Handelshøyskolen BI

GRA 19703 Master Thesis

Final Thesis Master of Science 100%

Predefinert informasjon

Startdato:	08-01-2024 09:00 CET
Sluttdato:	01-07-2024 12:00 CEST
Eksamensform:	T
Termin:	202410
Vurderingsform:	Norsk 6-trinns skala (A-F)
Flowkode:	202410||11436||IN00||W||T
External assessor:	External assessor 1
Internal assessor:	Internal assessor 1

Deltaker

Navn:	Synne Eline Loftesnes og Aqsa Waqas

Informasjon fra deltaker

Tittel *:	Managers' reliance on AI decision aids and their perceived trustworthiness
Navn på veileder *:	Elizabeth Anne Solberg

Inneholder besvarelsen konfidensielt materiale?:	Nei	**Kan besvarelsen offentliggjøres?:**	Ja

Gruppe

Gruppenavn:	(Anonymisert)
Gruppenummer:	11
Andre medlemmer i gruppen:	

Managers' reliance on AI decision aids and their perceived trustworthiness

GRA19703 Master Thesis

Supervisor:

Elizabeth Anne Solberg

Associate Professor Department of Leadership and Organizational

Study Program:

Master of Science in Leadership and Organizational Psychology

Acknowledgements

First, we extend our gratitude to our supervisor, Elizabeth Solberg, Associate Professor in the Department of Leadership and Organizational at BI Norwegian Business School. Her support, guidance, and encouragement have been indispensable throughout our journey.

Secondly, we would also like to express our sincere appreciation to Ulf Henning Olsson, Professor in the Department of Social Economics at BI Norwegian Business School. We are profoundly grateful for his invaluable assistance, flexibility, and motivation, which have played a pivotal role in shaping our research endeavor.

Third, we extend our heartfelt thanks to all the participants who generously contributed their time and insights to this study.

Lastly, we want to thank everyone who has supported and encouraged us along the way. Your belief in us has been a constant source of motivation and inspiration.

BI Norwegian Business School, Oslo
01 July 2024

Abstract

With the increasing integration of artificial intelligence (AI) in managerial roles, this study investigates how employees perceive the trustworthiness of managers relying on AI decision aids in personnel decision-making, and how these perceptions are influenced by the extent of explanation provided about the decisions. A quantitative research approach using an experimental vignette survey was employed, with a total of 230 respondents participating. Three sets of hypothetical vignettes were created to capture varying levels of managerial reliance on AI and the explanations provided by managers. The findings suggest that managers who do not rely on AI decision aids are perceived as having higher ability and benevolence compared to those who do. However, no significant results for the relationship between a manager's reliance on AI and perceived integrity was found, indicating that managerial reliance on AI decision aids does not negatively affect perceived integrity. Additionally, the provision of explanations for AI-generated decisions does not significantly impact perceptions of managerial trustworthiness. These results are reviewed and discussed in terms of their theoretical implications, limitations, directions of future research and practical implications.

Table of Contents

1.0 Introduction

1.1 Case identification

In recent years, considerable research has focused on how artificial intelligent (AI) decision aids can enhance the quality and efficiency of managerial decision-making, particularly personnel decision-making processes (Chen & Decary, 2020; Langer et al., 2021; Parent- Rocheleau & Parker, 2022; Parry et al., 2016; Raisch & Krakowski, 2021). The belief that technologies can provide innovative and competitive advantages is driving the organizational push towards transforming the workplace into a more digital environment (Makarius et al., 2020), with a particular emphasis on AI enhancing decision-making processes.

As AI is developing into one of the most impactful trends in the workplace, the majority of research has investigated the factors that make AI trustworthy (Glikson & Woolley, 2020; Höddinghaus et al., 2021; Hoff & Bashir, 2014; Solberg et al., 2022). Trust is unquestionably regarded as a fundamental requirement for effective leadership in classical leadership research (Colquitt et al., 2007; Dirks & Ferrin, 2002; Mayer et al., 1995). Building on this, studies have examined what makes AI trustworthy to managers and other decision-makers, referred to as the first party who use or interact with the output of AI decision aids to make decisions that affect others (Langer & Landers, 2021). Additionally, research on the perceived trustworthiness of AI from the perspective of those affected by its decisions, including the second parties targeted by automated or augmented decisions, is also gaining importance and attention (Langer & Landers, 2021).

However, significant gaps remain in understanding the broader implications of AI decision aids on perceptions of managerial trustworthiness. Specifically, there is an unaddressed area regarding whether the use of AI for personnel decision-making can potentially negatively influence managers' trustworthiness (Langer & Landers, 2021). Furthermore, there is a lack of exploration into the practices essential for maintaining trustworthiness when managers choose to rely on these tools (Glikson & Woolley, 2020; Solberg et al., 2022).

Given the growing interest in studying managers' reliance on AI for decision-making, this thesis investigates how employees perceive the trustworthiness of managers who use AI to make personnel decisions. Specifically, it examines the impact of the manner in which managers explain these AI-driven decisions. The focus is on how employees' perceptions of manager trustworthiness are influenced by the degree to which managers rely on AI decision aids and the extent of explanation provided about the decisions. Based on this, our research questions are as follows:

1. *How does the use of AI as a support aid for personnel decisions impact employee perceptions of managerial trustworthiness?*

2. *To what extent does the provision of explanations for AI-driven personnel decisions influence the perceived trustworthiness of managers?*

1.2 Study plan

To address the research questions, this study has followed Mayer, Davis, and Schoorman's (1995) integrative model of organizational trust, which examines factors contributing to trust in organizational settings. Additionally, the study incorporates theories from AI decision- making literature to contextualize the trust dynamics in AI-reliant management. A vignette methodology is employed, to study this in a controlled yet realistic manner (Bell et al., 2022). Participants have been presented with hypothetical scenarios depicting situations where managers utilize an AI decision aid for training and development decisions. These scenarios are varied in terms of the level of explanation provided for the AI-assisted decision. Furthermore, the thesis also details the analytical strategy employed to evaluate the data and the subsequent results, followed by a discussion of the findings in relation to existing theory. Finally, the study presents its limitations, offers suggestions for future research, and outlines the practical implications of the findings.

1.3 Intended contributions

In undertaking this research, we contribute to the human-AI literature by elucidating the dynamics of trust in the context of managers relying on AI decision aids for personnel decision-making and their provision of data explanation. Existing research often relies on deterministic assumptions about the consequences of technology (Parent-Rocheleau & Parker, 2022; Parry et al, 2016). It is essential to move beyond these assumptions and investigate how decision-makers can be encouraged to critically evaluate AI recommendations, meet ethical and legal obligations, and avoid blindly adhering to AI decisions. Understanding how leaders are perceived when using AI decision aids is closely related to what makes AI trustworthy for employees, which is crucial for effective human-AI collaboration in the workplace. Our study emphasizes the second-party perspective, by examining how employees, as recipients of AI-influenced decisions, perceive manager's trustworthiness, thus providing a more comprehensive understanding of trust dynamics in organizational settings.

Practically, our study offers valuable insights for managers on how to effectively integrate AI into their decision-making processes without compromising their perceived trustworthiness. By emphasizing the importance of combining AI with human judgment and transparent data explanations, we offer strategies to maintain and enhance trust. We aim to offer valuable insights into the complex interplay between trust, AI, and managerial decision-making, thus advancing our understanding of implementing AI in organizational contexts.

Furthermore, while much of the existing research on decision automation and augmentation, and trustworthiness predominantly originates from the field of medical decision-making (Gillespie et al., 2023; Langer & Landers, 2021), our research is not limited to a specific organizational context or industry. Thus, our study contributes to the broader application of AI in organizational settings. However, future research could extend this understanding specifically to human resources, as HR professionals- first-party users of AI decision support systems- face unique challenges in maintaining trustworthiness as AI evolves (Langer & Landers, 2021).

2.0 Literature review

2.1 Artificial intelligence

Artificial intelligence (AI) represents a new generation of highly sophisticated technologies capable of interacting with the environment, aiming to simulate human intelligence (Glikson & Woolley, 2020, p. 627). It is characterized as a technology that mimics human intelligence by gathering and interpreting data to perform cognitive tasks, generating solutions, decisions, instructions, and learning from feedback or new examples to improve and adapt (Enholm et al., 2021, p. 1712; Solberg et al., 2022, p. 188). Consequently, AI can sense, comprehend, act, and learn in complex environments, significantly impacting work and organizational processes (Kolbjørnsrud, 2023). This capability enables human decision-makers to collect and utilize new sets of information effectively (Jarrahi, 2018).

This thesis emphasizes the psychological construct of trust and its antecedents rather than the technological aspects of AI. It specifically focuses on AI decision aids, also known as AI-enabled decision support systems (Shrestha et al., 2021), which are computer programs that use AI to generate decision alternatives or recommend courses of action to achieve specific objectives (Solberg et al., 2022, p. 188). Today, AI decision aids are employed to enhance the quality and efficiency of personnel management decisions (Langer et al., 2021; Parent-Rocheleau & Parker, 2022; Parry et al., 2016). These systems can determine available decision alternatives and suggest optimal options to managers, such as optimal work schedules, task assignments, or employees' training needs (Parent-Rocheleau & Parker, 2022; Raisch & Krakowski, 2021; Solberg et al., n.d.)

The use of AI in managerial decision-making is often referred to as algorithmic management, a control system where algorithms are responsible for making and executing decisions affecting labor, thereby reducing human involvement and oversight in the labor process (Parent-Rocheleau & Parker, 2022). Furthermore, a distinction exists between AI augmentation and automation (Kolbjørnsrud, 2023; Raisch & Krakowski, 2021). According to several researchers, significant challenges are associated with AI, including the potential replacement of human leadership, and the subsequent reduction of human control (Langer et al., 2021;

Nagtegaal, 2021; Parry et al., 2016). In this context, automation refers to the autonomous performance of tasks by AI decision aids, excluding human involvement. In contrast, augmentation involves close collaboration between humans and machines to accomplish tasks (Raisch & Krakowski, 2021), also referred to as “hybrid decision-making” (Nagtegaal, 2021).

