust, perceived risk, and reliance on AI decision aids at work. Drawing from the trust in technology, trust in automation, and decision support systems literatures, we redefine central concepts in Mayer et al.'s (1995) model, expand the model to include new, relevant constructs (like perceived control over an AI decision aid), and refine propositions about the relationships expected in this context. The conceptual model put forward presents a framework that can help researchers studying trust in and reliance on AI decision aids develop their research models, build systematically on each other's research, and contribute to a more cohesive understanding of the phenomenon. Our paper concludes with five next steps to take research on the topic forward.

[1]Department of Human-Centred Digitalization, Institute for Energy Technology, Halden, Norway
[2]Department of Humans and Automation, Institute for Energy Technology, Halden, Norway

Corresponding Author:
[*]Elizabeth Solberg, Department of Human-Centred Digitalization, Institute for Energy Technology, Os Allé 5, Halden 1777, Norway.
Email: elizabeth.solberg@ife.no

Keywords
trust, perceived risk, reliance, artificial intelligence, AI decision aids

Research on organizational trust is most often concerned with understanding what facilitates one person's trust in and decision to rely on another person to carry out important work tasks or other decision-making responsibilities (e.g., Costa, Fulmer, & Anderson, 2018; Davis, Schoorman, Mayer, & Tan, 2000; Jarvenpaa, Knoll, & Leidner, 1998; Ladegard & Gjerde, 2014; Schoorman & Ballinger, 2006; Schoorman, Mayer, & Davis, 1996). Yet, in addition to trusting in and relying on other humans at work, people must also trust in and rely on technology, increasingly technology embedded with artificial intelligence (AI). Based on the rapid developments in AI, scholars foresee a near future where people will work interdependently with computer programs enabled with AI to facilitate organizational decision-making tasks (Metcalf, Askay, & Rosenberg, 2019; Parry, Cohen, & Bhattacharya, 2016; Shrestha, Ben-Menahem, & Von Krogh, 2019; Tambe, Cappelli, & Yakubovich, 2019). Research, however, indicates that people's trust in and reliance on AI decision aids can be tenuous (Glikson & Woolley, 2020). The growing interest in and adoption of this technology in organizations thus raises the question, what facilitates a person's trust in and choice to rely on an AI decision aid at work?

The term "artificial intelligence" was coined by John McCarthy in the 1950s to describe "the science and engineering of making intelligent machines, especially intelligent computer programs" (McCarthy, 2007, p. 2). Today, AI is described as technology capable of gathering and interpreting data to complete cognitive tasks and generate solutions, decisions, and instructions, and learning based on feedback from its actions or new proposed examples in order to improve (Glikson & Woolley, 2020; Haenlein & Kaplan, 2019). AI can be embedded within different technologies and have a variety of functions. In the present research, we focus on artificially intelligent decision aids (AI decision aids), computer programs that use AI to generate decision alternatives or recommended courses of actions to achieve a specific objective (Shrestha et al., 2019; Shrestha, Krishna, & von Krogh, 2021; Von Krogh, 2018). Table 1 provides examples of AI decision aids used in work settings.

As AI decision aids can significantly improve organizational decision-making and free up employees to engage in other important work, it is important that people trust in and rely on them to carry out the decision-making tasks they are programmed for.[1,] Yet, the nature of AI decision aids could create unique challenges for trust and reliance. For instance, employees are likely to have little insight into the processes AI decision aids use to generate decisions or recommendations. While there is increasing focus on the

Table 1. Examples of AI Decision Aids Used in Practice.

Application domain	Example
Agriculture	Farmers work with AI decision aids that analyze climate data and images captured by satellites and drones to help improve the quality and accuracy of harvests.
Finance	Credit risk professionals work with AI decision aids that monitor and analyze large amounts of data related to a lending request in order to better predict the probability of default.
Healthcare	Medics work with AI decision aids to evaluate information available from medical imaging (e.g., x-ray, MR, ECG) to determine optimized treatment protocols and to assist with decision-making during surgery.
Human resources	Corporate recruiters work with AI decision aids that gather and analyze information from a large pool of job applicants to help identify high-potential candidates.
Manufacturing	Plant managers work with AI decision aids that gather and analyze data from smart sensors to monitor manufacturing equipment conditions in order to estimate when equipment maintenance should be performed.
Sales	Sales managers work with AI decision aids that compile and assess a variety of data sources to identify high-potential sales leads.
Supply chain	Supply chain managers work with AI decision aids to predict the amount of supplies and goods it needs to address forecasted demand.
Utility sector	Plant operators work with AI decision aid's to gather and analyze data from smart meters to detect supply and demand issues and to identify measures needed to prevent power outages.

design of transparent and explainable AI (Rai, 2020; Shin, 2021), its current "black box nature" is known to create challenges for trust. Furthermore, as AI decision aids are embedded in a computer system, they are unlikely to have a distinguishable identity that can be manipulated to influence trust in the program, like AI embodied in a robot or virtual agent (i.e., chatbot) (Glikson & Woolley, 2020). Employees might also have limited control over the AI decision aid, particularly aids that can select and execute decisions autonomously (O'Neill, McNeese, Barron, & Schelble, 2020). The inability to direct the AI decision aid towards desired outcomes could reduce the willingness to rely on this aid, particularly when there are questions about the AI decision aid's trustworthiness. Furthermore, there are likely to be salient risks associated with using AI decision aids. For example, the risk of being made redundant by an AI decision aid (Frey & Osborne, 2017) or losing important

skills (Parasuraman, Sheridan, & Wickens, 2000) or meaningful tasks (Langer, König, & Busch, 2021).

With this background in mind, the objective of our paper is to connect research on the topics of trust in AI and trust in organizational relationships to develop a conceptual model aimed at understanding what influences people's trust in and choice to rely on an AI decision aid to carry out an important decision-making task. Mayer et al.'s (1995) model of organizational trust is selected as a foundation for this work for several reasons. First, the model is generalizable to contexts comprised of human and non-human agents (Schoorman, Mayer, & Davis, 2007) and is therefore relevant for the domain of human trust in AI decision aids (Glikson & Woolley, 2020). Second, the model is well-suited to study the relationship between trust and perceived situational risk that is indicated to be salient in decisions to cooperate with AI-enabled systems (Glikson & Woolley, 2020; Hoff & Bashir, 2015; Stuck, Holthausen, & Walker, 2020) and could explain why trust in an AI decision aid may not result in reliance on this technology in practice. Furthermore, the model, while often cited in research on trust in AI, has rarely been applied as a theoretical framework for studying the phenomenon. This could be because the model is general and requires specification to its research context (Schoorman et al., 2007). To our knowledge, this model specification has not been undertaken for the context of trust in and reliance on AI decision aids. Finally, Mayer et al.'s (1995) model distinguishes between a number of concepts that are often confounded in the extant research on trust in AI. Accordingly, we believe that the model helps to develop greater conceptual clarity as research on the topic of trust in AI advances in the organizational science literature.

In the sections that follow, first, we briefly review the theory and empirical research that informs our conceptual model. We then review the core constructs and relationships specified in Mayer et al.'s (1995) model and compare them to constructs and relationships studied in relation to trust in AI decision aids or other relevant AI-enabled or automated systems. Using insight gained from this review, we redefine constructs from Mayer et al.'s (1995) model and the relationships expected between them to arrive at a conceptual model of trust, perceived risk, and human reliance on AI decision aids (see Figure 1) that addresses this unique human–AI working relationship. Our paper concludes with five next steps to take future research on the topic forward.

Theoretical and Empirical Foundations

In 1995, Mayer, Davis, and Schoorman published a conceptual paper that integrated several bodies of literature and diverse disciplines to arrive at

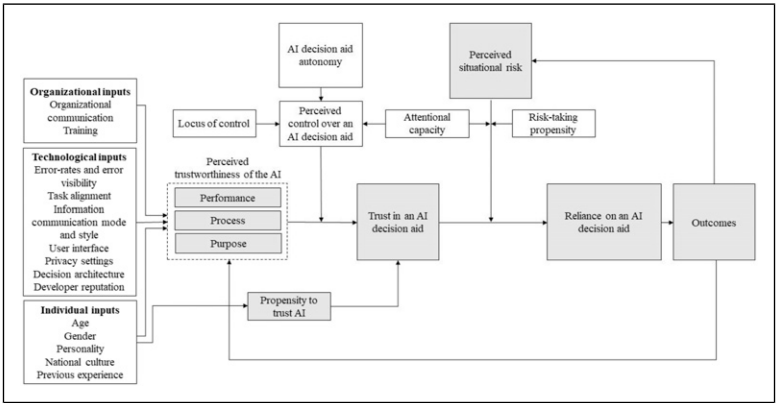

Figure 1. A conceptual model of trust, perceived risk, and human reliance on AI decision aids.

a single model of trust in organizations. According to their theorizing, trust is the willingness to be vulnerable to the actions of another party, which is informed by the trustor's perceptions of the other party's trustworthiness and of their own general propensity to trust. Trust, in turn, is posited to interact with the risk the trustor perceives in the situation to predict risk-taking in the relationship, that is, the decision to allow the other party to carry out an activity important to the trustor, regardless of the trustor's ability to monitor or control the trustee. The decision to rely on the trustee results in certain performance outcomes, the evaluations of which are expected to feedback into the system to provide further evidence for or against the trustee's trustworthiness. We will review the main concepts and mechanisms specified in their model in more detail in subsequent sections of the paper.

