Differentiation of Trust and Compliance in Artificial Intelligence in a working environment – A literature review

Author: Isabelle Hecht
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands

ABSTRACT,
As artificial intelligence (AI) continues to permeate organisational settings, understanding the dynamics of trust and compliance in human-AI interactions becomes increasingly crucial. This paper presents a comprehensive literature review that examines the factors influencing cognitive trust and compliance in the context of AI systems deployed in workplaces. Drawing on cognitive trust theory and social psychology frameworks, the study distinguishes between cognitive trust, which reflects individuals' beliefs in AI systems' reliability and competence, and compliance, which pertains to observable behavioural changes in response to AI directives. Practical insights are provided for enhancing cognitive trust in AI systems, including strategies to improve transparency, reliability, explainability, accuracy, and perceived competence. Furthermore, the paper offers guidance on managing human-AI interaction and addressing ethical considerations in AI design and implementation. The contributions to theory and practice outlined in this paper provide a valuable framework for organisations seeking to leverage AI technologies effectively while fostering employee trust.

Graduation Committee members:
Dr. S.D. Schafheitle, L.C. Lamers Msc

Keywords
Artificial Intelligence, AI, cognitive trust, compliance, authority, reliability, persuasiveness, accuracy, competence, transparency

Table of Contents

1. INTRODUCTION

Integrating Artificial Intelligence (AI) technologies into modern workplaces has brought profound changes, reshaping traditional workflows and redefining human-machine collaboration dynamics. McKinsey (2022) highlights this transformative trend, indicating a notable increase in AI adoption within organisations, from 20% in 2017 to a striking 50% by 2022, which is still ongoing. While AI adoption promises enhanced productivity, efficiency, and decision-making capabilities, it raises crucial questions regarding the future work dynamics and the trust relationship between humans and technology, especially since automated systems still have their faults (Singh et al., 2023).

Recent studies have examined various facets of AI implementation across different domains. For instance, Zhou et al. (2024) investigated the consistency of ChatGPT responses in a medical context, revealing both alignment with clinical standards and cases of deviation and inconsistency. Habbal, Ali, and Abuzaraida (2024) propose the AI Trust, Risk, and Security Management (AI TRiSM) framework, which addresses regulatory compliance, defence against adversarial attacks, skill gap management, and adaptation to evolving threat landscapes in AI integration. Moreover, Economou-Zavlanos et al. (2023) provide a framework for evaluating AI technologies in healthcare, emphasising principles such as clinical value and safety, usability and adoption, fairness and equity, regulatory compliance, and transparency and accountability.

The existing literature on AI adoption within organisational contexts has predominantly focused on technical and organisational aspects, with relatively few studies delving into human dimensions such as trust and compliance till 2021 (Özkizitan & Hassel, 2021). While these studies contribute valuable insights into specific aspects of AI implementation, there remains a gap in understanding employees' and employers' perceptions and utilisation of such changes and the cognitive processes underlying resulting behaviours (Özkizitan & Hassel, 2021).

1.1 Knowledge Gap

It seems that the current focus of research is the connection between trusting the AI and the employer (Weibel et al., 2023), the different versions of trust and how they connect to different forms of AI (Glikson & Woolley, 2020), and future work implications of AI adoption (Haenlein et al., A., 2019). However, the risk of employees just following the orders of AI is not often mentioned, besides missing trust, since it is expected that employees would not follow instructions mindlessly. However, that aspect should be considered, too, given that research shows how inconsistent AI can be (Zhou et al., 2024). Thus, we should differentiate between complying with AI and trusting AI instead of viewing the two as the same thing. For example, Settinger et al. (2024) used the term trust in their study to describe if their participants followed the AI instructions, even though following instructions could simply be compliance. They mixed the two terms, ignoring the cognitive processes leading to the resulting behaviour, which worked in their study, but can be further developed within future research.

It should be noted that, while not clearly differentiated in the business and management literature, there is a clear differentiation between compliance and trust in psychology literature. For instance, Du, Huang and Yang (2019) stated the difference in the context of human-automated teaming by explaining that trust is dependent on reliability, while compliance is more about following systems recommendations. Another difference they stated is that trust is built by clear and comprehensive information, while compliance can occur even if the provided information is misunderstood or incomplete. Another study of the psychology domain that investigated these differences was the study by Hofmann et al. (2017), who also found that a difference between trust and compliance is the influence of coercive power, which decreases trust but enforces compliance.