Conversely, studies highlight the positive impact of AI decision aids on leadership, suggesting that AI can redefine leadership roles by automating routine tasks, allowing leaders to concentrate on strategic, interpersonal, and value-creating tasks (Höddinghaus et al., 2021; Hoff & Bashir, 2014; Kolbjørnsrud, 2023; Raisch & Krakowski, 2021). Additionally, research indicates that people perceive automated systems as less biased and less likely to discriminate compared to human decision-makers (Bigman et al., 2020; Höddinghaus et al., 2021; Langer & Landers, 2021; Nagtegaal, 2021).

2.2 Trustworthiness

The landscape of workplace dynamics and organizational structures has undergone a profound transformation in recent years, particularly with the integration of technology and artificial intelligence into various aspects of daily operations. Initially, research has primarily focused on understanding interpersonal relationships within organizational settings. However, with the widespread adoption of AI systems in workplaces, the focus has shifted towards exploring the dynamics between individuals utilizing AI for decision-making and those affected by these AI-driven decisions. This evolution underscores the critical importance of trust in the context of AI decision aids within organizations.

2.2.1 Interpersonal trust

Mayer and Normann (2004) differentiates between trust, which is a behavioral intention marked by a willingness to be vulnerable to another party, and trustworthiness, which involves the conditions that lead one party to trust another. The exploration of trustworthiness has garnered significant attention, resulting in the development of various indicators over the years. In 1995, Mayer, Davis, and Schoorman introduced an integral model of trust in organizations, drawing from multiple bodies of literature and diverse disciplines. This model remains one of the most widely used frameworks for studying trust in organizational settings

(Schoorman et al., 2007; Solberg et al., 2022). Trust, as defined by Mayer et al. (1995), is a multifaceted construct encompassing the trustor's general disposition to trust, the perceived trustworthiness of the trustee, and the stakes involved in the interaction. Thus, three components are essential; there must be a trustor to give trust, there must be a trustee to accept trust, and something must be at stake (Hoff & Bashir, 2014). While this definition was initially conceived within the framework of human-human interactions in organizational contexts, it is not restricted to interpersonal relationships alone (Solberg et al., 2022). Additionally, trust is described as a dynamic concept subject to change based on the behavior of the trusted agent (Glikson & Woolley, 2020).

Furthermore, Mayer et al. (1995) identified three general bases of trustworthiness: ability, benevolence, and integrity, which serve as antecedents to trust in interpersonal relationships (Höddinghaus et al., 2021; Hoff & Bashir, 2014; Mayer et al., 1995). Ability pertains to the perceived competency and expertise of the trustee (Colquitt et al., 2007; Mayer et al., 1995), benevolence is defined as the extent to which a trustee is believed to want to do good for the trustor, characterized by loyalty, openness, supportiveness and caring (Colquitt et al., 2007). Lastly, integrity refers to the trustee's adherence to moral and ethical principles, encompassing fairness, justice, consistency and promise fulfillment (Colquitt et al., 2007; Mayer et al., 1995). These factors reflect a consistency that reduces employees' perceived risk in trusting their managers (Whitener et al., 1998).

2.2.2 Human-AI trust

In the realm of AI and technology, the concept of trust has evolved into new dimensions. Glikson and Woolley (2020) emphasize trust as a dominant theme in the literature on human-AI relationships. For studying trust in AI, Lee & Moray (1992) developed a conceptual framework emphasizing three bases of trust; performance-based trust, process-based trust and purpose-based trust. These have been related to Mayer et al.'s (1995) dimensions of perceived trustworthiness (Hoff & Bashir, 2014; Solberg et al., 2022), specifically in the context of human-AI relationships. Performance-based trust parallels the concept of ability, reflecting how well an automated system executes a task. Purpose-based trust is based on the understood purpose of automation, aligning with the dimension of

benevolence, as it relates to the perceived helpfulness of AI decision aids in enhancing job performance (Solberg et al., 2022). Lastly, process-based trust is comparable to integrity-based trust in humans, involving the perception that the AI processes are transparent, dependable and adhere to normative values (Solberg et al., 2022).

To foster process-based trust, it is essential for individuals to understand AI systems and the rationale behind their outputs (Kolbjørnsrud, 2023). Without this understanding, the inability to guide AI decision aids toward desired outcomes can undermine the AI decision aid's trustworthiness (Solberg et al., 2022). Consequently, transparency becomes crucial, referring to the availability and clarity of data and decision-making processes within an AI system (Schoenherr et al., 2023). Similarly, explainability- a shared meaning-making process between an explainer and an explainee- is defined as the understanding of the rationale behind AI's successes or failures, its use of data, and its decision-making processes (Ehsan et al., 2021; Schoenherr et al., 2023). In discussions about AI, the terms explainability and transparency are often used interchangeably (Arrieta et al., 2020; Balasubramaniam et al., 2022; Ehsan et al., 2021; Höddinghaus et al., 2021).

2.3 Explaining decisions made with AI

In the context of guidance related to explaining decisions made by AI decision aids, the Information Commissioner's Office (ICO) and the Alan Turing Institute provide practical advice about how to explain decisions made with AI (ico.org.uk). Six main main types of explanations that are crucial for ensuring transparency and compliance, are identified.

The first is rationale explanation, which involves clarifying the logic and reasoning behind the AI decision aids. This requires a clear presentation of how and why certain decisions are reached by the AI. Additionally, responsibility explanation details the roles and accountabilities of the stakeholders involved in the development and management of the AI decision aid, helping to determine who is ultimately responsible for the decisions made by the AI. Furthermore, data explanation provides comprehensive information about the data used by the AI system, including its sources, quality, and relevance. Fairness explanation

addresses measures taken to ensure that the AI system operates without bias and discrimination. Moreover, safety and performance explanation outline the mechanisms in place to monitor and maximize the AI system's accuracy, reliability, security, and robustness of its decisions and behaviors. Lastly, impact explanation considers the broader effects of the AI decision aid's decisions on individuals and society, highlighting both positive and negative consequences and the strategies employed to manage these impacts.

Additionally, Balasubramaniam et al., (2022) identify four key components of explainability: aspects, context, explainers and addressees. These components involve determining which aspects of a system should be explained, the contexts for explanations, the entities providing the explanations (explainers), and the recipients of these explanations (Balasubramaniam et al., 2022; Chazette et al., 2016). The aspect component focuses on explaining the AI decision aid's purpose, the data used, and roles and capabilities to ensure clarity on when AI makes the actual decision and when it only supports managers in making the decisions. This aligns with rationale, responsibility, and data explanations identified by the Information Commissioner's Office (ICO) and the Alan Turing Institute (ico.org.uk). Context is also crucial, determining the appropriate situations for explanations, such as during the use of AI decision aid, or while building, auditing, or testing it (Balasubramaniam et al., 2022).

Guidance from the European Commission identifies three critical considerations for explaining AI decision aids (commission.europa.eu). Transparency is essential for individuals to understand the decision-making process, which involves clarifying the AI's functionality and the rationale behind its determinations (Arrieta et al., 2020; Ehsan et al, 2021; Schoenherr et al., 2023). Another vital aspect is protecting individual rights by informing individuals of their entitlement to human intervention, allowing them to contest decisions made solely by automated methods. This provision aligns with the fairness explanation (ico.org.uk), ensuring decisions are not arbitrary or unfair without human oversight. Additionally, managers must effectively communicate the implications of automated processing, detailing how AI-driven decisions impact individuals' legal rights, personal circumstances, behavior, and choices to reduce potential negative outcomes.

3.0 Theory and hypotheses

3.1 Managerial reliance on AI and employee perception of managerial trustworthiness

The adoption of AI decision aids by managers can significantly influence how their trustworthiness is perceived by employees. In particular, relying on AI in personnel decision- making may raise concerns about a manager's ability, benevolence and integrity, potentially reducing trust perceptions among subordinates subjected to these decisions.

When managers rely on AI for personnel decision-making, employees may question the manager's ability, as reliance on AI could indicate a lack of personnel competence. Without the belief that a manager possesses the competence or ability to fulfill the managerial role, an employee is unlikely to develop trust in that manager (Whitener et al., 1998, p. 526). The introduction of AI decision aids may be perceived by some as indicative of the manager's inability to independently make sound decisions or exercise judgment effectively. Dependence on AI systems could imply a lack of confidence in the manager's own skills and expertise, thereby diminishing perceptions of their competence (Mayer et al., 1995). Therefore, we hypothesize the following:

H1a: *There will be a negative relationship between managers' reliance on AI decision aids and employees' perceptions of managerial ability.*

Managerial benevolence is another critical facet, as employees may feel that AI-driven decisions are impersonal and do not adequately consider their individual needs (Lee, 2018). The increased reliance on AI decision aids by managers could potentially signify a shift in their priorities, placing greater importance on efficiency and optimization, while potentially neglecting human-centric aspects such as empathy, compassion, and relational trust (Lee & Moray, 1992). This shift in focus could possibly erode employees' perceptions of managerial benevolence. Hence, our hypothesis is as follows:

H1b: *There will be a negative relationship between managers' reliance on AI decision aids and employees' perceptions of managerial benevolence.*

Managerial integrity might also be questioned if employees believe AI-driven decisions are biased or lack transparency. Previous research has highlighted the importance of transparency and comprehensibility in decision-making processes, especially when technological aids like AI are involved (Arrieta et al., 2020; Balasubramaniam et al., 2022; Colquitt & Salam, 2009; Ehsan et al., 2021; Höddinghaus et al., 2021; Schoenherr et al., 2023). The opaque nature of AI decision-making processes may lead to concerns about the fairness, accountability, and transparency of decisions made using AI (Solberg et al., 2020). Given these potential issues, it is reasonable to assume that managers who rely on AI decision aids for personnel decision- making, risk being perceived as having lower integrity by their employees. Thus, we formulate the following hypothesis:

H1c: *There will be a negative relationship between managers' reliance on AI decision aids and employees' perceptions of managerial integrity.*

3.2 The moderating role of data explanation

Prioritizing transparency and providing explanations for AI-generated decisions are essential for building and maintaining trust. Managers' provision of explanation in AI decision-making processes may affect their perceived ability, benevolence, and integrity among their employees (Scoenherr et al., 2023). Specifically, when data explanation is provided, employees can challenge the outcomes of AI decision aids if they believe they are flawed (ico.org.uk). This reduces the likelihood of unknown input data, often referred to as the "black box" phenomenon (Samek et al., 2017), thus enhancing transparency and comprehensibility (Ehsan et al., 2021; Schoenherr et al., 2023). Consequently, this thesis emphasizes the importance of data explanation.