Though we have selected Mayer et al.'s (1995) model as the theoretical framework for our paper, we acknowledge that other theories of organizational trust are also relevant for understanding trust in AI decision aids. Notably, theories that emphasize the role that positive affect and emotions play in the formation of trust (e.g., McAllister, 1995) are highlighted as important to consider in the context of human trust in AI-enabled systems (Glikson & Woolley, 2020; Hoff & Bashir, 2015). While Mayer et al.'s (1995) model was originally developed as a cognitive model of trust, Mayer and colleagues acknowledged in their later work that recognizing the role of emotions adds an important dimension to their model (Schoorman et al., 2007). Therefore, in elaborating Mayer et al.'s (1995) conceptual model to domain of human trust

in AI decision aids, we take care to identify how features of and interaction with AI decision aids could shape both cognitive and emotional trust.

Much of the theorizing and empirical research that informs the AI-oriented aspect of our conceptual model comes from the trust in technology and trust in automation literature. Research exploring the diverse factors that shape trust in and interactions with computer-aided decision support systems (DSS) in different practical contexts also provides input to our work. AI decision aids in the present research are equivalent to AI-enabled DSS. Research in these literatures is largely informed by the same theories of trust that informed Mayer et al.'s (1995) integrative model. As such, they stand ripe to develop a conceptual model on trust in AI decision aids in organizational settings. However, these research domains are also very broad and there remain many inconsistencies with how trust and behavioral displays thereof are defined, conceptualized, and studied. It is outside of the scope of our paper to conduct a systematic review of, or to resolve all conflicts present in, these literatures. Our focus is rather on linking relevant material from these literatures to Mayer et al.'s (1995) model of organizational trust to specify and elaborate a model that better fits the context of human–AI decision aid work relationships.

In taking this selective focus, we also acknowledge that we are limited in the extent to which we link empirical research applying different theories to understand people's trust in and choice to rely on an AI decision aid with Mayer et al.'s (1995) model. We make note, in particular, of Davis's (1989) influential technology acceptance model (TAM), which specifies the perceived usefulness of a new technology and its perceived ease-of-use as key determinants of a person's intentions to use new technology and, thus, its subsequent usage. In the years since the TAMs introduction, scholars acknowledging the trust in technology and trust in automation literatures have extended the model to incorporate trust and perceived risk as correlates of perceived usefulness and ease-of-use (e.g., Beldad & Hegner, 2018; Featherman, 2001; Ghazizadeh, Lee, & Boyle, 2012; Im, Kim, & Han, 2008; Pavlou, 2003; Schnall, Higgins, Brown, Carballo-Dieguez, & Bakken, 2015). However, in this research, there are significant differences in how trust and risk are conceptualized and positioned in relation to each other and other concepts in the TAM as compared to in Mayer et al.'s (1995) model. Incorporating research that applies the TAM into our conceptual model would require a systematic review of these differences, something that is outside the scope of this paper.

A Conceptual Model of Trust, Perceived Risk, and Reliance on AI Decision Aids

Trust in AI Decision Aids

Mayer et al. (1995, p. 712) 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." The emphasis on vulnerability in their definition implies that there is a degree of risk for the trustor derived from uncertainty regarding the motives, intentions, and future actions of the other party (Kramer, 1999). Thus, trust is evident when a trustor is willing to give an important task or decision responsibility to another party despite perceiving a risk in doing so related to uncertainty about the trustee's behavior (Mayer & Davis, 1999).

Unlike other definitions of trust that also emphasize the willingness to accept vulnerability (e.g., Rousseau, Sitkin, Burt, & Camerer, 1998), Mayer et al.'s (1995) definition does not include any indication of positive beliefs or expectations about the trustee that make the willingness to give the trustee control over a particular action less risky or uncertain. As such, Mayer et al.'s (1995) definition of trust can also be differentiated from those that define trust only in terms of the positive expectations about the trustee's future actions (e.g., Robinson, 1996). This also differentiates Mayer et al.'s (1995) definition of trust from McAllister's (1995) definitions of cognition- and affect-based trust. According to McAllister (1995), cognition-based trust is confidence in a trustee based on positive beliefs about the trustee's competency and dependability, while affect-based trust is confidence based on positive beliefs about there being a reciprocal relationship between the trustor and the trustee that emphasizes personal care and concern. In Mayer et al.'s (1995) model, positive beliefs or expectations about the trustee are captured in the construct of perceived trustworthiness.

Research on trust in AI decision aids and other similar technology is also broad, and many definitions of trust exist in this literature. It is also the case that some studies do not formally define trust. Rather, they conceptualize trust in line with its dictionary definition, to rely on the technology, as it is generally conceptualized in the trust in technology literature (McKnight, Carter, Thatcher, & Clay, 2011). Trust is thus reflected in positive attitudes about relying on the AI-enabled system to perform its task (Gillath et al., 2021), expressed intentions to rely on the AI-enabled system (Höddinghaus, Sondern, & Hertel, 2021), or actual reliance behavior

(Oksanen, Savela, Latikka, & Koivula, 2020), which is also more generally viewed as the behavioral expression of trust (Lee & See, 2004).

When trust in AI decision aids and other similar technology is defined in research, Mayer et al.'s (1995) definition is often referred to as it does not specify trust in a way that is limited to interpersonal relationships (Chancey, Bliss, Yamani, & Handley, 2017; Glikson & Woolley, 2020). Lee and See's (2004, p. 54) definition of trust in automation, "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability," also influences research on trust in AI decision aids and other similar technology. Like Mayer et al. (1995), Lee and See's (2004) definition implies some degree of risk arising from uncertainty, and thus requiring the individual to accept vulnerability in trusting the non-human agent. However, Lee and See's (2004) definition also emphasizes the expectation of positive outcomes; the belief that the agent will help to achieve one's goals, which is emphasized in earlier definitions of trust in automation (e.g., Muir, 1994). Informed by this research, scholars studying trust in AI-enabled systems have defined trust in terms of confidence, based on the belief that the technology will perform its expected work and display favorable behavior (e.g., Aoki, 2021). Expanding the confidence-oriented conceptualization, Glikson and Woolley (2020) draw on McAllister's (1995) work to differentiate cognitive trust in AI, confidence based on rational evaluations of its performance quality and reliability, from emotional trust in AI, confidence based on affect or emotions.

Moreover, in their recent meta-analysis of the trust in AI research, Kaplan et al. (2021, p. 2) defined trust in AI as "the reliance by an agent that actions prejudicial to their well-being will not be undertaken by influential others." In this definition, the "agent" refers to the human and the "influential other" refers to an AI-enabled system. "Reliance" does not reflect a behavioral outcome of trust in an AI-enabled system, but rather the confidence that the AI-enabled system will not behave in ways that are detrimental to the trustor. Accordingly, Kaplan et al.'s (2021) definition is similar to others in the field, only it emphasizes expectations of non-negative outcomes as compared to positive outcomes resulting from the interaction with artificially intelligent technology.

Reviewing this literature, we see that there are several ways of defining trust that can be relevant in future research on trust in AI decision aids. However, Mayer et al.'s (1995) definition is applicable and accepted in research on trust in AI. It also has the benefit of not confounding the willingness to be vulnerable to an AI decision aid based on expectations of what the aid will do with factors that increase positive beliefs about the AI decision aid. Neither does it confound the willingness to be vulnerable to an AI decision aid

with a person's decision to rely on the aid (i.e., behavioral displays of trust). We, therefore, adopt Mayer et al.'s (1995) definition in our conceptual model, defining *trust in an AI decision aid* as the willingness of a person to be vulnerable to the actions of an AI decision aid, based on the expectation that it will perform a decision-making task important to the trustor. Beliefs about the AI decision aid and its outputs will be addressed in the concept of perceived trustworthiness and the factors that contribute to it, as discussed in the section that follows.

Perceived Trustworthiness of AI Decision Aids

A considerable contribution of Mayer et al.'s (1995) work is their literature review identifying the factors that contribute to the belief that the trustee will perform and behave in positive ways, that is, the perceived trustworthiness of the trustee. According to their review, perceived trustworthiness is a function of three things: First *ability*, or the extent to which the trustor perceives the trustee to have the ability to successfully carry out the tasks and responsibilities expected in the relevant domain or context. Ability is evident when the trustee is believed to have the knowledge, capabilities, influence, and qualifications for carrying out the task or responsibility (Mayer & Davis, 1999). The second factor, *integrity*, refers to the extent to which the trustee is perceived to adhere to a set of normative values and principles. The integrity of the trustee is evident when the trustor believes the trustee to be fair, consistent, and reliable in terms of doing what they say they will do (i.e., is dependable) (Mayer & Davis, 1999). The third factor, *benevolence*, refers to the extent to which the trustee is perceived to be caring, wanting to help, and having good intentions. A trustee's benevolence is evident when he or she is believed to be loyal, genuinely concerned with the trustor's needs and welfare, and dedicated to helping the trustor (Mayer & Davis, 1999). As would be expected, Mayer et al.'s (1995) model predicts a positive relationship between the perceived trustworthiness of the trustee based on perceptions of their ability, integrity, and benevolence and the trustors' trust in the trustee. This positive relationship is predicted because there should be an inverse relationship between the perceived trustworthiness of the trustee and the perceived risk of the trustee performing unsatisfactorily. Thus, the trustee's perceived trustworthiness is predicted to be positively related to the willingness to be vulnerable to his or her actions, because it reduces perceived risk in this relationship.