Noting this, this research aimed to fill this gap, adding to the current literature by making a clear distinction between trust and compliance with AI in a working environment. Through that, a better understanding of the employees' cognitive processes was created, resulting in a more controlled integration of AI as well as a better understanding of the future work with AI and the cognitive processes of employees that follow up with this.

1.2 Research Question

The research question guiding this paper is: “What factors influence individuals’ compliance with and cognitive trust in embedded artificial intelligence systems at work?”. This question aims to uncover the psychological and behavioural determinants driving individuals’ acceptance and utilisation of AI technologies within organisational contexts, highlighting the distinct yet interrelated nature of compliance and cognitive trust. In this context, the interrelated nature means that things influence compliance and trust, which this paper determined too. Notably, this research question represents an original contribution, as it distinguishes compliance from cognitive trust, unlike conventional approaches that often treat compliance as a subset of trust (Glikson & Woolley, 2020).

1.3 Research Objective

This research investigates the factors influencing individuals’ compliance and cognitive trust in embedded artificial intelligence systems. Embedded AI systems mean that the AI is integrated within another system, like the algorithmic one used by Facebook (Glikson & Woolley, 2020). By understanding the origin of the trust relationship between humans and AI and comparing it to reasons for compliance in a working environment, this study provides actionable insights for organisations aiming to understand and foster positive human-AI interactions. Another goal of this Paper is to differentiate better the cognitive factors influencing the human-AI relationship at work and find variables that distinguish cognitive trust and compliance from each other to reduce mistakes when incorporating AI.

1.4 Academic Relevance

This study contributes to the academic literature by addressing a critical gap in understanding human dimensions in AI adoption within organisational settings. By synthesising existing knowledge and identifying underexplored determinants, this research aims to advance scholarly understanding of the trust relationship between humans and AI and distinguish this from compliance with AI since current research does not provide this specific distinction in Business Literature, currently viewing both as the same. Furthermore, the focus on embedded AI systems adds specificity to the research, contributing nuanced insights to the existing body of literature (Glikson & Woolley, 2020; Weibel et al., 2023). The focus on embedded AI was decided based on the fact that emotional attachment is less likely and less beneficial in a work setting, and reliance has a higher relevance (Glikson & Woolley, 2020), which should be the primary focus for AI when integrated into the work field, thus being one of the factors investigated.

2. THEORETICAL FRAMEWORK

This section discusses existing research on trust, compliance, and human-AI interactions. It differentiates between trust and

compliance to provide a nuanced understanding of their roles and implications. Artificial intelligence (AI), trust, compliance, and cognitive trust theory were defined to set the specific framework by which this study operated throughout this paper.

2.1 Artificial Intelligence (AI)

Contemporary artificial intelligence (AI) applications, commonly referred to as 'narrow AI', 'applied AI', or 'weak AI', are understood as specialised systems designed for specific tasks, such as chess games, speech recognition, or image processing, often demonstrating capabilities that rival or surpass human intelligence in their designated domains (Özkizitan & Hassel, 2021). Expanding on this notion, Glikson and Woolley (2020) propose that AI embodies a sophisticated technology that emulates human intelligence, particularly in functions such as reasoning and learning. For this research, the focus was limited to embedded AI without a physical appearance since it is assumed that this will be the most likely AI system to be integrated with a corporate environment.

Anastasi et al. (2021) elaborate on embedded AI technologies, which involve integrating AI into other systems. Examples of this concept include Google Maps, Alexa, or Siri, where AI functionalities seamlessly augment everyday products. Embedded AI can interpret external data accurately, learn from such data, and adapt flexibly to achieve specific goals and tasks (Kaplan & Haenlein, 2019). This integration underscores AI's versatility and its potential impact across diverse domains.

2.2 Trust

Trust is fundamental in human interactions, encompassing beliefs, expectations, and behaviours in various contexts (Moorman et al., 1992). It involves a willingness to rely on others' actions, expecting goodwill and benevolent intentions (Wang et al., 2016). In organisational settings, trust is crucial in facilitating cooperation, collaboration, and effectual functioning (McAllister, 1995). Trust can be differentiated into various forms, including cognitive trust, affective trust, and behavioural trust (Fabrigar et al., 2012). But nowadays, it is mainly divided into either cognitive trust or affectionate trust (Chen et al., 2021).