Managers who thoroughly explain AI-generated decisions demonstrate competence by understanding and communicating what data and how it is used behind these decisions (ico.org.uk). This enhances perceptions of their ability to navigate and interpret complex technological processes effectively (Mayer et al., 1995; Whitener et al., 1998). Additionally, employees are more likely to trust managers who show proficiency in understanding and explaining AI systems, as it

instills confidence in their decision-making capabilities (Whitener et al., 1998). Therefore, we hypothesize the following:

H2a: *Managers who provide data explanations for AI-generated decisions will be perceived by employees as having more ability than managers who do not provide such explanations.*

Transparency and explainability in decision-making signal a manager's genuine concern for the well-being and understanding of their employees (Colquitt et al., 2007). By providing clear explanations, managers address employees' need to understand what data was used behind an AI-driven decision and how it was utilized. Knowledge about the actions taken when collecting and preparing the training and test data for the AI model reassures employees that appropriate and responsible choices were made to develop an understandable, fair, and accurate AI decision aid (ico.org.uk). This fosters perceptions of benevolence, as employees feel supported and valued by managers who prioritize transparency and communication. Hence, we hypothesize the following:

H2b: *Managers who provide data explanations for AI-generated decisions will be perceived by employees as more benevolent than managers who do not provide such explanations.*

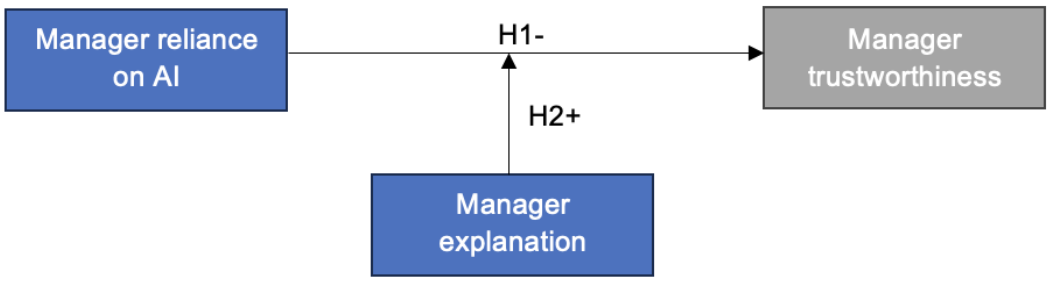
Providing explanations for AI decision aids demonstrates a commitment to ethical and moral principles by promoting fairness, accountability, and transparency in decision-making (Mayer et al., 1995). Managers who provide explanations uphold their integrity by ensuring decisions are made transparently, allowing employees to comprehend the underlying logic and reasoning. This enhances perceptions of integrity, as employees are more likely to trust managers who adhere to principles of fairness and openness in their leadership approach (Colquitt & Salam, 2009). Based on this, we formulate the following hypothesis*:*

H2c: *Managers who provide data explanations for AI-generated decisions will be perceived by employees as having more integrity than managers who do not provide such explanations.*

3.3 Research model

The research model (figure 1) illustrates the relationships to be tested between manager's reliance on AI decision aid, as independent variable, and perceived managerial trustworthiness captured in the facets of ability, benevolence and integrity, as dependent variable (Hypothesis 1a-c). Additionally, manager provision of data explanation is examined as a potential moderator of this relationship, corresponding to Hypothesis 2a-c.

Figure 1: Research model



4.0 Methodology

To ensure the best possible answers to our research question, we have opted for a quantitative research approach using a survey vignette experiment. This approach is particularly well-suited for establishing cause-and-effect relationships and testing hypotheses. Moreover, an experimental vignette design allows researchers to investigate the main effects of each variable and their interaction effect. According to Steiner et al. (2016), a vignette experiment involves presenting a series of vignettes—systematically varied descriptions of subjects, objects, or situations—to understand respondents' beliefs, attitudes, or intended behaviors in response.

4.1 Vignettes developed

For this study, three sets of vignette scenarios were developed (see Appendix 1), each comprising different conditions that depicted interactions between a manager and an employee at a multinational consumer product company utilizing an AI decision aid for training and development decisions. The vignettes were inspired by cases from Gary A. Yukl's *Leadership in Organizations* (2019) and resources available on the website gking.harvard.edu/vign/eg.

The scenario introduction was common to all the conditions. This was followed by three sets of scenarios systematically combining the independent variables, manager's reliance on AI decision aid (yes/no) and manager provision of data explanation (no explanation provided/explanation provided). Each participant was randomly assigned to one of the three conditions to control for order effects and potential bias, leading to a between-subject design (Pannucci & Wilkins, 2010). Table 1 provides a comprehensive overview of the three conditions, with control condition (condition 0) containing no experimental manipulation. For both condition 1 and 2, manager's reliance on AI decision aid is present, however the latter scenario entails increased manipulation of the explanation variable. This was followed by an attention check question where participants were asked to confirm that their answers will be about the manager in the scenario, not their actual manager.

Table 1: Independent variables

Manager reliance on AI decision aid	Manager Explanation	
	No	Yes
Yes	Relies on AI; provides no explanation for decision (Condition 1)	Relies on AI; provides explanation for decision (Condition 2)
No	No reliance on AI decision aid, and no explanation for the decision (Control condition; Condition 0)	

A possible factor that could impact the evaluation of trustworthiness was the gender of the manager in the vignette scenarios. To control for such bias, which could distort measurement and affect investigations and their results (Pannucci & Wilkins, 2010), we put up two sub-conditions for each condition, one for each gender. Participants were then randomly assigned to these.

Ensuring reliability in vignette-based studies, which refers to the consistency and stability of a measure of a concept (Bell et al., 2022), was essential to guarantee that the same vignettes, when presented to different participants, elicit consistent responses. This consistency strengthens the study's overall reliability. Validity, on the other hand, which assesses the extent to which a research instrument measures what it intends to measure (Bell et al., 2022, p. 48-50), can help confirm that the scenarios presented accurately represent the real-world situations or concepts the researcher aims to explore, and it manipulates what it intends to manipulate. The external validity and generalizability of experimental vignette results are often limited because vignette scenarios only provide an approximation of real-life experiences (Höddinghaus et al., 2021). As a result, it is crucial to prioritize the believability of these scenarios (Bell et al., 2022, p.262). In our study, we placed a strong emphasis on presenting vignette scenarios that were as realistic as possible, incorporating relevant information to enhance their credibility.

4.2. Sample and procedure

In this study, we employed non-probability sampling using the snowball effect. To gather primary data, we utilized our social networks by sharing the survey on platforms such as Facebook and LinkedIn. As a result of this sampling method,

certain segments of the population are more likely to be selected than others, as a random process of data collection was not implemented (Bell et al., 2022). Additionally, due to our network consisting of Norwegian and English speakers, we distributed the survey in both languages. The data collection process utilized Qualtrics, with participants being provided with a brief introduction that explained the study's purpose and how their data would be used for academic research. The conditions and sub-conditions were set up in different blocks within Qualtrics and randomly assigned to the participants.

A total of 250 respondents were gathered at the end of the data collection. We removed 6 participants based on them answering “I have no work experience”, and 1 participant because they did not tick off the consent part. We also excluded 13 responses due to missing value and respondents using exceptional long or short time answering the survey. In the end, the final dataset consisted of 230 respondents which provided us with 68 participants in condition 1, 62 in condition 2 and 98 in the control condition. Higher responses in the control condition were due to an initial error in the random allocation process, which was thereafter corrected. At the subcondition level, there were approximately 30 respondents for each gender subcondition in condition 1 and 2. For the control condition, there were 47 respondents in the female sub-condition and 51 respondents in the male sub-condition.

Among the 230 respondents, there was a skewed distribution between genders, with approximately a quarter being men and over three-quarters being women, and one person choosing not to disclose their gender, as shown in Table 2. The age distribution showed that the majority of participants were between 20 and 29 years old, while the rest were evenly distributed across other age groups, including a few under 20 years and one person over 70 years old. Regarding work experience, the largest group of respondents had more than 11 years of experience, followed by those with 3-5 years of experience. A quarter of the participants held managerial positions with personnel responsibility, while three-quarters did not. When asked about their experience with artificial intelligence, responses were almost evenly split between yes and no, with a few respondents unsure.