A long history of research in the trust in technology and trust in automation literature suggests that the factors that influence the trustworthiness of humans and the factors that influence the trustworthiness of technology are very

similar in nature (Lee & Moray, 1992; Madsen & Gregor, 2000; McKnight et al., 2011; Muir, 1994; Taddeo, 2009). Therefore, it is not surprising that research building on this literature to study trust in an AI decision aid has evaluated perceptions of its ability (the AI decision aid is competent), integrity (the AI decision aid is free from bias, is fair), and benevolence (the AI decision aid puts my interests first) (Höddinghaus et al., 2021). However, it is more common to see the trustworthiness of AI decision aids and other similar technology conceptualized as positive beliefs about unique technological features. Relating again to the trust in technology and trust in automation literatures (Lee & Moray, 1992; Madsen & Gregor, 2000; McKnight et al., 2011; Muir, 1994; Taddeo, 2009), these features include, but are not limited to, competence (the belief that the technology has the ability to perform its tasks), reliability (the belief that the technology will perform consistently over time), understandability (the belief that what the technology is doing, why it is doing it, and how it works is understandable), dependability (the belief that the technology can be counted on to do its job), and helpfulness (the belief that the technology helps the people who work with it).

The technological features identified above are also emphasized in Lee and Moray's (1992) three "bases of trust," a conceptual framework developed for studying trust in automation. In this framework, performance-based trust refers to the belief that an automation will perform in positive ways based on the perceived competency it has for carrying out a task. Performance-based trust has been likened to ability in Mayer et al.'s (1995) model (Chancey et al., 2017; Hoff & Bashir, 2015), as it is reflected in the belief that the automation performs its task capably and reliably.[2] The importance of performance-based trust in relation to AI decision aids and other AI-enabled systems is supported by Kaplan et al.'s (2021) meta-analysis, which finds that performance quality and reliability contribute positively to beliefs about AI's trustworthiness. Performance-based trust is particularly important for what Glikson and Woolley (2020) refer to as cognitive trust in AI.

The second factor identified by Lee and Moray (1992, p. 1246) is process-based trust or trust in automation based on "an understanding of the underlying qualities or characteristics that govern [its] behavior," such as data reduction methods, rule bases, or control algorithms. Process-based trust is held to be enhanced when a person understands what the automation does, why it does it, and how it works (Chancey et al., 2017; Hoff & Bashir, 2015). The understandability of AI-enabled systems is sometimes conceptualized in terms of perceived transparency, or the insight a person believes they have over the operating rules and inner logics used by the AI (Glikson & Woolley, 2020). Kaplan et al.'s (2021) meta-analysis supports that transparency is positively associated with the perceived trustworthiness of AI-enabled

systems. Like its performance, transparency is identified as important for building cognitive-based trust in AI (Glikson & Woolley, 2020).

Scholars also sometimes liken process-based trust to integrity in Mayer et al.'s (1995) model (Chancey et al., 2017; Hoff & Bashir, 2015). While the connection is not made explicit, it could be because Lee and Moray (1992) compare the data reduction methods, rule bases, and control algorithms that govern how automation behaves to the stable dispositions and character traits that explain human behavior. Knowing a person's dispositions and character traits will influence our beliefs about if they will actually do what they say they will do, which is aligned with perceived integrity in Mayer et al.'s (1995) model. Similarly, understanding automation's data reduction methods, rule bases, and algorithms could influence our beliefs that the automation will do what it says it will do and can thus be depended on in a way that transcends its performance reliability. Research on trust in AI-enabled systems indicates a further connection between integrity and process-based trust. Just as a trustee's perceived adherence to a set of normative values and principles is important for the integrity factor of trustworthiness in Mayer et al.'s (1995) model, research finds that the belief that AI-enabled systems comply with contextual norms and rules is important for its perceived trustworthiness (Kaplan et al., 2021). Glikson and Woolley (2020) suggest that AI's compliance with norms and standards should enhance emotional trust in AI. Accordingly, it seems reasonable that process-based trust can be enhanced through cognitive mechanisms related to the understandability/transparency of AI's data gathering, analysis, and learning processes and through emotional mechanisms related to the perceived qualitative characteristics of these processes.

Moving on, the third factor identified by Lee and Moray (1992), purpose-based trust, is based on the understood purpose of the automation, that is, perceptions of why it was designed and who it is intended to benefit. Purpose-based trust is held to be evident when a person believes that an automation is intended to make their work more efficient and/or effective, that is, to help them perform well (Chancey et al., 2017). Therefore, it is largely equivalent to the helpfulness attribute specified in the trust in technology literature, which is positioned in relation to benevolence in Mayer et al.'s (1995) model (McKnight et al., 2011). The notion of purpose-based trust is evident in research on AI-enabled systems that finds trust in the technology to be reduced when it is believed to serve the purpose of monitoring a person's activity or coercing their behavior for someone else's benefit (Alan et al., 2014; Möhlmann & Zalmanson, 2017), as this elicits negative cognitions and emotions about the technology (Glikson & Woolley, 2020). Thus, purpose-based trust should be evident when a person believes that an AI-enabled

system is intended to help them, not when it is perceived to harm or coerce them. It is important to point out that purpose-based trust does not refer to the purpose the AI itself is perceived as intending to serve. AI cannot (at this point) be ascribed intentions. Rather, it is based on perceptions of why the AI was designed and who it is intended to benefit, as specified by its developers. Perceived intentions could also extend to those who have implemented it in a given context (e.g., management).

Based on this review, we conclude that future research can conceptualize the perceived trustworthiness of an AI decision aid as a function of its perceived competence, integrity, and benevolence (Höddinghaus et al., 2021). This route may be practical when comparing trust in AI decision aids to trust in human decision makers. However, Lee and Moray's (1992) conceptualization of performance-, process-, and purpose-based trust is more likely to capture the unique technological features important for facilitating positive beliefs about and trust in AI decision aids (Chancey et al., 2017). Accordingly, in our conceptual model, we define the *perceived trustworthiness of an AI decision aid* as the belief that the AI decision aid will perform and behave in positive ways, based on its perceived *performance* (perceptions that an AI decision aid performs its task capably and reliably), *processes* (perceptions that the processes used by an AI decision aid are transparent, dependable, and adhere to normative values and principles), and *purpose* (perceptions that an AI decision aid is intended to help those who work with it perform their job better). As in Mayer et al.'s (1995) model, we expect that positive beliefs about an AI decision aid's performance, processes, and purpose will reduce the perceived risk of the AI decision aid performing unsatisfactorily or behaving unfavorably. Thus, these factors are predicted to be positively related to the willingness to be vulnerable to the actions of the AI decision aid (i.e., trust in the AI decision aid) because they reduce the perceived risk in this relationship. Accordingly, we propose:

Proposition 1: *Beliefs about an AI decision aid's performance, processes, and purpose will be significantly related with trust in the AI decision aid.*

However, in making this proposition, we also acknowledge that the dynamic relationship between perceptions of an AI decision aid's performance, processes, and purpose and trust in an AI decision aid could be different than the dynamic between perceived ability, integrity, and benevolence in predicting trust in interpersonal relationships. In Mayer et al.'s (1995) model, perceived ability, integrity, and benevolence are proposed to reinforce each other in contributing to perceived trustworthiness and, in turn, interpersonal trust, a proposition also supported by empirical research (Colquitt, Scott, &

LePine, 2007). Yet, beliefs about an AI decision aid's performance and processes could offset each other. For example, an AI decision aid that elicits positive beliefs about its performance might use complex algorithms that undermine beliefs about the understandability of its processes (Lee & See, 2004). It is also reasonable to believe that perceptions of an AI decision aid's purpose could color perceptions of its performance and processes (Muir, 1994). Therefore, unlike Mayer et al. (1995), who predict a simple, additive model of the factors contributing to perceived trustworthiness, we refer to Muir (1994) and suggest that a more complex model may be appropriate when considering how the perceived performance, processes, and purpose relate to the trustworthiness of AI decision aids.

Furthermore, how perceptions of an AI decision aid's performance, processes, and purpose influence trust in the technology over time could differ from how perceived ability, integrity, and benevolence are expected to influence interpersonal trust longitudinally. Mayer et al. (1995) propose that the effect of perceived ability and integrity on trust will be most salient early in the relationship between the trustor and the trustee, prior to the development of meaningful benevolence data. However, over time, interpersonal trust based on perceived ability and integrity will give way to exchanges based on perceived benevolence of the trustee, similar to how an economic exchange relationship transforms into a social exchange relationship over time (Blau, 1968; Cropanzano & Mitchell, 2005). Evidence from the trust in automation literature, on the other hand, suggests that trust in an AI decision aid should begin with faith in the technology based on purpose-based trust (Lee & See, 2004), which gives way to a trial-and-error period where perceptions of its performance and processes can develop (Zuboff, 1988). Thus, how a person comes to trust in an AI decision aid is held to align more closely with emotion-oriented theories of trust evolution (Jones & George, 1998). The reason why trust in an AI decision aid is anticipated to begin with purpose-based trust is because it is expected that a description of the intended purpose and benefits of using the technology will likely be communicated to a person by their organization prior to its usage (Lee & See, 2004). On the other hand, it is also likely that organizational inputs such as communication or training could inform initial perceptions of an AI decision aid's performance and processes (Hoff & Bashir, 2015). Accordingly, while we expect that the longitudinal relationship between factors contributing to perceived trustworthiness of an AI decision aid and trust in the aid will not necessarily follow the same pattern predicted in research on trust in interpersonal relations, we also predict that relationships could vary depending on several factors.