Cognitive trust refers to the rational aspect of trust based on perceived reliability, competence, and dependability (Moorman et al., 1992). It involves individuals' beliefs in the competence and dependability of others or systems, such as technology (Wang et al., 2016). In artificial intelligence (AI), cognitive trust becomes pertinent, especially concerning complex technologies like embedded AI systems in workplaces (Glikson & Woolley, 2020). Unlike affective trust, which is based on emotional connections and rapport, cognitive trust is rooted in rational assessments of capability and reliability (Moorman et al., 1992). This rational evaluation aligns well with workplace demands, where objective criteria and outcomes often drive decisions.

Considering the focus of this study on the dynamics of human-machine interaction in a working environment, cognitive trust emerged as the most relevant form of trust to be investigated. In a professional setting, where decisions and actions have tangible implications for productivity and outcomes, the rational assessment of trustworthiness becomes paramount (Glikson & Woolley, 2020). Employees' trust in AI systems' reliability, transparency, and competence is crucial for their acceptance and effective utilisation in workplace tasks (Glikson & Woolley, 2020). Therefore, adopting a cognitive trust perspective allowed for a comprehensive understanding of the factors influencing individuals' trust in embedded AI systems at work.

Recent studies have underscored the significance of cognitive trust as a critical determinant of value creation through digital technologies such as AI (Hengstler et al., 2016). It extends beyond human-to-human interactions and encompasses trust in technology, including AI (Wang et al., 2016). However, studies have also highlighted the impact of erroneous AI functions on cognitive trust, indicating that initial trust may decrease over time due to perceived inaccuracies (McKnight et al., 2020).

2.3 Cognitive Trust Theory

This paper utilised the cognitive trust theory as the theoretical framework to examine individuals' trust in embedded AI systems in the workplace. Focusing on cognitive trust, the study uncovered the underlying cognitive processes and perceptual factors that shape employees' trust in embedded AI technology and differentiate trust from compliance regarding the employee-AI working relationship. Drawing on insights from cognitive trust theory, this paper investigated how factors such as **perceived reliability, transparency, and accuracy** (Shamim et al., 2023) influence employees' trust in AI systems. Moreover, the paper explored the dynamics of trust development and maintenance over time, considering the impact of feedback and experience on cognitive trust in AI technology. Through a cognitive trust lens, this paper contributes to a deeper understanding of the human dimensions of AI adoption in organisational contexts.

2.4 Compliance

Compliance can be explained as “changes in behaviour elicited by direct requests”(Baron et al., 2006). Another term describing compliance is “public conformity”, which is defined as “a superficial change in overt behaviour without a corresponding change of opinion that is produced by real or imagined group pressure” (Baron et al., 2006). The latter is more focused on the social connection of compliance, whilst the former is the general term definition. However, both describe the same phenomenon, meaning a change in behaviour upon request without changing on a cognitive level. Thus, compliance is temporary and quickly gained. In social psychology, many different ways of eliciting compliance are described, which can be used in everyday and professional situations (Baron et al., 2006, pp. 286–287). Compliance in human-AI interactions refers explicitly to individuals' behavioural changes in response to requests or directives from AI systems (Fabrigar et al., 2012). Unlike trust, which pertains to individuals' beliefs in AI systems, compliance focuses on observable actions prompted by AI directives. Research has shown that compliance with AI can be influenced by various factors, including the **clarity and persuasiveness of AI directives, individuals' perceptions of AI competence, and the perceived legitimacy of AI authority** (Fabrigar et al., 2012).

2.5 Differentiating Trust and Compliance

While cognitive trust reflects individuals' beliefs in the reliability and competence of AI systems, compliance entails observable behavioural changes in response to AI directives. Trust is a foundation for cooperation and collaboration in human-AI interactions, influencing individuals' willingness to rely on AI systems for decision-making and task execution. It is harder to gain trust than compliance, but trust is more permanent, given the changed mindset. While compliance is often called “mindlessness” (Baron et al., 2006, pp. 286-287), cognitive trust is anything but. It is carefully evaluated and based upon former experiences and knowledge of the functioning and reliability of the AI. Cognitive trust is earned, not just given.

3. METHODOLOGY

3.1 The variables – a research model

During this study, the response of humans in a working environment to embedded artificial intelligence was investigated,

whereby the relationship could either be described as trust or compliance (dependent variables). The specific topic of interest is how those two different responses are triggered, so those influences are treated as independent variables within this paper. For this, the cognitive trust theory, as well as known factors that influence compliance, as stated by Fabrigar et al. (2012), were utilised to determine potential independent variables, which were investigated within the result section. A visualisation of the research model is provided in Figure 1, which can also be viewed in Appendix A.