Table 2: Participant characteristics (in percentages)

	Study (n=230)
Gender	23.9% male; 75.7% female; 0.4% prefer not to say
Age	< 20= 1.7% 20-29= 56.6% 30-39= 17.4% 40-49= 14.8% 50-59= 7.4% 60-69= 1.7% > 70= 0.4%
Work experience	<1= 4.4% 1-2= 13.1% 3-5= 27.5% 6-10= 18.3% 11+ = 36.7%
Managerial position	24% with; 76% without
Experience working with AI	48.9% answering yes; 48.5% answering no, 2.6% answering don't know

4.3 Measures

We aimed to identify well-established items from past research to represent our concepts, ensuring reliability and validity (Bell et al., 2022). In assessing trustworthiness, we drew from the foundational work of Mayer et al. (1999). Our questionnaire items measuring trustworthiness were rooted in this seminal research. However, we omitted two items from each dimension of trustworthiness due to their lack of relevance to our study. Consequently, we included four items for ability, three for benevolence, and four for integrity. To avoid potential misinterpretations and ensure clarity, we removed intensifiers such as "very," "much," "strong," and "hard," which could lead participants to feel that there was insufficient information to answer accurately, particularly in relation to the vignettes (Low, 1996). Additionally, “top management” was replaced with “the manager” to ensure realistic vignette scenarios.

Ability was measured using four items: “The manager is capable of performing its job,” “The manager has knowledge about the work that needs to be done,” “I feel confident about the manager’s skills,” and “The manager is well qualified.” Benevolence was assessed through three items: “The manager is concerned about my welfare,” “My needs and desires are important to the manager,” and “The manager looks out for what is important to me.” Integrity was evaluated with four

items: “The manager has a sense of justice,” “The manager tries to be fair in dealings with others,” “The manager’s actions and behaviors are consistent,” and “Sound principles seem to guide the manager’s behavior.”

We adopted the 5-point Likert scale from Mayer et al. (1999), where participants rated their agreement from 1 (strongly disagree) to 5 (strongly agree), with a neutral option at 3. This adaptation ensured consistency with prior research while effectively capturing participants’ perceptions of trustworthiness.

Participants’ age, gender, work experience, experience working with AI, and managerial status were included as demographic variables in the study. Both experience with AI and managerial status were assessed using binary questions (Yes/ No). Additionally, an attention check question was included, which stated, “Please confirm that your answers to the statements below will be about the manager in the scenario, not your actual manager”, which participants had to agree to in order to proceed.

4.4 Research ethics

Research ethics applies to all aspects of data collection and analysis (Bell et al., 2022). The study is based on the ethical guidelines drawn up by the National Research Ethics Committees. The principles are about respect, good consequences, justice, and integrity (De nasjonale forskningsetiske komiteene, 2019). Additionally, the principle of informed consent and anonymity are highly important (Bell et al., 2022). Respondents in our study were informed in the introduction part of the survey, its use and assurance of anonymity and confidentiality. The provision of information concerning voluntariness, including the option to withdraw from the survey at any point without providing justification, was also evident. Moreover, participants were also provided a debrief, as it is an essential part of the informed consent process (ovpr.uconn.edu), where they got aware of the manipulations made.

4.5 Pilot

Prior to mass distributing the survey we sent out a pilot to 14 different participants, ensuring that the questions and measurements made sense for further analyses. Enclosed with the survey distributed, were questions regarding time

spent on the survey, remarks on difficult or non-understandable questions and alike. The respondents participating in the pre-test were excluded from the final study.

4.6 Analytical strategy

In the process of explaining, describing and analyzing the data collected in the questionnaire, we primarily used the IBM SPSS Statistics. Additionally, Rstudio (2020), integrated development for R, was used for the confirmatory factor analysis (CFA). Also, the PROCESS macro by Hayes (2022) was used for moderation analysis. Prior to conducting the analysis, we prepared the data, ensuring sufficient results in the initial stages before proceeding to test our hypotheses.

Firstly, an exploratory factor analysis (EFA) was conducted to assess the factor structure of the items, a commonly employed approach in this context (Sass, 2010). The EFA was performed using Principal Component Analysis (PCA) with Promax Rotation on the collected data, providing an efficient means of reducing a large set of variables into a smaller set of components that capture the majority of the variance (Jolliffe, 2002). Promax rotation facilitates easier interpretation by allowing for correlated factors. Examination of the factor loadings of each item provided a good indication of how well the items measuring ability, benevolence and integrity represented perceived managerial trustworthiness. To confirm the factor structure and get a more rigorous test of the model fit, we performed a confirmatory factor analysis (CFA) using RStudio (2020).

Subsequently, we conducted a reliability analysis to assess the internal consistency of the items measuring perceived trustworthiness, by applying Cronbach's alpha. For a reliable measure, we used the general rule of thumb to have values in the region of about .70-.80 (Field, 2018). Furthermore, a correlation analysis was conducted using Pearson’s correlation coefficient (Pearson’s r) to explore the relationships between the dependent variables of trustworthiness, and their correlation with the conditions and demographic variables.

To test our hypotheses 1a, 1b and 1c, we used Ordinary Least Squared (OLS) regression analysis. This method was used to determine differences in the dependent variable, manager trustworthiness, between the three vignette experiment groups. OLS estimates the best-fitting linear regression line, allowing researchers to determine the strength, significance, and direction of relationships between variables, thereby understanding causal relationships (Cohen et al., 2013). For Hypotheses 2a, 2b and 2c, an independent-samples t-test was conducted to determine if there were differences in means between condition 1, providing no data explanation for the AI decision and condition 2, providing data explanation.

We used Haye's (2022) PROCESS macro version 4.2 (model 4) to examine whether the provision of data explanations by managers moderates the relationship between managers' reliance on AI decision aids and perceived managerial trustworthiness. A 95% confidence interval was selected. The three dimensions of trustworthiness (perceived manager ability, benevolence, and integrity) were entered as the dependent variables. Manager explanation was set as the moderator variable, where the control condition (no reliance on AI, no explanation provided) and condition 1 (reliance on AI, no explanation provided) were coded as 0, and condition 2 (reliance on AI, explanation provided) was coded as 1. Additionally, reliance on AI was entered as the predictor variable, with the control condition coded as 0 and conditions 1 and 2 coded as 1. In this coding scheme, 0 indicates the absence and 1 indicates the presence of the respective conditions.

5.0 Results

5.1 Descriptive statistics

Table 3 presents the means, standard deviations (SD), and correlations, for the conditions, trustworthiness related variables, as well as the demographic variables in the study. Manager's perceived ability, benevolence and integrity were strongly interrelated. Participants' gender, age, and work experience had some notable correlations with other variables, indicating demographic influences on perceptions, especially the integrity dimension. Additionally, "explanation provided" had a significant negative correlation with perceived benevolence (r= -.21, p<.01). "Reliance on AI" was also found to have a significant negative correlation with perceived benevolence (r=-.22, p<.01), and perceived ability (r=-.14, p<0.5).

Table 3: Descriptive statistics and correlations

Variable	Mean	SD	1	2	3	4	5	6	7	8	9
1. Explanation provided	.28	.45									
2. Reliance on AI	.57	.49	.53**								
3. Ability	3.44	.78	-.09	-.14*	(.88)						
4. Benevolence	3.25	.86	-.21**	-.22**	.65**	(.87)					
5. Integrity	3.30	.78	-.08	-.09	.66**	.70**	(.86)				
6. Gender	1.76	.42	-.06	-.05	-.03	.02	-.03				
7. Age	2.77	1.12	-.01	-.05	-.00	-.01	-.15*	-.28**			
8. Work Experience	3.70	1.21	-.03	-.10	.02	.03	-.13*	-.14*	.70**		
9. Managerial position	1.76	.42	.07	.03	-.10	.04	.02	.37**	-.31**	-.35**	
10. Experience working with AI	1.54	.55	.02	.09	-.06	-.10	-.06	.14*	.08	.08	-.02

Notes. N=230 Cronbach's alpha for each measure is provided in the parantheses.

"Explanation provided" as control condition and condition 1= 0, condition 2 =1, (0=no, 1=yes).

"Reliance on AI" as control condition=0, condition 1 and condition 2=1, (0=no, 1=yes).

p <.05,p <.01.*

5.2 Factor analysis: EFA and CFA

The Principal Component Analysis (PCA) supported the hypothesized factor structure of the scale. Specifically, the analysis revealed three unique factors corresponding to the four items related to perceived manager ability, the three items related to manager benevolence, and the four items linked to manager integrity. Each factor demonstrated strong loadings for their respective items, with loadings ranging from .77 to .91 for ability, .78 to .88 for benevolence, and .82 to .87 for integrity, without any significant cross-loadings on other factors.

The hypothesized CFA model provided a good fit to the data, as evidenced by the $\chi 2(41) = 88.2$, p-value <0.001, RMSEA = .07, CFI = .97, and TLI = .96.

5.3 Reliability analysis

The results demonstrated high levels of reliability across all dimensions, indicating that the items used were consistent in their measurement. The alpha coefficients were .88, .87, .86, for perceived managerial ability, benevolence, and integrity, indicating strong internal consistency.

5.4 Hypothesis testing

5.4.1 Hypothesis 1 - Managerial reliance on AI

In the present study we ran an OLS regression analysis to test the hypotheses predicting negative relationships between managers' reliance on AI decision aid and perceived managerial trustworthiness captured in the facets of ability, benevolence, and integrity,

Hypothesis 1a predicted a negative relationship between a managers' reliance on AI decision aids and employees' perception of the manager's ability. As shown in Table 4, a significant positive relationship was observed between the control condition, where the manager did not rely on AI for decision-making, and perceived manager ability, while a negative relationship was observed between condition 1, where the manager relied on AI but provided no explanation and perceived ability. However, no significant relationship was observed between condition 2, where the manager relied on AI but provided an explanation, and perceived managerial ability. The F-statistic for the model was also insignificant, and the R^2 statistic indicated a low proportion of explained variance. These findings suggest support for hypothesis 1a when no AI explanation is provided for the decision.