In identifying perceptions of an AI decision aid's performance, processes, and purpose as key factors contributing to its trustworthiness, it is also relevant

to make note of some technological attributes that are likely to shape these perceptions. For example, just as poor performance can negatively influence the perceived ability of humans, factors such as error rates and error visibility can negatively influence the perceived performance of AI decision aids (Glikson & Woolley, 2020). Furthermore, like ability in Mayer et al.'s (1995) model, the perceived performance of an AI decision aid hinges on how well-suited the aid is perceived to be for the task. AI is assumed to perform better on analytical, data-intensive tasks than on tasks that require softer "human skills" such as communication and creativity (Glikson & Woolley, 2020). Accordingly, the perceived performance of an AI decision aid is likely to be more positive when it is used to perform a decision-making task appropriate for its analytical and information processing capabilities (Lee, 2018).

Similarly, positive perceptions about an AI decision aid's processes could be enhanced when the decision aid provides helpful explanations about the data and algorithms used or the outputs delivered, as this is important for increasing the understandability of the technology (Glikson & Woolley, 2020). However, how this information is presented by the decision aid matters. While AI decision aids may not have a distinct physical identity, like a robot or chatbot, research indicates that communication cues in the AI decision aid's user interface can influence the understandability of information presented as well as the acceptance of the information provided (Pak, Fink, Price, Bass, & Sturre, 2012).

As previously addressed, beliefs about the purpose of the AI decision aid are likely to be influenced by communications made by the organization (J. D. Lee & See, 2004). However, certain technological attributes, such as privacy settings, could help people working with an AI decision aid feel assured that their own behavior is not being monitored while using the aid (Wang, Sun, & Bertino, 2014). Furthermore, research indicates that features of an AI decision aid that direct the choice of a human partner (nudges, boosts) should do so in ways that are perceived as helping the person achieve personal work objectives (Burr, Cristianini, & Ladyman, 2018; Glikson & Woolley, 2020), otherwise positive beliefs about the purpose of the AI decision aid could be undermined. Finally, the technology developer's reputation should be important for forming positive beliefs about an AI decision aid (Kaplan et al., 2021), as it can inform perceptions of the aid's intended purpose, as well as its performance and processes.

We add the organizational and technological factors identified in this section to our conceptual model as inputs that contribute to the trustworthiness of an AI decision aid. However, we acknowledge that this is just a small selection of what research indicates is important for beliefs about an AI decision aid's performance, processes, and purpose.

Propensity to Trust AI Decision Aids

Beyond perceived trustworthiness, the propensity to trust others is also important for explaining trust in Mayer et al.'s (1995) model. The propensity to trust is "a generalized expectation about the trustworthiness of others" (Mayer et al., 1995, p. 715) that is evident when an individual believes that most people are capable, have integrity, and are benevolent (Mayer & Davis, 1999). It is described as a stable individual trait influenced by developmental experiences, personality types, and cultural backgrounds (Rotter, 1967). Mayer et al. (1995) propose that trust will be a function of the trustee's perceived trustworthiness and the trustor's propensity to trust, with the propensity to trust being particularly important prior to forming perceptions about the trustee's ability, integrity, or benevolence. Meta-analytical research supports that the propensity to trust has a stronger, positive relationship with trust prior to the inclusion of information about the perceived ability, integrity, and benevolence of the trustee; however, it also continues to have a small positive influence on trust beyond these perceptions (Colquitt et al., 2007).

Similarly, in the trust in technology and trust in automation literatures, it is understood that people do not simply base their trust in a technology on its technological attributes (McKnight et al., 2011; Muir, 1994). In this literature, "dispositional trust" has been identified as the tendency to be willing to depend on a technology across a broad spectrum of situations (McKnight et al., 2011), or as a person's tendency to trust automation, independent of context or a specific system (Hoff & Bashir, 2015), which predicts trust in a technology beyond its technological attributes. Hoff and Bashir (2015) identify national culture, age, gender, and personality as four primary sources of dispositional trust. Thus, the factors that lead to a generalized expectation about the trustworthiness of humans (Rotter, 1967) are also likely to contribute to generalized beliefs about the trustworthiness of specific technologies.

Accordingly, in the present research, we define the *propensity to trust AI decision aids* as a person's general tendency to perceive AI decision aids as trustworthy, independent of their perceptions of a specific AI decision aid's performance, processes, or purpose, and propose:

> **Proposition 2:** *Trust in an AI decision aid will be a function of its perceived trustworthiness and the trustor's general propensity to trust AI decision aids.*

However, going beyond Mayer et al.'s (1995) model, we believe there is a need to account for how individual factors that contribute to the propensity to trust AI decision aids could also create a perceptual filter that influence beliefs

about the AI decision aid's trustworthiness. Presently, a number of individual factors are identified as influencing trust in AI-enabled systems, including national culture, gender, personality, and expertise (the understanding of a specific domain that results from long experience) (Kaplan et al., 2021). However, it is unclear if these factors influence trust directly or through enhanced perceptions of AI's performance, processes, or purpose. It is reasonable to expect that factors such as gender, age, and personality could influence the perceived performance, processes, and purpose of a specific AI decision aid just as it could influence general beliefs that AI decision aids are trustworthy. Accordingly, our conceptual model indicates that individual factors likely influence both the propensity to trust AI decision aids as well as beliefs about their trustworthiness.

Perceived Control over AI Decision Aids

Control is defined as "the mechanisms used to specify, measure, monitor, and evaluate other's work in ways that direct them towards the achievement of desired objectives" (Long & Sitkin, 2018, p. 725). Mayer et al. (1995, p. 106) did not include control as a distinct concept in their model, beyond defining trust as "the willingness to be vulnerable to the actions of another party-…*irrespective of the ability to monitor or control that party"* (italics not in original text). However, they later acknowledged control as "an alternate mechanism for dealing with risk in relationships" (Schoorman et al., 2007, p. 346). Indeed, having ways to direct another party's performance and behavior towards desired outcomes should reduce the perceived risk that the party may perform unsatisfactorily or behave unfavorably (Das & Teng, 2001). Hence, as a trustor's ability to control the performance and behavior of the trustee increases, the willingness to be vulnerable to the trustee's actions should also increase, even at lower levels of perceived trustworthiness. Alternatively, as control over the trustee decreases, the willingness to be vulnerable to their actions should also decrease, unless perceived trustworthiness is high enough to reduce this perceived risk.

Just as control over a human trustee will naturally be limited, so too will people's control over an AI decision aid. While people may be able to specify the data inputs fed into an AI decision aid, they will never have full control over the processes it uses to complete cognitive tasks, generate decisions or decision alternatives, and learn based on feedback. On the other hand, people might have some control over the AI decision aid based on the autonomy granted to the technology. As the autonomy granted to an AI decision aid is likely to be more formalized than the autonomy granted to human work

partners, we feel it is important to include control as an explicit concept in our model.

O'Neill et al. (2020) discuss different levels of autonomy that could be granted to an AI-enabled system, which should in part inform the level of control a person has over it. Adapting this discussion to the context of AI decision aids, we could first envision an AI decision aid that is tasked with determining the complete set of decision alternatives that their human partner must select from. In this scenario, the AI decision aid has no autonomy and the person working with the AI decision aid has significant control, as they can evaluate all alternatives and specify the choice they believe is best for the achievement of desired objectives. As the autonomy granted to an AI decision aid increases, it may narrow down the decision alternatives or even suggest one best alternative. It may even execute that alternative if its human partner approves or fails to veto the decision within a given time frame. A person's control over the AI decision aid is increasingly reduced in these scenarios because they cannot evaluate the alternatives to ensure the superiority of the alternative(s) selected or specify which alternative from the narrower range of choices to select. However, research indicates that some additional control might be gained if the human partner has the autonomy to modify the AI decision aid's outputs, that is, adjust its decision alternative(s) within a specified range (Dietvorst, Simmons, & Massey, 2018). Finally, at its highest level of autonomy, an AI decision aid may determine the set of decision alternatives, select the best one, and execute it, without involving a human partner. In this scenario, the human partner is "out of the loop."[3] They have no control over the AI decision aid as there are no mechanisms through which they can direct the AI decision aid's outputs towards desired outcomes.

Given the inverse dynamics expected between perceived risk and control (Das & Teng, 2001), we expect that people should express greater trust in an AI decision aid having lower levels of autonomy even when the perceived trustworthiness of the decision aid is also low. On the other hand, when the perceived trustworthiness of the AI decision aid is high, people should be more willing to allow it to perform its task autonomously as there is lower perceived risk that the AI decision aid will not perform or behave as desired. However, there is also some indication that high control-low trustworthiness conditions elicit greater trust in AI decision aids than low control-high trustworthiness conditions. For example, research conducted by Ho, Pavlovic, Myers, and Arrabito (2013) found that people trusted automation with lower levels of autonomy more than automation with higher levels of autonomy, even when low-autonomy automation was less reliable than high-autonomy automation. Accordingly, control could be the more dominant

factor in influencing trust in AI decision aids, underlining the importance of including it in our conceptual model.