Factors that are suspected to either cause compliance or trust in/with AI based on the theoretical framework

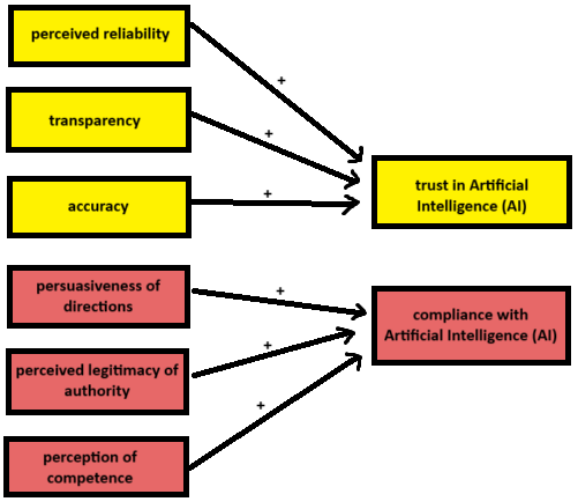


Figure 1: A visualisation of the suspected variable relationship

Figure 1 shows the suspected independent variables on the left side and the dependent variables (which are the two reactions to working with AI that are investigated) on the right side. The relationships suspected are shown through arrows, which are all assumed to be positive. To make the model more understandable, the independent variables are coloured the same as the dependent variable they are suspected to influence.

3.2 Finding Data

A literature review was conducted to investigate the independent variables influencing the dependent variables. By gathering literature from the scientific website Scopus, this Paper used a narrative approach to outline the current state of the field and its complexity, following the approach of Aguinis et al. (2023). This means the paper provides a comprehensive synthesis and critical analysis of existing literature on the topic, highlighting key findings, gaps, and future research directions. Unlike systematic reviews, it offers a more flexible and interpretative approach, allowing for a broader exploration and integration of diverse perspectives and methodologies.

3.2.1 The search terms

Nine different search terms were used during the data gathering process. First, a general search for compliance and cognitive trust was conducted to identify unconsidered variables as well as validate the suspected ones.

For compliance, the general search terms used were “ai AND compliance AND employee”, “ai AND compliance AND work” and “ai AND compliance AND experiment”. Afterwards, the variables “persuasion” and “authority” were investigated specifically since there wasn’t enough evidence in the already found data to validate them. Therefore, the search terms “persuasion AND ai”, “persuasiveness AND ai” and “ai AND authority” were used. Put together, this resulted in a total of 18 sources. The specific filters used during the search can be found in Appendix C.

For cognitive trust, the general search term “cognitive AND trust AND ai” was used. Afterwards, the variable reliability was further investigated through the search term “reliability AND ai AND work.” Together, these search terms provided 11 sources usable for this paper. Again, specific filters can be found in Appendix C.

3.3 Data Selection

A specific selection for this paper was made using different criteria such as time, relevance, correlation, and contribution to this paper. The sources found were analysed more closely, and a few were deemed unusable after closer inspection. The factors considered for that are stated in the following part.

3.3.1 Time and Relevance

Given that artificial intelligence is a developing field that has grown significantly in recent years, the focus was set on recent findings (2022-2024), but older sources (published between 2018-2022) that have high relevance to the topic were also viewed as acceptable, in case of discovery through a backward search, which happened with the variable reliability. The study of Felzmann et al. (2019) was discovered during a backward search and evaluated for inclusion, even though it was not part of the set focus time for publishing, given its significant contribution to this paper’s topic.

3.3.2 Correlation and Contribution

Besides that, the findings were evaluated based on the definitions since there are different ways to define trust, compliance, and artificial intelligence. The latter was investigated carefully since the different types of AI used during a study are not always stated clearly in the beginning, and the different forms of AI might result in different dynamics. The relationship between a human and an AI with a physical appearance (like robots or visualisations) might be completely different from that of a human and an embedded AI. However, the definitions for trust and compliance vary, too, so it was essential to check all three before including a source.

Lastly, the contribution value of the identified papers regarding this paper was also considered. The researcher has read the findings and sorted through the value they might add to the paper (see Analysis-Evaluation Protocol, Appendix D). This paper was only interested in research that included information regarding what determines compliance with AI or cognitive trust in AI. This does not mean that the findings always had to analyse that specific topic, but that they included some kind of information usable to answer the research question.