Table 4: Results of regression analysis for ability

Manager's reliance on AI	*b*	*Std. Error*
Constant (No reliance on AI, C0)	3.574**	.079
Reliance on AI, no explanation (C1)	-.291*	.126
Reliance on AI, explanation (C2)	-.164	.123
R^2	.025	
F(3, 227)=2.853, p=.060		

Note. N=230. b=unstandardized coefficent
C0=Control conditon; C1= Condition 1; C2=Condition 2
$*p<.05, **p<.01$

Hypothesis 1b proposed a negative relationship between a managers' reliance on AI decision aids and employees' perception of the manager's benevolence. As illustrated in table 5, a significant positive relationship exists between control condition (no reliance on AI) and perceived manager benevolence. Conversely, significant negative relationships were observed for both condition 1(reliance on AI, no explanation provided) and condition 2 (reliance on AI, explanation provided) with perceived managerial benevolence. Altogether, the findings yield support for Hypothesis 1b.

Table 5: Results of regression analysis for benevolence

Manager's reliance on AI	*b*	*Std. Error*
Constant (No reliance on AI, C0)	3.469**	.085
Reliance on AI, no explanation (C1)	-.357**	.136
Reliance on AI, explanation (C2)	-.396**	.134
R^2	.054	
F(3, 227)=5.592 , p=.004		

Note. N=230. b=unstandardized coefficent
C0=Control conditon; C1= Condition 1; C2=Condition 2
$*p<.05, **p<.01$

Hypothesis 1c posited a negative relationship between managers' reliance on AI decision aids and employees' perception of the manager's integrity. As illustrated in Table 6, the control condition was positively related to perceived manager integrity. However, neither condition 1 (reliance on AI, no explanation provided) nor condition 2 (reliance on AI, explanation provided) were significantly related to perceived manager integrity. The F-statistic for the model was also insignificant, and the R^2 statistic indicated a low proportion of explained variance. Thus, Hypothesis 1c is not supported.

Table 6: Results of regression analysis for integrity

Manager's reliance on AI	*b*	*Std. Error*
Constant (No reliance on AI, C0)	3.393**	.079
Reliance on AI, no explanation (C1)	-.106	.126
Reliance on AI, explanation (C2)	-.182	.124
R^2	.010	
F(3, 227)= 1.080 , p=.341		

Note. N=230. b=unstandardized coefficent
C0=Control conditon; C1= Condition 1; C2=Condition 2
$*p<.05, **p<.01$

In the moderation analysis using SPSS with the PROCESS macro (Hayes, 2022) it was tested for whether manager provision of data explanation moderates the relationship between managers' reliance on AI decision aids and perceived managerial trustworthiness (see appendix 2). The results indicate that the effect of managers' reliance on AI on perceived manager ability does not significantly differ based on whether explanations are provided or not (b=.02, SE=.44, Z=.06, p=.94). Similarly, providing explanations does not significantly modify the relationship between managers' reliance on AI decision aids and perceived manager benevolence (b=-.03, SE=.34, Z=-.09, p=.92). Lastly, there is no significant interaction between managers' reliance on AI and explanations provided by the managers in predicting perceived integrity (b=-.01, SE=.96, Z=-.03, p=.96).

5.4.2 Hypothesis 2 - Providing data explanation

To examine the impact of providing explanations for AI decisions on perceived trustworthiness, we tested Hypothesis 2 through independent-samples t-tests, comparing condition 1, where no explanation was provided, and condition 2, where an explanation was provided. These comparisons were made across the dimensions of perceived manager ability, benevolence, and integrity.

Hypothesis 2a proposed that managers who provide data explanations for AI-generated decisions would be perceived as having higher ability. The results indicated no statistically significant difference in the mean scores of perceived manager ability between condition 1 (M = 3.28, SD = 0.83) and condition 2 (M = 3.41, SD = 0.77), t(130) = -0.915, p = .362 (see Table 7). This suggests that providing data explanations does not significantly influence the perception of a manager's perceived ability, thus Hypothesis 2a is not supported.

Hypothesis 2b proposed that managers who provide data explanation for AI-generated decisions would be perceived as more benevolent. As shown in Table 7, the mean perceived benevolence scores for condition 1 (M=3.11, SD=.90) and condition 2 (M=3.07, SD=.87) did not differ significantly, t (130) =2.57, p=.798. Therefore, Hypothesis 2b is not supported.

Lastly, Hypothesis 2c posited that managers who provide data explanations for AI-generated decisions would be perceived as having greater integrity. The analysis presented in Table 7 found no significant differences in the mean scores for perceived manager integrity between condition 1 (M=3.29, SD= .81) and condition 2 (M=3.21, SD=.66), t(df=130) =.586, p=.559, leading to the conclusion that Hypothesis 2c is not supported.

Table 7: Independent-samples T-test

Trustworthiness Dimension	Mean	Std. Deviation	t	p-value	Mean difference	Std. Error difference
Ability						
No Explanation provided (C1)	3.28	0.83	-0.915	0.362	-.13	.14
Explanation Provided (C2)	3.41	0.77				
Benevolence						
No Explanation provided (C1)	3.11	0.90	0.257	0.798	.04	.15
Explanation Provided (C2)	3.07	0.87				
Integrity						
No Explanation provided (C1)	3.29	0.81	0.586	0.559	.08	.13
Explanation Provided (C2)	3.21	0.66				

Note. Equal variances assumed. C1= Condition 1; C2=Condition 2
df=130

5.4.3 Manager's gender

Gender of the managers, set as the sub-conditions, did not have any significant results for the three dimensions of trustworthiness. However, it is worth mentioning that the mean score of perceived manager benevolence for when the manager is referred to as "her" is .418 higher than the mean score of perceived benevolence for when the manager is referred to as "him". For manager integrity, the mean score is .192 higher for manager referred to as "her" than for manager referred to as "him". Conversely, for perceived manager ability, the mean score for the manager being a female is .069 points lower than the mean score for manager referred to as a male.

6.0 General discussion

The main purpose of this study was to investigate how the use of AI as a support aid for personnel decisions impacts employees' perceptions of managerial trustworthiness, and to what extent the provision of data explanation by managers moderates the relationship.

Our findings revealed nuanced insights across the dimensions of trustworthiness. Although Hypothesis 1a was not conclusively supported, our data indicate that managers in the control condition (C0), who did not rely on AI decision aids, are perceived as having higher ability compared to condition 1, where AI is utilized without providing data explanation. Our hypothesis regarding the negative relationship between managers' reliance on AI decision aids and employees' perception of the managerial benevolence (Hypothesis 1b) is supported. The findings reveal that managers who do not rely on AI decision aids are perceived as having higher benevolence compared to those who rely on AI, regardless of whether they provide an explanation. Contrary to our predictions, managerial reliance on AI did not result in any significant differences in perceptions of the manager's integrity (Hypothesis 1c), hence there was no support for the hypothesis.

Our expectations regarding the moderating influence of managers' provision of data explanation on the relationship between managers' reliance on AI decision aids and perceived managerial trustworthiness were not supported. Surprisingly, there was no significant support for Hypothesis 2 (a, b and c), indicating that providing or not providing data explanations for AI-generated decisions does not significantly impact perceptions of manager ability, benevolence or integrity.

6.1 Theoretical implications

6.1.1 Hypothesis 1- Managerial reliance on AI

Our findings, which indicate that managers who do not rely on AI decision aids are perceived as having higher ability compared to those who rely on AI without providing explanation, align with existing research (Dietvorst et al., 2014; Gillespie et al., 2023; Langer & Landers, 2021). Studies suggest that employees who are targeted by automated or augmented decisions and those who observe the

effects of such decisions, ascribe lower ability to managers who use these systems in their decisions (Langer & Landers, 2021). Consequently, managers relying on AI might be seen as overly dependent on technology, thereby diminishing their perceived expertise and decision-making skills.

Moreover, previous research indicates that AI systems are perceived differently from humans even when performing the same roles. For instance, trust was found to be lower when participants imagined receiving medical treatment from an AI rather than humans (Georganta & Ulfert, 2024; Yokoi et al., 2020). Thus, there is a general reluctance to fully accept AI in decision-making roles (Gillespie et al., 2023; Langer & Landers, 2021), with individuals questioning the validity of automated decisions (Mou & Xu, 2017). This reluctance can be attributed to a preference for human forecasts over algorithmic ones and a tendency to judge professionals more harshly for seeking advice from an algorithm rather than from a human (Dietvorst et al., 2014). This supports our findings, suggesting a negative relationship between managers' reliance on AI and perceived managerial ability.

A contribution of our study is support for the research done by Höddinghaus et al. (2021), which found that human agents were assessed as more adaptive and benevolent than automated leadership agents. Our findings reveal that managers who do not rely on AI decision aids are perceived as having higher benevolence compared to those who do, regardless of whether they provide an explanation. This can be attributed to the human capacity to consider qualitative factors in their decisions, while AI-based decisions are often perceived as objective and predictable, potentially disregarding the emotions and interests of employees (Höddinghaus et al., 2021; Parry et al., 2016). For instance, AI systems might fail to account for individual circumstances or the emotional state of employees, leading to perceptions of detachment and lack of empathy and intentionality, thus a dehumanizing experience of AI systems (Lee et al., 2018). Additionally, algorithmic decisions evoke less emotional response than human decisions (Lee et al., 2018), aligning with our findings that leaders who don't rely on AI decision aids are found to be more benevolent, characterized by loyalty and openness (Colquitt et al., 2007).

A further contribution of our study is the support it provides for the notion that AI systems are generally perceived as having high integrity, aligning with the

research by Höddinghaus et al. (2021). Our findings did not indicate a negative relationship between managerial reliance on AI decision aids and employees' perceptions of managers' integrity, nor did the provision of data explanations by managers affect this relationship. AI systems are less prone to discrimination than humans and are impartial in decision-making processes, as they lack personal agency and do not follow personal agendas (Höddinghaus et al., 2021; Lee & See, 2004). Additionally, AI is viewed as less biased than human decision-makers (Bigman et al., 2020; Langer & Landers, 2021; Nagtegaal, 2021), being perceived as objective and more consistent. Thus, the use of AI decision aids does not inherently undermine perceptions of fairness, accountability, and transparency. Consequently, a manager's perceived integrity does not get negatively affected by the use of AI.