There are also factors that could influence the control a person has or perceives to have, over an AI decision aid beyond the level of autonomy granted to the technology. Notably, in their review of research on trust in automation, Hoff and Bashir (2015) identify a person's attentional capacity as a factor that can affect the time and cognition they spend monitoring automation. Attentional capacity refers to the degree to which a person can focus on different tasks simultaneously (Kahneman, 1973) and is influenced by factors such as workload, task complexity, exhaustion, motivation, and boredom (Hoff & Bashir, 2015). When a person's attentional capacity is high, they should have the time and cognitive resources needed to monitor and evaluate the AI decision aid's outputs and specify the selection of decision alternatives in a way that directs performance towards desired outcomes, to the extent that it is possible based on the level of autonomy granted to the AI decision aid.

Furthermore, a person's locus of control (LOC; Rotter, 1966) should also be important for perceived control over an AI decision aid in this context. Research indicates that people having an internal LOC trust an AI decision aid more than individuals who have an external LOC (Sharan & Romano, 2020). This is likely because people having an internal LOC feel capable of influencing their environment. Therefore, they should be more likely to specify and direct the performance of an AI decision aid in ways that reduces the risk of it not performing as expected, more so than people with an external LOC, who believe they have little ability to control their life and things in it (Stanton & Young, 2000).

In light of the discussion above, we define *perceived control over an AI decision aid* as "the extent to which a person perceives the ability to direct the outputs of an AI decision aid towards the achievement of desired objectives" and elaborate its role in our model with the proposition:

> **Proposition 3:** *Perceived control over an AI decision aid will moderate the positive relationship between its perceived trustworthiness and trust, such that trust in the AI decision aid will be higher when perceived control over the decision aid is high even at lower levels of perceived trustworthiness.*

Furthermore, we indicate the influence different factors could have on perceived control in our conceptual model, including the autonomy granted to the AI decision aid as well as a person's attentional capacity and locus of control. As in earlier sections of this paper, we acknowledge that this is not

a comprehensive summary of relevant factors, but rather examples that could be examined and expanded on in future research.

Perceived Situational Risk

Beyond the concept of trust and the factors that contribute to it, perceived risk is a central concept specified in Mayer et al.'s (1995) model. Trust is generally held to reflect the willingness to put oneself at risk. However, Mayer et al. (1995) intentionally separate the perceived risk of the trustee performing poorly from the perceived risk of loss or other disappointing outcomes in the situation that makes the decision to entrust the trustee with an important task significant and uncertain. Perceived risk in their model therefore reflects perceived situational risk, that is the trustor's beliefs about the extent to which relying on another party in a particular situation will result in undesirable outcomes, even if they experience trust for the particular party in question based on positive beliefs about their expected performance. Mayer et al.'s (1995) model indicates that perceived situational risk moderates the relationship between trust and behavioral displays of trust, but a specific proposition is not made about the nature of this interaction. However, it is logical to predict that perceived situational risk creates a boundary condition where the relationship between trust and behavioral displays thereof will be restricted when perceived situational risk is greater than trust in the trustee. Alternatively, when perceived situational risk is low, behavioral displays of trust could proceed even at lower levels of trust.

As in Mayer et al.'s (1995) model, perceived situational risk is identified as important in research on trust in automated and AI-enabled systems (Glikson & Woolley, 2020; Hoff & Bashir, 2015), particularly in safety-critical contexts where situational hazards or felt responsibility for catastrophic consequences make the decision to trust this technology significant and uncertain. However, most research in this domain has not systematically separated the perceived risk inherent in the relationship with the technology under study from perceived situational risks, nor has it considered the interplay between trust and perceived situational risk in predicting behavior. Furthermore, as research in this domain has been largely experimental, the salience of the perceived situational risks that people might actually experience in practice has often been restrained (Glikson & Woolley, 2020). In simulation studies that directly attempt to manipulate perceived situational risk, it has been difficult to create conditions realistic enough to moderate the relationship between trust and behavioral displays of trust as predicted by theory (Chancey et al., 2017; Hösterey & Onnasch, 2021).

As a remedy to the current state of research, recent work by Stuck et al. (2020) is aimed at helping scholars studying trust in technology better account for perceived situational risk in their models. As an initial step in this direction, Stuck et al. (2020) have expanded on a framework developed by Jacoby and Kaplan (1972) to specify nine domains of situational risk that can be considered in research on trust in robots. These domains reflect beliefs about the extent to which relying on a robot in a particular situation will result in undesirable outcomes, regardless of the person's beliefs about the trustworthiness of the particular robot with which they could be interacting. We believe Stuck et al.'s (2020) framework could be helpful for facilitating future research concerned with the perceived situational risk of working with AI decision aids. Therefore, in Table 2, we adapt their framework to define and exemplify each risk domain within the context of human–AI decision aid work relationships.

To date, no research that we know of has attempted to determine what domain of perceived situational risk could be most important in decisions to rely on an AI decision aid to carry out a decision-making task. However, it is plausible that financial risk could be prominent, particularly the perceived risk of salary loss, given the number of people who are likely to believe that their role could be made redundant by an AI decision aid (Frey & Osborne, 2017). We also anticipate higher perceptions of performance risk, as people could believe that relying on AI decision aids could negatively influence their own performance through, for example, the loss of situational awareness of loss of important work skills (Parasuraman et al., 2000). Psychological risk related to the loss of task identity and job satisfaction could also be an issue if decision-making responsibilities associated with higher work motivation are delegated to AI (Langer et al., 2021). We believe it is likely that these and other perceived situational risks could result in a person's decision not to use an AI decision aid even if the person experiences trust in the AI decision aid based on positive beliefs about its performance, processes, and purpose.

Accordingly, in line with Mayer et al.'s (1995) model, we define *perceived situational risk* in this context as a person's beliefs about the extent to which relying on an AI decision aid to carry out its intended task will result in undesirable outcomes, outside of considerations about the trustworthiness of the AI decision aid. Furthermore, we make the following proposition:

Proposition 4: *Perceived situational risk will moderate the relationship between trust in an AI decision and reliance on it, such that the relationship will be less positive when perceived situational risk is high.*

Table 2. Domains of Situational Risk in the Context of Using AI Decision Aids.

Risk domain	Definition and example
Financial risk	The belief that one could lose money if they let AI decision aids take over important tasks. *Example*: A social benefits administrator believes that she will lose her job, and thus her financial security, if she allows AI decision aids to take over key parts of her current task work.
Performance risk	The belief that relying on AI decision aids could have negative performance implications. *Example*: A credit risk manager believes that using AI decision aids to carry out customer risk assessments will create issues with his customers, who expect a subjective consideration of their personal circumstances.
Physical risk	The belief that relying on AI decision aids could lead to damage, physical harm, or negative health impacts. *Example:* An air defense system operator believes that relying on AI decision aids to prioritize targets reduces his situational awareness of the airspace, which could lead to errors with potentially unfortunate outcomes.
Psychological risk	The belief that relying on AI decision aids does not align with a person's identity or may lead to negative psychological states. *Example:* A manager believes that using AI decision aids will take away interesting work and will thus diminish the job satisfaction he experiences.
Social risk	The belief that relying on AI decision aids could impact the way that people think about the person. *Example.* A doctor perceives that using AI decision aids to assist in diagnosis would undermine her credibility and expertise in the minds of her patients.
Time loss risk	The belief that one would be using time inefficiently or ineffectively by relying on AI decision aids to carry out a specific task. *Example:* An architect perceives that collecting and entering the parameters needed for AI decision aids to generate the best possible building configuration for a location is not worth his time when he and his colleagues have some good ideas already.
Ethical risk	The belief that relying on AI decision aids could be viewed as immoral or incongruent with the moral beliefs or values of an individual. *Example:* A manager perceives that using AI decision aids for remuneration decisions is unethical because too much employee surveillance is required to collect the personal data the system needs.

(continued)

Table 2. (continued)

Risk domain	Definition and example
Privacy risk	The belief that relying on AI decision aids could expose personal information about the user or their surroundings. *Example:* A sales manager perceives that using AI decision aids to optimize route planning would reveal that he drives home at lunch to spend time with his family.
Security risk	The belief that relying on AI decision aids could make the situation vulnerable to crime, sabotage, attack, or some other threat to safety. *Example:* A power plant operator perceives that relying on AI decision aids makes the system more susceptible to cyber attack.

However, with this proposition stated, we also note that there are many benefits associated with using AI decision aids that will likely be weighed against perceived situational risks to determine whether relying on an AI decision aid in a given situation is a good decision or if the risks are too high. It is therefore also important to identify the combination of factors contributing to the larger risk-benefit analysis associated with the decision to rely on an AI decision aid when considering what moderates the relationship between trust and behavioral displays thereof.

We also acknowledge that several factors may reduce the influence perceived situational risk has on the relationship between trust in an AI decision made and behavioral displays of trust. For example, Stuck et al. (2020) identify a person's risk-taking propensity (general tendency to take risks) as an individual difference that will positively influence a person's likelihood of relying on a robot, despite the perceived situational risk of doing so. Similarly, we could expect a person with a high risk-taking propensity to be more likely to rely on an AI decision aid, even when they perceive a situational risk associated with this behavior. Furthermore, we also find it possible that a person's attentional capacity, which we discussed earlier in relation to perceived control over an AI decision aid, could influence the extent to which a person takes heed of perceived situational risks. Research indicates that people over rely on automated systems when working under high workload conditions, regardless of the perceived risks of doing so (Biros, Daly, & Gunsch, 2004). Workload is a primary factor that affects a person's attentional capacity and thus the time and cognitive resources they can apply to process risks in their immediate environment. Accordingly, we identify a person's risk-

taking propensity and attentional capacity in our conceptual model as factors that could moderate the extent to which perceived situational risk impacts the relationship between trust in an AI decision aid and behavioral displays of trust. However, there are likely many other factors that could be relevant to consider in this context.