3.4 Final data distribution

After selecting the articles found, it was necessary to sort them to the correct variables since some sources found when searching for compliance were actually about cognitive trust and thus used in the trust section. After filtering and organising the sources, the final distribution turned out to be 9 sources for compliance and 12 sources for cognitive trust.

3.5 Data analysis

After the data had been determined, it was sorted through the model. First, Compliance and Trust were looked at separately. In that part, their possible predictors, so the independent variables suspected and additionally found influences, were considered.

Afterwards, the findings were combined to distinguish between the two, and the variables were re-evaluated.

3.6 Transparency of the study

To further increase transparency, a search log was added to this paper, outlining the specific search terms and websites used and to which sources they led. The log also includes a detailed description of each source, outlining its specific variables, findings, and importance for this paper. The search log was divided into two categories: Search Protocol (Appendix C) and Analysis-Evaluation Protocol (Appendix D). The Search Protocol shows how the sources were found precisely, while the Analysis-Evaluation Protocol details the different Papers used and what they are about.

4. RESULTS

In this section, compliance and trust were analysed separately to assess the usability of the model and its provided variables. Additionally, variables influencing the relationship between employees using AI and them either complying with or trusting it were discovered. After that, the findings for the different variables were compared in the last subpart to reevaluate the existing model and its usability. Overlaps were considered, and new variables were integrated.

4.1 Compliance

As mentioned in the introduction, three different variables were assumed to be independent variables that positively influence the compliance behaviour of employees, as provided by Baron et al. (2006). Namely, those variables are persuasiveness of AI directions, perceived legitimacy of AI authority and perception of AI competence. Besides analysing these assumed variables, additional ones mentioned in the reviewed articles were also looked into, providing an even better understanding of compliance behaviour with AI.

4.1.1 Persuasiveness of AI Directions

When looking at the variable persuasiveness, multiple studies were found that indeed proved the significance of this variable. Sharabati et al. (2024), for instance, proved the actual power of AI persuasiveness while also stating the risks of it, like AI bias. In their study, they stated that the clarity of AI commands has a high impact on perceived persuasiveness and that persuasiveness indeed results in higher compliance. They also discovered the factor of organisational compliance, which was further evaluated within the part of other influential factors.

Since Persuasiveness was already proven as an independent variable for compliance in interhuman relationships (Baron et al., 2006), it is now important to compare if AI has the same power regarding persuasiveness as humans. For that, a study by Huang and Wang (2023) was conducted. They studied the relative effectiveness of AI compared to humans concerning persuasion. Their findings show that both humans and AI have the same persuasive power since people seem to respond to AI as if they were human beings as well. When closer investigation was conducted, though, they found that AI was weaker in shaping behavioural intentions due to algorithmic aversion. On the other hand, it seemed equally successful in shaping elicit attitudes, perceptions, and actual behaviours. They mentioned that AI's persuasive effectiveness was related to its role and the context of communication.

Since the variable of persuasiveness was found to be true, it was also important to look into how the actual persuasion happens. Two studies were identified for that matter that discussed this topic concerning AI, one by Matz et al. (2024) and one by Zhu et al. (2022). One thing that Zhu et al. (2022) found is that reciprocity plays a significant role in AI persuasiveness. Their study suggests that individuals who perceive benefits and kindness from AI systems are more likely to comply with its directions. This aligns with the psychological principle that gratitude can foster reciprocal behaviours, translating to compliance with AI recommendations (Zhu et al., 2022). Another thing found was that personalisation plays a role in the persuasive power of AI. Matz et al. (2024) found that personalising the way of communication and making it more adapted to the user increased the users' willingness to comply with its commands. They showed that through a study about consumer behaviour, explaining how large language models like ChatGPT can personalise ads, even if the AI had minimal information about the targets. Even though this finding was in the context of consumer behaviour, it still gives us a better understanding of AI persuasion tactics, which could also be used in a working relationship, especially when considering the amount of data about the employees that the AI has access to. Putting the findings of these two studies together, we can conclude that the persuasiveness of AI highly depends on the way of communication, meaning that the friendliness, the perceived benefits and the adaptability of AI to the individual's character are highly significant.

Lastly, it is important to mention the ethical considerations of AI persuasiveness. Klenk (2024) emphasizes the need for responsible design and use of AI to avoid manipulation. While AI can effectively influence decisions through personalised messaging, frameworks that ensure transparency, accountability, and ethical use are needed to prevent misuse. This is particularly important as AI's persuasive power grows with technological advancements and more sophisticated personalisation techniques.