6.1.2 Hypothesis 2- Providing data explanation

Our findings regarding the lack of significant support for Hypothesis 2 (a, b and c) contrast with previous research suggesting that clear explanations of how an automated system operates can enhance users' trust and reliance on it (Lee & See, 2004; Schoenherr et al., 2023). Although these findings were unexpected, they still provide valuable insights that are worth pursuing further. The negative mean difference in ability observed in Hypothesis 2a, indicates that participants rated the manager's ability slightly lower when providing data explanation for AI decisions compared to when no explanation was given. This suggests that providing explanations may introduce skepticism or doubt regarding the manager's ability. Participants may perceive the reliance on explanations as an indication of insufficient personal expertise or decision-making skills on the part of the manager (Dietvorst et al., 2014). Additionally, studies have shown that detailed explanations can sometimes lead to skepticism or reduced trust, particularly when individuals are unfamiliar with the technology or when the explanations are overly technical (Binns et al., 2018). Individuals can experience information overload, causing them to lose interest or become confused, which can lead to distrust (Schoenherr et al., 2023; ico.org.uk). This aligns with our findings that perceived ability of the manager was lower when data explanation for the AI decision aid was provided compared to when no explanation was given.

6.2 Limitations and future research suggestions

In line with previous research, the current study also has its weaknesses and limitations. In the following sections, we will discuss limitations of our study and directions for future research.

Firstly, the sampling method may present certain caveats. Response biases could have influenced our findings, as participants may have provided answers they perceived as socially desirable (Bell et al., 2022). Despite our efforts to encourage candid responses, there remains a risk that participants rated the managers as more trustworthy than they genuinely believed. This type of bias is a common concern in self-report studies and can compromise the accuracy of the data (Van de Mortel, 2008), although some researchers argue that it may not be as problematic as anticipated (Ones et al., 1996).

Furthermore, the use of social media channels such as LinkedIn and Facebook for questionnaire distribution might have led to a somewhat homogeneous sample. The generalizability of our findings is further constrained by the small sample size (N=230), with a disproportionate representation of females (75.7%), and higher responses in the control condition (N=98) than the two other conditions. A more diverse and larger sample could yield more robust and generalizable insights (Bell et al., 2022; Pannucci, 2010).

Due to time and resource constraints, the study only examined the immediate effects and did not assess long-term follow-up. The vignettes used, while providing experimental control, do not fully capture the complexities of real-life scenarios, which limits the generalizability and external validity of our findings (Höddinghaus et al., 2021). Future research could employ longitudinal studies, offering valuable insights into the lasting impacts of AI as a support aid for personnel decisions and examine the changes in employee perceptions of manager trustworthiness.

Since this study focused solely on data explanation as the type of explanation provided by managers, future research should further explore the nature and context of different explanations. Examining various types of explanations identified by the Information Commissioner's Office (ICO) and the Alan Turing

Institute could offer a deeper understanding of their impact on employees' perceptions of managerial trustworthiness.

Our findings indicate slight differences in the perception of male and female managers, with a higher mean score for perceived manager benevolence when the manager was female, and a lower mean score for perceived ability when the manager was female. However, these differences were not significant and not part of our predicted relationships. The skewed gender distribution in our sample, with 75% of participants being women, further limits our ability to conclusively determine the reasons for these mean differences. Future research is necessary to explore the underlying mechanisms that perpetuate gender stereotypes within leadership contexts, particularly in the complex interplay between manager's reliance on AI decision aids and their perceived trustworthiness.

Furthermore, future research could examine the relationship between a manager's reliance on AI decision aid and perceived manager trustworthiness, by including control variables such as the manager's education, tenure, and organizational rank. Managers with higher education levels, longer tenure, and more senior ranks are often viewed as more trustworthy due to their potentially greater knowledge, capability, and experience (Lau et al., 2008). Investigating how these factors influence trustworthiness in the context of AI decision aids would yield valuable insights.

Additionally, it would be worthwhile to explore the similarity-attraction paradigm, which posits that subordinates perceive managers as more trustworthy when they share similar demographic attributes (McAllister, 1995). Determining whether this similarity-attraction effect persists when AI decision aids are involved would be a fascinating area of study. Further research could also explore the role of an individual's propensity to trust, as conceptualized in Mayer et al. 's (1995) model, to understand the extent to which this influences the decision to trust a manager relying on AI decision aids. Lastly, it would be valuable to investigate the moderating role of AI complexity, building on the study by Nagtegaal (2021), to determine if managers who rely on high-complexity AI decisions are perceived as less trustworthy. This could replace the focus on data explanation in our study and provide new insights into trust dynamics in AI-reliant management.

6.3 Practical implications

Despite its limitations, this study brings to light practical implications. The insights gained can provide organizations with valuable strategies on how managers can balance personnel decisions made in collaboration with AI while maintaining trustworthiness.

Our findings suggest that managers should balance AI use with visible demonstrations of their own expertise and judgment. Given the general preference for human forecasts over algorithmic ones (Dietvorst et al., 2014), managers should clearly articulate their own role in the decision-making process to reinforce their perceived ability. For example, when using AI decision aid for personalized training recommendations, managers should personally review the information and incorporate insights from their direct interactions with employees. Furthermore, managers should focus on building strong interpersonal relationships with employees and demonstrate empathy and understanding in decision-making, as these are qualities AI cannot replace (Höddinghaus et al., 2021; Lee et al., 2018; Parry et al., 2016). Even when relying on AI, emphasizing human values, and showing care for employees can enhance perceptions of manager benevolence.

Moreover, our findings show that perceived manager ability is lower when providing data explanation compared to when no explanation is given, suggesting that explanations can be technical and lead to more frustration. Therefore, it is important that managers ensure that these explanations are clear, meaningful, and contextually relevant. Additionally, combining explanations with other trust-building activities, such as transparent communication and involving employees in decision-making processes, might be effective (Balasubramaniam et al., 2022; Whitener et al., 1998).

Furthermore, despite the increasing trend in using AI for decision-making tasks (Jaharri, 2018; Hancock et al., 2023), the implementation of AI systems is still limited or non-existent in many workplaces. Thus, organizations should invest in training and development programs to educate employees about the benefits and limitations of AI. Providing opportunities for employees to voice their concerns and involving them in discussions about AI implementation can help build trust

towards managers using AI and foster greater acceptance of AI decision aids in general.

7.0 Conclusion

In the present study, we examined the impact of AI as a support aid for personnel decisions on employee perceptions of managerial trustworthiness, and the extent to which providing explanations for AI-driven decisions influence these perceptions.

We found that managers who do not rely on AI decision aids are perceived as having higher ability and benevolence compared to those who do. Conversely, no significant differences were found in the relationship between manager reliance on AI and perceived integrity, indicating that reliance on AI does not negatively affect perceptions of manager integrity. Additionally, providing or not providing explanations for AI-generated decisions does not significantly impact perceptions of managerial trustworthiness.

The study offers valuable insights into how the dimensions of managerial trustworthiness are affected by the use of AI in decision-making processes. We hope our findings can provide insights for future research and provide practical guidance for managers on effectively integrating AI into their decision-making processes without compromising their perceived trustworthiness.

8.0 References

Alarcon, M. G. & Jessup, A. S. (2023). Propensity to trust and risk aversion: Differential roles in the trust process. *Journal of Research in Personality, 103* https://doi.org/10.1016/j.jrp.2023.104349

Arrieta, A. B., Diaz-Rodriguez, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R. & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82-115. https://doi.org/10.48550/arXiv.1910.10045

Balasubramaniam, N., Kauppinen, M., Hiekkanen, K. & Kujala, S. (2022). Transparency and Explainability of AI Systems: Ethical Guidelines in Practice. *Requirements Engineering*. https://10.1007/978-3-030-98464-9_1

Bell, E., Bryman, A. & Harley, B. (2022). *Business Research Methods* (6th ed.). Oxford University Press.

Bigman, Y., Gray, K., Waytz, A., Arnestad, M., & Wilson, D. (2020). Algorithmic Discrimination Causes Less Moral Outrage than Human Discrimination. https://doi.org/10.31234/osf.io/m3nrp

Binns, R., Van Kleek, Max, Veale, M., Lyngs, U., Zhao, J. & Shadbolt, Nigel. (2018) 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18), https://doi.org/10.1145/3173574.3173951, Available at SSRN: https://ssrn.com/abstract=3114133

Chazette, L., Brunotte, W. & Speith T. (2021). Exploring Explainability: A definition, a Model, and a Knowledge Catalogue. *IEE International Requirements Engineering Conference.* https://doi.org/10.48550/arXiv.2108.03012

Chen, Mei & Decary, Michel (2020). Artificial intelligence in healthcare: An essential guide for health leaders. *Healthcare Management Forum 2020, 33(1), 10-18.* The Canadian College of Health Leaders. https://doi.org/10.1177/0840470419873123

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.