Risk-Taking in the Relationship: Reliance on (and Compliance With) AI Decision Aids

Coming to the predicted outcome of their model, Mayer et al. (1995, p. 724) refer to risk-taking in the relationship as "the behavioral manifestation of the willingness to be vulnerable" (i.e., trust) that is displayed when the trustor allows the trustee to perform a particular, important action. However, as indicated in the preceding section, risk-taking in the relationship is not simply viewed as the behavioral display of trust in their model, but the outcome of the interaction between trust and perceived situational risk.

Similarly, in research on automation, the decision to rely on a technology by allowing it to be responsible for important tasks or outcomes is often identified as the behavioral outcome of trust (Lee & See, 2004). Some research in this field, however, also goes on to distinguish between reliance, which is evident when a person lets the automated system do what it is supposed to do, and compliance, which is evident when the person responds in ways expected to a signal given by the automation (Chancey et al., 2017; Meyer, 2001). Exemplifying these concepts in context of AI decision aids, reliance would be evident when a person allows an AI decision aid to carry out its intended decision-making task while compliance would be evident when the person accepts and proceeds with the decision alternative recommended by the AI decision aid. Depending on the autonomy granted to an AI decision aid, both reliance and compliance may be important outcomes to consider in future research on the topic. We therefore include both in our conceptual model, defining *reliance in AI decision aids* as the decision to allow an AI decision aid to carry out an important decision-making task and *compliance with an AI decision aid* as the decision to comply with decision(s) made by an AI decision aid.

It is also relevant to acknowledge that much of the research on trust in AI and other relevant technology has been undertaken with the goal of correctly calibrating trust in the technology, such that it is appropriately relied on and complied with in practice (Hoff & Bashir, 2015; Lee & See, 2004). Accordingly, the level of trust a person has for an AI decision aid they work with should correspond to the actual level of performance the decision aid can

deliver in a given context to ensure safe interaction and optimal utilization of the technology's capabilities. Too high trust could lead to over-reliance on the AI decision aid and thus to misuse (Parasuraman & Riley, 1997), particularly if the person fails to notice an error or follows a faulty recommendation generated by the decision aid (Bahner, Elepfandt, & Manzey, 2008). On the other hand, too low trust may lead to under-utilization and AI decision aid disuse as the person chooses not to use the decision aid even though it could enhance performance (Parasuraman & Riley, 1997). Both misuse and disuse describe inappropriate levels of reliance could be important to consider in research on trust in AI decision aids, particularly where misuse/disuse could have negative implications for performance and safety outcomes.

Outcomes and Feedback to the System

Following from the previous section, Mayer et al. (1995) acknowledge that risk-taking in the relationship (i.e., behavioral displays of trust) should result in a particular outcome, which in turn will feedback into the system to inform perceptions of the trustee's perceived trustworthiness. When a trustor allows the trustee to perform an important action and this leads to positive outcomes, perceptions of the trustee's trustworthiness are expected to be enhanced. If outcomes of a behavioral display of trust are negative, it is expected that perceptions of the trustworthiness of the trustee will be diminished. Outcomes, in this case, can be both performance-related (the results achieved by the trustee) and affective responses or emotions experienced in relation to the process of allowing the trustee to carry out important actions and the results achieved (Jones & George, 1998; Schoorman et al., 2007).

Similarly, research on trust in automation refers to the concept of dynamic learned trust, which refers to the trust that develops or is attenuated during interaction with an automated system (Hoff & Bashir, 2015). Like predictions made by Mayer et al. (1995), research in this domain indicates that positive experiences using automated systems enhance trust, while negative experiences reduce it (e.g., Manzey, Reichenbach, & Onnasch, 2012; Yuviler-Gavish & Gopher, 2011). However, the term dynamic learned trust does not refer to experienced positive and negative outcomes, but to the effect these outcomes have on trust in the automated system via augmented perceptions of the system's performance, processes, or purpose. Dynamic learned trust is differentiated from initial learned trust, or trust based on pre-existing knowledge of the automated system's trustworthiness, from communication, training, and previous experience with similar technology (Hoff & Bashir, 2015).

To clarify the experienced outcomes of relying on an AI decision aid from their influence on the perceived trustworthiness of, or trust in, the AI decision aid, we retain the concept name "Outcomes" in our model, referring to the experience of positive or negative outcomes that result from relying on an AI decision aid. We also account for the feedback mechanism between outcomes and perceptions of AI decision aid trustworthiness, as indicated in Mayer et al.'s (1995) model. However, we believe it is also logical to expect that experience relying on an AI decision aid will also influence perceived situational risk, as experience should lend support or disconfirm beliefs about the potential losses or negative outcomes in the situation. As such, we propose that a feedback mechanism should also be accounted for between outcomes and perceived situational risk.

Proposition 5: *Outcomes of relying on an AI decision aid will feedback into the system to inform and adjust perceptions of the AI decision aid's trustworthiness and perceptions of situational risk.*

Discussion and Ways Forward

The conceptual model proposed in this paper specifies and elaborates Mayer et al.'s (1995) model of trust in organizations to the context of human–AI decision aid work relationships. As indicated, many studies on trust in AI and other relevant technology are available to inform this model and future research on the topic. Other scholars have provided a more comprehensive review of this research (e.g., Glikson & Woolley, 2020). However, they have not put forward a conceptual model that distinguishes between the multitude of concepts studied in this domain or the relationships between them and their boundary conditions. The present paper, therefore, complements the existing literature by providing a framework that can facilitate more systematic research activity on the topic.

While we believe that Mayer et al.'s (1995) model was an appropriate foundation for our work, we also identified some deficiencies in their model when translating it to the context of trust in AI decision aids. Notably, we found that the perceived trustworthiness of an AI decision aid could be better conceptualized in terms of beliefs about its performance, processes, and purpose. We also found that control over an AI decision aid and its implication for trust was important to account for explicitly in the model. Furthermore, Mayer et al.'s (1995) model was underdeveloped with regards to specifying the nature of perceived situational risk and how it interacted with trust in predicting reliance on a trustee. We also made note of a number of relationships not previously accounted for in Mayer et al.'s (1995) model and specified a number of inputs that could be important for eliciting

perceived trust and displays thereof in our particular context of interest. But where should researchers interested in studying trust in AI decision aids go from here? Below, we outline five next steps needed to take research on the topic forward.

Operationalizing Core Concepts and Developing Valid Measures

As indicated in the paper, scholars engaged in research on trust in AI have not always been precise in defining and operationalizing trust and related constructs, such that "there is an urgent need for addressing the great variance in measures used to assess human trust in AI" (Glikson & Woolley, 2020, p. 651). Our conceptual model provides a framework to differentiate trust in AI decision aids from other related concepts and thus sets the path for more systematic measure development. However, additional care is still needed with regards to ensuring that concept definitions are operationalized accurately in empirical research. Existing survey measures are available to measure many of the concepts defined in our model (e.g., Chancey et al., 2017), as are discussions of how to measure these concepts (Kohn, De Visser, Wiese, Lee, & Shaw, 2021; Wei, Bolton, & Humphrey, 2020). However, researchers must ensure the content validity of any measures they use by checking that they correspond accurately and uniquely with the target construct. Available measures of trust and other related constructs may need to be adapted substantially and thus warrant additional validation work in line with expert guidelines (e.g., Heggestad et al., 2019).

Applying Theory to Develop Arguments for the Relationships Between Constructs

As previously noted, while Mayer et al.'s (1995) model of organizational trust is often referred to in research on trust in AI, it is rarely applied in practice. Nor are other theoretical frameworks widely applied by researchers studying trust in AI, as evident by a relative absence of the word "theory" in comprehensive reviews of the literature (Glikson & Woolley, 2020; Hoff & Bashir, 2015). Lack of a theoretical model is less problematic when the relationship being examined is the one between the reliability of an AI decision aid and perceptions of its trustworthiness. However, tests of more complex indirect or conditional effect relationships could be much improved with theory-based arguments. The conceptual model put forward in this paper provides a framework to base such arguments, but it must be applied for this purpose. As mentioned earlier in the paper, the Technology Acceptance Model, particularly expanded models including trust and perceived risk, could be another

theoretical framework that could facilitate more systematic research on the topic.

Conducting Field Studies with Real AI Decision Aids and Real Situational Risks

Much of the research on trust in AI decision aids to date has been conducted in the laboratory, often using "Wizard-of-Oz" or prerecorded technology rather than actual intelligent systems (Kaplan et al., 2021). Alternatively, it has been conducted using vignette studies where short descriptions about the AI decision aid are presented to respondents within an online survey to elicit their judgments and perceptions (e.g., Aoki, 2021; Gillath et al., 2021; Höddinghaus et al., 2021). Research using these methods is severely limited in the extent to which they can elicit the interplay between trust and perceived situation risk that is important for explaining people's reliance on/compliance with this technology. Thus, "there is a growing need for research in real-life settings…," such as organizations that are preparing to use, or already using, AI decision aids, "…where using AI is associated with greater personal risk for users" (Glikson & Woolley, 2020, p. 647). Field research in industries characterized as having greater risks and thus higher decision stakes, such as healthcare, manufacturing, energy, human resources, and finance, are prime areas to start.