Putting all this together, AI's persuasive capabilities profoundly influence compliance, leveraging non-rational methods to subtly guide behaviour and decision-making, which means that ethical considerations must be addressed to prevent misuse. As AI technology advances, its potential for personalized persuasion is expected to grow, offering significant opportunities for enhancing compliance across various domains. This multifaceted influence underscores the importance of carefully integrating AI into decision-making processes to augment rather than replace human judgment, ensuring that AI remains a supportive tool rather than an authoritative decision-maker.

4.1.2 Perceived Legitimacy of AI Authority

While studies about the legitimacy of authority are still limited in the context of Artificial Intelligence, there was still a study that hinted at it, namely the study of Agudo et al. (2024). They found that people tend to follow AI's directions when given directly. This shows that individuals generally perceive AI as having authority when providing clear and direct instructions. The study further suggests that people are usually inclined to comply with AI, which factors like perceived legitimacy and competence can influence. However, they also found that the timing of the AI suggestion plays a significant role (before or after their own decision). This could be explained through the assumed role of AI that is caused by the timing since the AI receives higher authority by directly stating what to do instead of giving suggestions after the individual has thought about it. Besides that, they emphasise the importance of critically analysing interactions between AI and humans in decision-making processes. Agudo et al. (2024) propose that AI systems should enhance rather than replace human judgment, reinforcing the idea that AI is a supportive tool rather than an authoritative decision-maker. This perspective aligns with the broader goal of

integrating AI to enhance human capabilities and promote organisational collaboration.

Putting these findings together, it can be concluded that there is not enough evidence of the independent variable authority. Further empirical research is required to determine if the AI's legitimacy of authority has a significant effect and how it influences compliance.

4.1.3 Perception of AI Competence

When looking at the perception of competence, three studies were identified that validated the independent variable, proving that it influences compliance behaviour with AI. The first one being by Choudhury et al. (2024), who found strong evidence that the perceived competence of AI, such as ChatGPT, significantly aids decision-making and influences compliance. If the users think the AI is competent, they are more likely to follow its instructions. Furthermore, Zhu et al. (2023) found that the perceived operational capabilities of AI positively affect compliance and the practical attitudes of employees towards AI. This finding suggests that employees are more likely to follow AI directives when they recognise AI systems' technical proficiency and reliability. In addition, Zhu et al. (2022) highlight that investing in AI and improving employees' recognition of AI's cognitive capabilities can enhance thriving at work and compliance behaviour. Organisations can increase employees' confidence and willingness to engage with AI tools by fostering a better understanding of AI's potential. Lastly, it was found that satisfaction with AI outcomes significantly influences compliance. Agudo et al. (2024) found that satisfying AI results lead to higher levels of compliance, whereas unsatisfying results diminish compliance. This indicates that the effectiveness of AI in producing favourable outcomes is crucial for maintaining compliance and adherence to AI recommendations.

4.1.4 Other Influencing Factors

After looking into the different assumed factors, this section relays the additional factors found within the analysed articles that should be considered. The first one being **reward mechanisms**. Reward mechanisms have been shown to positively impact compliance by enhancing employees' self-esteem and reducing anxiety. This motivational strategy can foster a more receptive attitude towards AI directives, leading to higher adherence (Zhu et al., 2023). The same study showed that punishments were also effective, but far less so since they resulted in undesirable consequences such as lower self-esteem and heightened anxiety among employees. So, the study's findings suggest focusing on positive enhancement instead of negative ones, which aligns with the article findings of Zhu et al. (2022), who suggested the kindness of AI and perceived benefits as factors influencing compliance. While Zhu et al. (2022) talked about persuasiveness, they used that to explain compliance, so the article also fits here.

Another variable found was **personality traits**. Personality traits such as conscientiousness play a crucial role. Individuals with high conscientiousness are more likely to comply with AI systems due to their inherent tendency to follow rules and fulfil responsibilities diligently (Zhu et al., 2022). Matz et al. (2024) also found evidence for this in their study. They specifically looked into the possibility of using the "Big 5" personality traits for AI persuasion, showing that matching the user's personality correctly can indeed increase compliance. Extrovertism and Openness proved to be sufficient factors, but they also stated that others might be working as well since their use of social media limited the AI's perception of users' personality traits (Agreeableness is not as easily detected through social media) (Matz et al., 2024). That means that if the AI can access even more detailed information about the individual, which could be through an employee folder and other gathered data about the individual at the workplace, it can further influence the individual's compliance.