Commision.europa.eu (n.d.). *Are there restrictions on the use of automated decision-making?* Retrieved from commision.europa.eu: https://commission.europa.eu/law/law-topic/data-protection/reform/rules-business-and-organisations/dealing-citizens/are-there-restrictions-use-automated-decision-making_enpa.eu

Cook, J., & Wall, T. (1980). New work attitude measures of trust, organizational commitment, and personal need non-fulfillment. *Journal of Occupational Psychology*, 53, 39-52. https://doi.org/10.1111/j.2044-8325.1980.tb00005.x

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

Colquitt, J. A., & Salam , S. C. (2009). Foster trust through ability, benevolence and integrity. In E. Locke, *Handbook of Principles of Organizational Behavior: Indispensable Knowledge for Evidence-Based Management*, 389-404. https://doi.org/10.1002/9781119206422.ch21

De nasjonale forskningsetiske komiteeene (2019). *Generelle forskningsetiske retningslinjer.* Retrieved from: forskningsetikk.no: https://www.forskningsetikk.no/retningslinjer/generelle/

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114-126. https://doi.org/10.1037/xge0000033

Dirks, K. T., & Ferrin, D. L. (2002). Trust in leadership: Meta-analytic findings and implications for research and practice. *Journal of Applied Psychology, 87*(4), 611-628. https://doi.org/10.1037/0021-9010.87.4.611

Ehsan, U, Liao, Q. V, Muller, M, Riedl, M. O & Weisz, J.D. (2021). Expanding Explainability: Towards Social Transparency in AI systems. *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems,* 1-19. https://doi.org/10.1145/3411764.3445188

Enholm, M. I., Papagiannidis, E., Mikalef. P., Krogstie. J. (2021). Artificial Intelligence and Business Value: A literature review. *Information Systems Frontiers*, 1709-1734. https://doi.org/10.1007/s10796-021-10186-w

Field, A. (2018). *Discovering Statistics Using IBM SPSS Statistics* (5th ed.). Sage Publication.

Georganta, E., & Ulfert, A. S. (2024). Would you trust an AI team member? Team trust in human–AI teams. *Journal of Occupational and Organizational Psychology*. https://doi.org/10.1111/joop.12504

Giffin, K. (1967). The contribution of studies of source credibility to a theory of interpersonal trust in the communication department. *Psychological Bulletin*, *68*(2), 104-120. https://doi.org/10.1037/h0024833

Gillespie, N., Lockey, S., Curtis, C., Pool, J., & Akbari, A. (2023). Trust in Artificial Intelligence: A Global Study. The University of Queensland and KPMG Australia. https://doi.org/10.14264/00d3c94

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

Hancock, B., Schaninger, B., & Yee, L. (2023, 5th of June). *Generative AI and the future of HR* [Audiopodcast]. McKinsey. https://www.mckinsey.com/capabilities/people-and-organizational-performance/our-insights/generative-ai-and-the-future-of-hr

Hayes, A. (2022). The PROCESS macro for SPSS, SAS, and R. https://www.processmacro.org/index.html.

Hoff, K. A., & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors*, *57*(3), 407-434. https://doi.org/10.1177/0018720814547570

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

Ico.org.uk. (n.d.). *What goes into an explanation?* Retrieved from ico.org.uk: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/part-1-the-basics-of-explaining-ai/what-goes-into-an-explanation/

Jarrahi, M. H (2018). Artificial intelligence and the future of work: Human- AI symbiosis in organizational decision making. *Business Horizons,* 577-586. https://doi.org/10.1016/j.bushor.2018.03.007

Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). Springer Series in Statistics. Springer-Verlag. https://doi.org/10.1007/0-387-22440-8_13

Kaplan, A. & Haenlein, M (2020). Rulers of the world, unite! The challenges and opportunities of artificial intelligence. *Business Horizons,* 37-50. https://doi.org/10.1016/j.bushor.2019.09.003

King, E.B., Hebl M.R., Morgan, W. B., and Ahmad, A. S. (2012). Field Experiments on Sensitive Organizational Topics. *Organizational Research Methods*, *1*6(4), 501-521. https://doi.org/10.1177/1094428112462608

Kolbjørnsrud, V. (2023). Designing the Intelligent Organization: SIX PRINCIPLES FOR HUMAN-AI COLLABORATION. *California Management Review*. https://doi.org/10.1177/00081256231211020

Kolbjørnsrud, V., Amico, R., & Thomas, R. J. (2017). Partnering with AI: How organizations can win over skeptical managers. *Strategy and Leadership*. *45*(1), 37-43. https://doi.org/10.1108/SL-12-2016-0085

Kull, T.J., Ellis, S.C., Narasimhan, R. (2013). Reducing Behavioral Constraints to Supplier Integration: A Socio-Technical Systems Perspective. *Journal of Supply Chain Management* https://onlinelibrary.wiley.com/doi/10.1111/jscm.12002

Langer, M., & Landers, N.R. (2021). The future of artificial intelligence at work: A review on effects of decision automation and augmentation on workers targeted by algorithms and third-party observers. *Computers in Human Behavior.* https://doi.org/10.1016/j.chb.2021.106878

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*, 751–769. https://doi.org/10.1007/s10869-020-09711-6

Lau, D. C., Lam, L. W., & Salamon, S. D. (2008). The Impact of Relational Demographics on Perceived Managerial Trustworthiness: Similarity or Norms? *The Journal of Social Psychology*, *148*(2), 187–209. https://doi.org/10.3200/SOCP.148.2.187-209

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

Lee, 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, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society, 5*(1), 1–16. https://doi.org/10.1177/2053951718756684

Low, G. (1996). Intensifiers and Hedges in Questionnaire Items and the Lexical Invisibility Hypothesis, *Applied Linguistics*, *17*(1), 1–37, https://doi.org/10.1093/applin/17.1.1

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

Makarius, E. E., Mukherjee. D, Fox, J. D & Fox, A. K (2020). Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. *Journal of Business Research,* 262-273. https://doi.org/10.1016/j.jbusres.2020.07.045

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *The Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.2307/258792

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., & Norman, P. M., (2004). Exploring Attributes of Trustworthiness: A classroom exercise. *Journal of Management Education*, *28*(2), 224-249 https://doi.org/10.1177/1052562903252641

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267,* 1-38. https://doi.org/10.1016/j.artint.2018.07.007

Mou, Y., & Xu, K (2017). The media inequality: Comparing the initial human-human and human-AI social interactions. *Computers in Human Behavior, 72,* 432-440. http://dx.doi.org/10.1016/j.chb.2017.02.067

Nagtegaal, R (2021). The impact of using algorithms for managerial decisions on public employees' procedural justice. *Government Information Quarterly, 38*(1) https://doi.org/10.1016/j.giq.2020.101536

Nols, T., Ulfert-Blank, A. S., & Parush, A. (2023, June 26-27). *Trust Dispersion and Effective Human-AI Collaboration: The Role of Psychological Safety.* [paper presentation]. Workshops at the Second International Conference on Hybrid Human-Artificial Intelligence (HHAI), Munich, Germany https://www.researchgate.net/publication/373195481_Trust_Dispersion_and_Effective_Human-AI_Collaboration_The_Role_of_Psychological_Safety

Ovpr.uconn.edu. (n.d). *Deception/ Debriefing* Retrieved from ovpr.uconn.edu: https://ovpr.uconn.edu/services/rics/irb/researcher-guide/deception/#:~:text=The%20debriefing%20is%20an%20essential,was%20necessary%20to%20deceive%20them.

Pannucci, C. J., & Wilkins, E. G. (2010). Identifying and avoiding bias in research. *Plastic and reconstructive surgery*, *126*(2), 619–625. https://doi.org/10.1097/PRS.0b013e3181de24bc

Parent- Rocheleau, X., & Parker, S. K. (2022) Algorithms as work designers: How algorithmic management influences the design of jobs. *Human Resource Management Review.* https://doi.org/10.1016/j.hrmr.2021.100838

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

Powell, G.N., Butterfield, A. D., Alves, J. C. & Bartol, K. M. (2004) Sex effects in evaluations of transformational and transactional leaders. *Academy of Management Proceedings*, *2004* (1), https://doi.org/10.5465/ambpp.2004.13863020

Powell, G.N., Butterfield, D.A. & Parent, J.D. (2002). Gender and managerial stereotypes: Have the times changed? *Journal of Management*, *28*(2), 177-193 https://doi.org/10.1016/S0149-2063(01)00136-2

Qiu, J., Kesebir, S., Günaydin, G., Selçuk, E. S. & Wasti, W. (2022). Gender differences in interpersonal trust: Disclosure behavior, benevolence sensitivity and workplace implications. *Organizational Behavior and*

Human Decision Processes, *169* (2022), https://doi.org/10.1016/j.obhdp.2022.104119

Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. *The Academy of Management Review, 46*(1), 192–210. https://doi.org/10.5465/amr.2018.0072

Samek, T., Wiegand, T., & Müller, K. (2017). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. https://doi.org/10.48550/arXiv.1708.08296

Sass, D. A. (2010). Factor Loading Estimation Error and Stability Using Exploratory Factor Analysis. *Educational and Psychological Measurement*, *70*(4), 557-577. https://doi.org/10.1177/0013164409355695

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors*, *58*(3), 377–400. https://doi.org/10.1177/0018720816634228

Schoenherr, J. R., Abbas, R., Michael, K., Rivas, P. & Anderson, T, D. (2023). Designing AI Using a Human-Centered Approach: Explainability and Accuracy Toward Trustworthiness. *IEEE Transactions on technology and society, 4* (1), 9-23. https://doi.org/10.1109/TTS.2023.3257627

Schoorman, F. D., Mayer, R. C., & Davis, J. H. (2007). An Integrative Model of Organizational Trust: Past, Present, and Future. *The Academy of Management Review*, *32*(2), 344–354. http://www.jstor.org/stable/20159304

Solberg, E., Kaarstad, M., Eitrheim, M. H. R., Bisio, R., Reegård, K., & Bloch, M. (2022). A Conceptual Model of Trust, Perceived Risk, and Reliance on AI Decision Aids. *Group & Organization Management*, *47*(2), 187-222. https://doi.org/10.1177/10596011221081238

Solberg, E., Adamska, K., Traavik, L.E.M., & Wong, S. I. (n.d.). *Manager's Digital Mindset and Perceived Risk Using AI Decision Aids.* [Manuscript in preparation]. BI Norwegian Business School

Steiner, Peter. M., Atzmüller, Christiane & Su, Dan (2016). Designing Valid and Reliable Vignette Experiments for Survey Research: A Case Study on the Fair Gender Income Gap. *Journal of Methods and Measurement in the Social Sciences*, *7*(2), 52-94. https://doi.org/10.2458/v7i2.20321