Building Bridges Across Levels

Our conceptual model focuses primarily on the micro-level processes that explain trust, perceived risk, and reliance on/compliance with an AI decision aid. However, as this relationship is situated in the multilevel work system, a multilevel perspective is needed to fully understand what contributes to trust in AI decision aids and behavioral displays of trust in this context (Hitt, Beamish, Jackson, & Mathieu, 2007). For example, in addition to features of the AI decision aid itself and the dispositional qualities of the person, a number of meso- and macro-level factors should be important for influencing trust in AI decision aids. The communication and training offered by organizations (Hoff & Bashir, 2015) could influence perceptions of an AI decision aid's trustworthiness, or work design could influence a person's attentional capacity and thus their perceived control over an AI decision aid. There are also multilevel factors that contribute to perceived situational risks. Building on an example in Table 1, an account manager's belief that relying on an AI decision aid could have negative implications for their relationship with a client could be influenced by factors related to the individual (e.g., their conscientiousness), perceived expectations of client service behavior in the organization

(e.g., work climate), and the likelihood of being reprimanded by their supervisor if client service ratings drop (e.g., leadership styles and leader-member relationships). It is important that future research on trust in AI decision aids also addresses the larger context in which work with this technology is embedded.

Engaging in Interdisciplinary Research

Glikson and Woolley (2020) encourage researchers studying trust in AI in more technical domains to join forces with organizational researchers when studying the trust in AI in organizational settings. The reverse is also warranted. Organizational researchers studying trust in AI decision aids and other AI-enabled systems would be wise to involve researchers working in more technical fields such as human factors and information systems in their projects, as they are likely to have a more comprehensive overview of the empirical research that has been carried out on the topic of trust in AI or other relevant technology. They are also likely to contribute important considerations relevant for research on the topic. For example, what level of trust in the AI decision aid is appropriate? What human-, technological-, or organizational-factor could contribute to AI decision aid misuse/disuse? What can be done to ensure that trust in the AI decision aid is appropriately calibrated to ensure safe interaction and optimal utilization of the technology's capabilities?

Conclusion

Understanding what facilitates a person's trust in and choice to rely on an AI decision aid is important for the organizations who introduce this technology to improve decision-making and for the employees who are expected to work with it. The aim of the present paper was to specify and elaborate on Mayer et al.'s (1995) model of organizational trust to more adequately address the human–AI decision aid work relationship. Using insight gained from the trust in technology and trust in automation literatures, we redefined central concepts in Mayer et al.'s (1995) model for this context, including trust, perceived trustworthiness, propensity to trust, perceived risk, and risk-taking in the relationship. Further, we expanded the model to include new constructs, such as a perceived control in the AI decision aid, as well as other factors indicated in our review to have implications for trust in and reliance on an AI decision aid. Earlier propositions made by Mayer et al. (1995) were refined to reflect the relationships expected between constructs identified in our model.

The conceptual model developed in this paper presents a framework that can help researchers studying trust in AI in organizational settings develop

their research models, build systematically on each other's research, and contribute to a more cohesive understanding of human–AI work relationships. However, there is still much work to do. We outlined five next steps needed to take research on the topic forward.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Elizabeth Solberg https://orcid.org/0000-0003-3325-3015

Notes

1. For the moment, we ignore that the level of trust a person has for the AI decision aid should correspond to the actual level of performance the aid can deliver. While underutilizing an AI decision aid that performs its task well is a problem, so too is overtrusting the aid if there is a failure to notice a decision error made by the application, or if a faulty recommendation from the application is followed. We return to this point in a later section of the paper.
2. Reliability here refers to performance stability, that good performance is demonstrated over time and across different circumstances, within a reasonable range of contexts. This is different from the reliability associated with integrity in Mayer et al.'s (1995) model, which refers to trusting that a person will actually do what they say they will do, something attributed to the person's disposition and character rather than to their ability.
3. In research on automated technologies, a user is said to be "out of the loop" (OOTL) when they lack the ability to directly control or monitor functions being performed by a technology. It is often discussed in relation to the "out-of-the-loop performance problem" (Endsley & Kiris, 1995), where system operators working with automation have been found to have reduced capability to detect system errors and to perform tasks manually in the face of automation failures, compared with operators who perform the same tasks manually.

References

Alan, A., Costanza, E., Fischer, J., Ramchurn, S., Rodden, T., & Jennings, N. R. (2014, May 5–9). A field study of human-agent interaction for electricity tariff switching.

13th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS), Paris, France.

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*(4), Article 106572. https://doi.org/10.1016/j.chb.2020.106572.

Bahner, J. E., Elepfandt, M. F., & Manzey, D. (2008, September 22–26). Misuse of diagnostic aids in process control: The effects of automation misses on complacency and automation bias. Human Factors and Ergonomics Society 52nd Annual Meeting, Los Angeles, CA. https://doi.org/10.1177/154193120805201906.

Beldad, A. D., & Hegner, S. M. (2018). Expanding the technology acceptance model with the inclusion of trust, social influence, and health valuation to determine the predictors of German users' willingness to continue using a fitness app: A structural equation modeling approach. *International Journal of Human–Computer Interaction*, *34*(9), 882–893. https://doi.org/10.1080/10447318.2017.1403220.

Biros, D. P., Daly, M., & Gunsch, G. (2004). The influence of task load and automation trust on deception detection. *Group Decision and Negotiation*, *13*(2), 173–189. https://doi.org/10.1023/b:grup.0000021840.85686.57.

Blau, P. M. (1968). Social exchange. *International Encyclopedia of the Social Sciences*, *7*(4), 452–457.

Burr, C., Cristianini, N., & Ladyman, J. (2018). An analysis of the interaction between intelligent software agents and human users. *Minds and Machines*, *28*(4), 735–774. https://doi.org/10.1007/s11023-018-9479-0.

Chancey, E. T., Bliss, J. P., Yamani, Y., & Handley, H. A. (2017). Trust and the compliance–reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence. *Human Factors*, *59*(3), 333–345. https://doi.org/10.1177/0018720816682648.

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

Costa, A. C., Fulmer, C. A., & Anderson, N. R. (2018). Trust in work teams: An integrative review, multilevel model, and future directions. *Journal of Organizational Behavior*, *39*(2), 169–184. https://doi.org/10.1002/job.2213.

Cropanzano, R., & Mitchell, M. S. (2005). Social exchange theory: An interdisciplinary review. *Journal of Management*, *31*(6), 874–900. https://doi.org/10.1177/0149206305279602.

Das, T. K., & Teng, B.-S. (2001). Trust, control, and risk in strategic alliances: An integrated framework. *Organization Studies*, *22*(2), 251–283. https://doi.org/10.1177/0170840601222004.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008.

Davis, J. H., Schoorman, F. D., Mayer, R. C., & Tan, H. H. (2000). The trusted general manager and business unit performance: Empirical evidence of a competitive advantage. *Strategic Management Journal*, *21*(5), 563–576. https://doi.org/10.1002/(sici)1097-0266(200005)21:5<563::aid-smj99>3.0.co;2-0.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643.

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors*, *37*(2), 381–394. https://doi.org/10.1518/001872095779064555.

Featherman, M. (2001). Extending the technology acceptance model by inclusion of perceived risk. *AMCIS 2001 Proceedings*, *148*, 758–760.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, *114*(2017), 254–280. https://doi.org/10.1016/j.techfore.2016.08.019.

Ghazizadeh, M., Lee, J. D., & Boyle, L. N. (2012). Extending the technology acceptance model to assess automation. *Cognition, Technology & Work*, *14*(1), 39–49. https://doi.org/10.1007/s10111-011-0194-3.

Gillath, O., Ai, T., Branicky, M. S., Keshmiri, S., Davison, R. B., & Spaulding, R. (2021). Attachment and trust in artificial intelligence. *Computers in Human Behavior*, *115*(52), Article 106607. https://doi.org/10.1016/j.chb.2020.106607.

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

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, *61*(4), 5–14. https://doi.org/10.1177/0008125619864925.

Heggestad, E. D., Scheaf, D. J., Banks, G. C., Monroe Hausfeld, M., Tonidandel, S., & Williams, E. B. (2019). Scale adaptation in organizational science research: A review and best-practice recommendations. *Journal of Management*, *45*(6), 2596–2627. https://doi.org/10.1177/0149206319850280.

Hitt, M. A., Beamish, P. W., Jackson, S. E., & Mathieu, J. E. (2007). Building theoretical and empirical bridges across levels: Multilevel research in management. *Academy of Management Journal*, *50*(6), 1385–1399. https://doi.org/10.5465/amj.2007.28166219.

Ho, G., Pavlovic, N., Myers, V., & Arrabito, R. (2013, September). Reducing false alarms in automated target recognition by lowering the level of automation. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 57, No. 1, pp. 374-378). Sage CA: Los Angeles, CA: SAGE Publications.

Höddinghaus, M., Sondern, D., & Hertel, G. (2021). The automation of leadership functions: Would people trust decision algorithms? *Computers in Human Behavior*, *117*, Article 106635. https://doi.org/10.1016/j.chb.2020.106635.

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

Hösterey, S., & Onnasch, L. (2021, September 01). Manipulating situational risk in human-automation research–Validation of a new experimental paradigm in virtual reality. Human Factors and Ergonomics Society Annual Meeting, Los Angeles, CA.