The third critical factor found is the **perceived accountability** of AI systems. Compliance increases when users know that AI systems are accountable for their actions and decisions. Knowing that there are mechanisms to ensure AI accountability can reassure users and enhance their willingness to follow AI recommendations (Novelli et al., 2023). The article mentioned here that perceived reliability is also heightened through the assurance of accountability, even though this definition matches more with this paper's definition of perceived competence. Another thing affected is the perception of authority, which also heightens if the AI can be held accountable.

Agudo et al. (2024) additionally add the variable **timing of AI suggestion**. They found that human judgment can be influenced depending on the time the AI suggestion is received. The study participants seemed more inclined to follow AI directions if received directly instead of after making their own judgment of the situation. If they made their judgment beforehand, they questioned the AI more critically, instead of complying with its suggestions (Agudo et al., 2024).

The fifth and last factor found to influence compliance is the **work culture**. Sharabati et al. (2024) found evidence of work culture's effect while investigating AI bias. They concluded from their empirical study that fostering an inclusive and ethic-focused culture reduces AI bias since employees have the confidence to question the AI commands. This means, in conclusion, that an ethic-focused and inclusive work culture reduces compliance with AI but enhances a trust relationship with it.

4.2 Cognitive trust

After investigating the different variables influencing compliance, this part is about the different independent variables influencing cognitive trust. The cognitive trust theory will be the lens for this part, but additional variables mentioned in the literature that could influence the employee's cognitive trust in AI are also stated, similar as executed in the compliance section.

4.2.1 Perceived reliability

Perceived reliability denotes employees' confidence in the consistency and dependability of AI systems, making it a very important variable in the human-AI relationship. An empirical study by Shamin et al. (2023) shows the positive as well as significant correlation between cognitive trust and reliability at work. The article proves that the variable is, as suspected, a valid factor and should be used in further models.

Two articles were identified that further explored the perceived reliability, namely the studies of Tejeda et al. (2022) and Shamin et al. (2023). The first paper that was considered was that by Tejeda et al. (2022). They explored how fluctuations in AI performance impact human confidence in Artificial Intelligence and their own abilities during their paper. Their findings revealed that subpar AI performance notably diminishes both human confidence in AI and their self-assurance. Moreover, confidence in AI takes longer to restore following poor performance compared to the swiftness with which it dissipates, underscoring an asymmetrical effect attributed to loss aversion (Tejeda et al., 2022). That means that users have more difficulty trusting if the AI is unreliable or makes a mistake initially. The study of Shamin et al. (2023) was utilised to examine this more closely. They found that factors such as error rates, visibility of errors, task appropriateness, communication cues, privacy protocols, and the reputation of

technology developers significantly influence the trust employees place in AI decision aids (Shamim et al., 2023). If the user can see the reliability score, he/she will choose more carefully if he/she will trust the results/commands given by the AI.

4.2.2 Transparency

Another suspected independent variable of the model that needed investigation was the variable transparency. For this variable, the paper of Shamin et al. (2023) was also utilised since they tested the usability of this variable as a factor to enhance cognitive trust through their empirical data too. Transparency was the variable they proved to be most significant during their study. Theis et al. (2023) added to this by noting that explaining AI results to foster trust is crucial, as users often need information about the results or behaviour of AI systems. The reason for that when looking at non-expert users primarily is to understand the decisions made by AI, specifically which factors the AI consider before making the decision/providing the solution (Theis et al., 2023). Unlike algorithms, human judgment offers a level of transparency and accountability due to their fallibility (Gravett, 2023), which makes it harder to trust an AI, given that algorithms often do not provide these insights. These findings suggest the importance of considering the benefits caused by openness about AI, specifically its failures and limitations. Users feel more comfortable trusting an AI with a sufficient reliability score, which they can check anytime (Stettinger et al., 2024).

When examining Transparency more closely, the study of Felzmann et al. (2019) should also be mentioned. They offer a comprehensive analysis of transparency in AI, highlighting its multifaceted nature and the critical relations between transparency, informed consent, and individual autonomy. They argue for a relational approach to transparency that acknowledges its role as a signal of trustworthiness and willingness to be accountable to those affected by AI systems.

Putting all this together, we can argue that Transparency offers multiple advantages for the user, like a better understanding of the AI outputs/decisions/results as well as feeling more comfortable with the usage of AI. All this adds positively to the cognitive trust experienced concerning AI.