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

Van de Mortel, T. F. (2008). Faking it: Social desirability response bias in self-report research. *Australian Journal of Advanced Nursing*, *25*(4), 40-48. https://search.informit.org/doi/10.3316/informit.210155003844269

Whitener. E.M., Brodt, S.E., Korsgaard, M.A., Werner, J.M. (1998) Managers as Initiators of trust: an exchange relationship framework for understanding managerial trustworthy behavior. *Academy of Management Review*, *23* (3), 513-530. https://doi.org/10.5465/amr.1998.926624

Yokoi, R., Eguchi, Y., Fujita, T., & Nakayachi, K. (2020). Artificial intelligence is trusted less than a doctor in medical treatment decisions: Influence of perceived and value similarity. *International Journal of Human-Computer Interaction,* 1-10. https://doi.org/10.1080/10447318.2020.1861763

Zhang, R., McNeese, N. J., Freeman, G., & Musick, G. (2021). "An ideal human" expectations of AI teammates in human-AI teaming. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3), 1–25. https://doi.org/10.1145/3432945

9.0 Appendices

Appendix 1: Survey

Information letter

Title of the project:
Managers' reliance on AI decision tool and how this reflects on the employees

Purpose of the project:
To investigate how employees perceive the trustworthiness of the managers when they use artificial intelligent (AI) decision aids to make personnel decisions

Who is responsible?
Master students Aqsa Waqas and Synne Eline Loftesnes from the Department of Leadership and Organizational Behavior at BI Business School under the supervision of Elizabeth Solberg.

What does participation require from you?
Participants in this study should be presently employed. The survey will take approximately 3-5 minutes. Your answers will be recorded electronically in the online survey platform Qualtrics. No prior experience with using AI is necessary for you to take the survey. The data will remain confidential and be used for research purposes only.

At the top right, you can choose English or Norwegian as your preferred language to take the survey in.

Participation in the project is voluntary and anonymous. If you choose to participate, you can withdraw your consent at any time. No personal information will be collected that could possibly identify you. Your response is fully anonymous if you choose to participate.

Yours sincerely,

Aqsa Waqas and Synne Eline Loftesnes

By going forward in the survey, you confirm that you are presently employed and give your consent to participate in this study. Please confirm your consent below

[] I give my consent to participate

Thank you for participating in our study! We will start with few general background questions, which has the aim to capture the different backgrounds and experiences represented by the survey respondents. This is so that we can analyze our study data in a good way.

Background questions

How old are you?
Under 20 years old (1), 20-29 years old (2), 30-39 years old (3), 40-49 years old (4), 50-59
years old (5), 60-69 years old (6), 70 years or older (7)

What gender do you identify as?
Male (1), Female (2), Other (3), Prefer not to say (4)

How many years of work experience do you have?
Less than a year (1), 1-2 years (2), 3-5 years (3), 6-10 years (4), 11+years (5), I have no work experience (6)

Are you currently in a managerial position?
Yes (1), No (2)

Do you have experience working with artificial intelligence (i.e., using AI to help execute work tasks)? AI is defined as "Algorithm-based technologies that solve complex tasks by carrying out functions that previously required human thinking."
Yes (1), No (2), I don’t know (3)

Vignette

AI decision aids are computer programs capable of gathering and interpreting large amounts of data to generate decision alternatives or recommend the best course of action. AI decision aids are increasingly available to support managerial decision making, including personnel decisions. For example, they could be used to determine the decision alternatives available and suggest the best options to the manager about who the best job candidate to hire is, how to schedule work most

effectively, or what training modules are best for a particular employee. The manager can take the information received to arrive at a decision.

On the next screen, you will be presented with a short scenario describing a situation where an AI decision aid is available to help a manager make a decision about their employees training and development needs. Please read the scenario thoroughly before answering the questions that follow. There are no right or wrong answers. Just answer as sincerely as you can.

Control condition 0: Manager doesn't rely on AI decision aid

Imagine you work for a multinational consumer product company that is looking to leverage AI solutions to improve the efficiency of different managerial tasks. Recently, the organization acquired a license to use an AI decision aid called DecisionCraft. This tool is designed to assist managers in making different personnel decisions. The first module to be activated supports decisions related to training and development. It suggests personalized training recommendations for each employee based on an analysis of the employee's existing skills, career aspirations, and performance feedback, aiming to find the best training configuration for each employee based on its algorithms.

When you meet with your manager to go through your own training recommendation, she/he clarifies that this is her/ his personal recommendation based on personal knowledge of your skills, career aspirations, and feedback. She/He did not use the *DecisionCraft* tool for this purpose.

Control condition 1: Manager relies on AI decision aid, provides no explanation

Imagine you work for a multinational consumer product company that is looking to leverage AI solutions to improve the efficiency of different managerial tasks. Recently, the organization acquired a license to use an AI decision aid called DecisionCraft. This tool is designed to assist managers in making different personnel decisions. The first module to be activated supports decisions related to training and development. It suggests personalized training recommendations for each employee based on an analysis of the employee's existing skills, career

aspirations, and performance feedback, aiming to find the best training configuration for each employee based on its algorithms.

When you meet with your manager to go through your own training recommendation, she/he clarifies that she/he used the *DecisionCraft* tool, and this was the recommendation it generated based on its analysis of the relevant data. She/He provides no further explanation.

Control condition 2: Manager relies on AI decision aid, provides explanation

Imagine you work for a multinational consumer product company that is looking to leverage AI solutions to improve the efficiency of different managerial tasks. Recently, the organization acquired a license to use an AI decision aid called DecisionCraft. This tool is designed to assist managers in making different personnel decisions. The first module to be activated supports decisions related to training and development. It suggests personalized training recommendations for each employee based on an analysis of the employee's existing skills, career aspirations, and performance feedback, aiming to find the best training configuration for each employee based on its algorithms.

When you meet with your manager to go through your own training recommendation, she/he clarifies that she/he used the *DecisionCraft* tool, and this was the recommendation it generated based on its analysis of the relevant data. She/He provides an explanation for the specific data that was taken into consideration and how it was weighted in the analysis to come to this recommendation.

Survey questions related to the vignette

Please confirm that your answers to the statements below will be about the manager in the scenario, not your actual manager:
( ) Yes, I confirm

Please respond to the statements below about your manager in this scenario, based on what was presented.

The manager is capable of performing her job

Strongly disagree (1), Disagree (2), Neither agree nor disagree (3). Agree (4). Strongly agree (5)

The manager has knowledge about the work that needs to be done

Strongly disagree (1), Disagree (2), Neither agree nor disagree (3). Agree (4). Strongly agree (5)

I feel confident about the manager's skills

Strongly disagree (1), Disagree (2), Neither agree nor disagree (3). Agree (4). Strongly agree (5)

The manager is well qualified

Strongly disagree (1), Disagree (2), Neither agree nor disagree (3). Agree (4). Strongly agree (5)

The manager is concerned about my welfare

Strongly disagree (1), Disagree (2), Neither agree nor disagree (3). Agree (4). Strongly agree (5)

My needs and desires are important to the manager

Strongly disagree (1), Disagree (2), Neither agree nor disagree (3). Agree (4). Strongly agree (5)

The manager looks out for what is important to me

Strongly disagree (1), Disagree (2), Neither agree nor disagree (3). Agree (4). Strongly agree (5)

The manager has a sense of justice

Strongly disagree (1), Disagree (2), Neither agree nor disagree (3). Agree (4). Strongly agree (5)

The manager tries to be fair in dealings with others

Strongly disagree (1), Disagree (2), Neither agree nor disagree (3). Agree (4). Strongly agree (5)

I like the manager's values

Strongly disagree (1), Disagree (2), Neither agree nor disagree (3). Agree (4). Strongly agree (5)

Sound principles seem to guide the manager's behavior

Strongly disagree (1), Disagree (2), Neither agree nor disagree (3). Agree (4). Strongly agree (5)

Debrief to study participants

In this study, we varied the manager's use of the *DecisionCraft* AI decision aid to make a training recommendation (used versus not used). In the conditions where managers were presented as using the *DecisionCraft* AI decision aid, we also varied the explanation provided by the manager for the decision (no explanation versus data explanation). These manipulations were made so that we can test how using AI for personnel decisions could influence perceptions of a manager's trustworthiness.

If you have any questions about the survey, please do not hesitate to contact us at aapiaqsa@gmail.com or synne_v@yahoo.com!

Please proceed to record your response.

Appendix 2: Moderation effect of data explanation on the relationship between reliance on AI and perceived trustworthiness

				CI95%		
Dependent variable (Trustworthiness)		*b*	*Z*	*Lower*	*Upper*	*p*
Ability	Reliance on AI	-.19 (.10)*	-1.90	-.40	.00	.05
	Explanation provided	-.05 (.16)	-.36	-.37	.25	.71
	Reliance on AI x Explanation provided	.02 (.44)	.06	-.83	.89	.94
Benevolence	Reliance on AI	-.25 (.11)*	-2.26	-.48	-.03	.02
	Explanation provided	-.24 (.12)*	-2.01	-.48	-.00	.04
	Reliance on AI x Explanation provided	-.03 (.34)	-.09	-.70	.63	.92
Integrity	Reliance on AI	-.09 (.10)	-.92	-.30	.11	.35
	Explanation provided	-.09 (.11)	-.81	-.31	.13	.41
	Reliance on AI x Explanation provided	-.01 (.96)	-.03	-.63	.61	.96

Notes. N=230. b=unstandardized coefficient; CI95%= confidence interval

"Explanation provided" as control condition and condition 1= 0, condition 2 =1, (0=no, 1=yes).

"Reliance on AI" as control condition=0, condition 1 and condition 2=1, (0=no, 1=yes).

$*p<.05$, $**p<.01$. Standard Errors are in parantheses.