Im, I., Kim, Y., & Han, H.-J. (2008). The effects of perceived risk and technology type on users' acceptance of technologies. *Information & Management*, *45*(1), 1–9. https://doi.org/10.1016/j.im.2007.03.005.

Jacoby, J., & Kaplan, L. B. (1972). *The components of perceived risk*. Chicago, IL: Third Annual Conference of the Association for Consumer Research.

Jarvenpaa, S. L., Knoll, K., & Leidner, D. E. (1998). Is anybody out there? Antecedents of trust in global virtual teams. *Journal of Management Information Systems*, *14*(4), 29–64. https://doi.org/10.1080/07421222.1998.11518185.

Jones, G. R., & George, J. M. (1998). The experience and evolution of trust: Implications for cooperation and teamwork. *Academy of Management Review*, *23*(3), 531–546. https://doi.org/10.5465/amr.1998.926625.

Kahneman, D. (1973). *Attention and effort* (Vol. *1063*). State College, PA: Citeseer.

Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. (2021). Trust in artificial intelligence: meta-analytic findings. *Human Factors*. Online First https://doi.org/10.1177%2F00187208211013988.

Kohn, S. C., De Visser, E. J., Wiese, E., Lee, Y.-C., & Shaw, T. H. (2021). Measurement of trust in automation: A narrative review and reference guide. *Frontiers in Psychology*, *12*, Article 604977. https://doi.org/10.3389/fpsyg.2021.604977.

Kramer, R. M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual Review of Psychology*, *50*(1), 569–598. https://doi.org/10.1146/annurev.psych.50.1.569.

Ladegard, G., & Gjerde, S. (2014). Leadership coaching, leader role-efficacy, and trust in subordinates. A mixed methods study assessing leadership coaching as a leadership development tool. *The Leadership Quarterly*, *25*(4), 631–646. https://doi.org/10.1016/j.leaqua.2014.02.002.

Langer, M., König, C. J., & Busch, V. (2021). Changing the means of managerial work: Effects of automated decision support systems on personnel selection tasks. *Journal of Business and Psychology*, *36*(5), 751–769. https://doi.org/10.1007/s10869-020-09711-6.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1), 2053951718756684. https://doi.org/10.1177/2053951718756684.

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

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392.

Long, C. P., & Sitkin, S. B. (2018). Control–trust dynamics in organizations: identifying shared perspectives and charting conceptual fault lines. *Academy of Management Annals*, *12*(2), 725–751. https://doi.org/10.5465/annals.2016.0055.

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. 11th Australasian Conference on Information Systems, Brisbane, Australia.

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making*, *6*(1), 57–87. https://doi.org/10.1177/1555343411433844.

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

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.5465/amr.1995.9508080335.e

McAllister, D. J. (1995). Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal*, *38*(1), 24–59. https://doi.org/10.5465/256727.

McCarthy, J. (2007). *What is artificial intelligence?* Stanford, CA: Stanford University. http://www-formal.stanford.edu/jmc/whatisai.pdf.

McKnight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems (TMIS)*, *2*(2), 1–25. https://doi.org/10.1145/1985347.1985353.

Metcalf, L., Askay, D. A., & Rosenberg, L. B. (2019). Keeping humans in the loop: pooling knowledge through artificial swarm intelligence to improve business decision making. *California Management Review*, *61*(4), 84–109. https://doi.org/10.1177/0008125619862256.

Meyer, J. (2001). Effects of warning validity and proximity on responses to warnings. *Human Factors*, *43*(4), 563–572. https://doi.org/10.1518/001872001775870395.

Möhlmann, M., & Zalmanson, L. (2017, December 10–13). Hands on the wheel: Navigating algorithmic management and Uber drivers' autonomy. 38th International Conference on Information Systems (ICIS 2017), Seoul, South Korea.

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

Oksanen, A., Savela, N., Latikka, R., & Koivula, A. (2020). Trust toward robots and artificial intelligence: An experimental approach to human–technology interactions online. *Frontiers in Psychology*, *12*, Article 568256. https://doi.org/10.3389%2Ffpsyg.2020.568256.

O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2020). Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature. *Human Factors*. Online First https://doi.org/0018720820960865.

Pak, R., Fink, N., Price, M., Bass, B., & Sturre, L. (2012). Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics*, *55*(9), 1059–1072. https://doi.org/10.1080/00140139.2012.691554.

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

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *30*(3), 286–297. https://doi.org/10.1109/3468.844354.

Parry, K., Cohen, M., & Bhattacharya, S. (2016). Rise of the machines: A critical consideration of automated leadership decision making in organizations. *Group & Organization Management*, *41*(5), 571–594. https://doi.org/10.1177/1059601116643442.

Pavlou, P. A. (2003). Consumer acceptance of electronic commerce: Integrating trust and risk with the technology acceptance model. *International Journal of Electronic Commerce*, *7*(3), 101–134. https://doi.org/10.1080/10864415.2003.11044275.

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5.

Robinson, S. L. (1996). Trust and breach of the psychological contract. *Administrative Science Quarterly*, *41*(4), 574–599. https://doi.org/10.2307/2393868.

Rotter, JB (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, *80*(1), 1–28. https://doi.org/10.1037/h0092976.

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

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

Schnall, R., Higgins, T., Brown, W., Carballo-Dieguez, A., & Bakken, S. (2015). Trust, perceived risk, perceived ease of use and perceived usefulness as factors related to mHealth technology use. *Studies in Health Technology and Informatics*, *216*, 467–471. https://www.ncbi.nlm.nih.gov/pubmed/26262094.

Schoorman, F. D., & Ballinger, G. (2006). *Leadership, trust and client service in veterinary hospitals*. Unpublished Working paper. New York, NY: Purdue University.

Schoorman, F. D., Mayer, R. C., & Davis, J. H. (1996). Empowerment in veterinary clinics: The role of trust in delegation. *Journal of Trust Research*, *6*(1), 76–90.

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

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

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

Shrestha, Y. R., Ben-Menahem, S. M., & Von Krogh, G. (2019). Organizational decision-making structures in the age of artificial intelligence. *California Management Review*, *61*(4), 66–83. https://doi.org/10.1177/0008125619862257.

Shrestha, Y. R., Krishna, V., & von Krogh, G. (2021). Augmenting organizational decision-making with deep learning algorithms: Principles, promises, and challenges. *Journal of Business Research*, *123*, 588–603. https://doi.org/https://doi.org/10.1016/j.jbusres.2020.09.068.

Stanton, N. A., & Young, M. S. (2000). A proposed psychological model of driving automation. *Theoretical Issues in Ergonomics Science*, *1*(4), 315–331. https://doi.org/10.1080/14639220052399131.

Stuck, R. E., Holthausen, B. E., & Walker, B. N. (2020). The role of risk in human-robot trust. In C. S. Nam & J. B. Lyons (Eds.), *Trust in human-robot interaction* (pp. 179–194). Amsterdam, Netherlands: Elsevier.

Taddeo, M. (2009). Defining trust and E− trust: Old theories and new problems. *International Journal of Technology and Human Interaction*, *5*(2), 23–35. https://doi.org/10.4018/jthi.2009040102.

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

Von Krogh, G. (2018). Artificial intelligence in organizations: New opportunities for phenomenon-based theorizing. *Academy of Management Discoveries*, *4*(4), 404–409. https://doi.org/10.5465/amd.2018.0084.

Wang, H., Sun, L., & Bertino, E. (2014). Building access control policy model for privacy preserving and testing policy conflicting problems. *Journal of Computer and System Sciences*, *80*(8), 1493–1503. https://doi.org/10.1016/j.jcss.2014.04.017.

Wei, J., Bolton, M. L., & Humphrey, L. (2020). The level of measurement of trust in automation. *Theoretical Issues in Ergonomics Science*, *22*(3), 274–295. https://doi.org/10.1080/1463922x.2020.1766596.

Yuviler-Gavish, N, & Gopher, D (2011). Effect of descriptive information and experience on automation reliance. *Human Factors*, *53*(3), 230–244. https://doi.org/10.1177/0018720811406725.

Zuboff, S. (1988). *In the age of the smart machine: The future of work and power*. New York, NY: Basic Books.

Submitted Date: June 25, 2021
Revised Submission Date: January 31, 2022
Acceptance Date: February 1, 2022

Author Biographies

Elizabeth Solberg is a Senior Researcher in the department of Human-Centred Digitalization at the Institute for Energy Technology. Elizabeth's areas of expertise include employee cognition, engagement, and adaptive performance in changing and digitalized workplaces.

Magnhild Kaarstad is a Senior Researcher in the department of Human-Centred Digitalization at the Institute for Energy Technology. Magnhild's areas of expertise include human factors, cognitive psychology, user experience, usability, and teamwork.

Rossella Bisio is a Senior Researcher in the department of Humans and Automation at the Institute for Energy Technology. Rossella's areas of expertise include the application of artificial intelligence for improving human-automation interaction, especially oriented to safety critical industries.

Maren H. Rø Eitrheim is a Researcher in the department of Humans and Automation at the Institute for Energy Technology. Maren's areas of expertise include human factors and cognitive and experimental psychology.

Kine Reegård is a Senior Researcher in the department of Human-Centred Digitalization at the Institute for Energy Technology. Kine's areas of expertise include work systems design, teams, and organizational capabilities.

Marten Bloch is a Researcher in the department of Humans and Automation at the Institute for Energy Technology. Marten's areas of expertise include human machine interaction in safety critical systems, human performance measurements, and human robot interaction.