4.2.3 Accuracy

When analysing this variable, it became clear that it overlaps immensely with reliability. Again, Shamim et al. (2023) can be utilised for this variable since they also analysed it in relation to cognitive trust in AI. They used accuracy to determine reliability, but it still proved the importance of accuracy, even if integrated into reliability. Elder et al. (2022) also combine accuracy and reliability, showing that the accuracy of AI outputs determines the willingness to further rely on it in the future, especially in high-risk scenarios. They argue that accuracy is part of reliability, and both should be considered as one. While Tursunalieva et al. (2024) highlight the importance of accuracy by showing its importance in relation to the decision-making process, they mainly focus on the usability of AI instead of the development of cognitive trust, so utilising their article only provides the knowledge of the variable's importance, but not its importance in regards to the model.

Putting this together, the variables' accuracy and reliability must be combined, as they are too similar to be separated by the model. Accuracy should be considered part of the reliability variable, as stated by the different sources, but it should not be forgotten.

4.2.4 Other factors

The first additional variable found was **explainability**. Sovrano and Vital (2023) emphasised the importance of explainability in AI systems, providing a system for enhancing explainability (Sovrano & Vital, 2023). They argue that understanding the process the AI uses in order to determine its conclusions is crucial. While this can be seen as part of transparency, it could also be argued that it is indeed more since explainability is not just about reviewing the process but also getting an explanation for it. Stettinger et al. (2024) highlighted something similar by stating consistency as an important factor since it makes the AI more predictable in the eyes of the users. This can also be seen as a form of explainability since it serves the purpose of understanding the decision-making process of AI.

Another variable that should be added is the variable of **organisational altruism**. This variable was found through an article about compliance. Zhu et al. (2022) broadened the understanding of compliance behaviour in their article and brought cognitive factors into the compliance variable through their description of compliance. This makes it more fitting to the definition of trust, which is why the factor is considered in this part instead of compliance. They added the variables intrinsic motivation and organisational altruism when analysing the willingness to follow AI directions, both of which had a positive effect. While they called this result increased compliance, it will be viewed here as increased cognitive trust within this paper since they did include intrinsic motivation to follow the AI directives, which is part of cognitive trust as per this paper's definition. Utilising these findings, the new variable "organisational altruism" will be added.

Furthermore, the variable **experience** was discovered. Solberg et al. (2022) showed in their study that trust evolves over time when using AI, like in human relationships. Through positive experiences with Artificial Intelligence, trust grows, and a positive perception of AI is created.

Additionally, the variable **work culture** can be added. As mentioned in the compliance section, Sharabati et al. (2024) found evidence for the effect of work culture while investigating AI bias and concluded that fostering an inclusive and ethic-focused culture reduces AI bias since employees are confident to question the AI commands, which enhances cognitive trust.

Lastly, the variable **personality traits** was considered to play a role here as well. Even though it was not stated in any of the selected literature, it stands to reason that if they influence compliance, as proven by Matz et al. (2024) and Zhu et al. (2022), there is a big chance they will influence cognitive trust as well. So, for now, it is also assumed to be an independent variable for cognitive trust. Future research should look further into this, determining how exactly personality traits play a role in employees' interaction with AI.

4.3 Cognitive Trust Vs Compliance

After investigating the factors influencing cognitive trust and compliance separately, this section compares the two, identifying similarities and distinctions between the dependent variables and determining the final independent variables for both.

4.3.1 Factors Affecting Both Compliance and Cognitive Trust

Some factors influence both compliance and cognitive trust. Reliability, accuracy and perceived competence were very similar and can thus be seen as one. Choudhury et al. (2024) and Zhu et al. (2023) highlight that recognising AI's competence increases compliance and builds trust. Transparency also plays a dual role. Shamim et al. (2023) and Theis et al. (2023) note that transparency enhances compliance and cognitive trust by ensuring users are informed about AI operations. Work culture also influences both compliance and cognitive trust, as

highlighted by Sharabati et al. (2024). Lastly, the personality traits of the Artificial Intelligence users, whose influence was discovered by Matz et al. (2024) and Zhu et al. (2022), are also considered to influence both, given that they determine how an individual reacts to an AI.

4.3.2 Factors Specific to Compliance

Certain factors are uniquely significant for compliance. Persuasiveness, the ability of AI to craft persuasive messages and influence decisions, is particularly impactful. Matz et al. (2024) demonstrate that persuasive AI messages effectively guide user behaviour. Additionally, reward mechanisms are crucial since they can enhance self-esteem and reduce anxiety while working with AI, in addition to promoting adherence to AI directives (Zhu et al., 2022). Another factor uniquely mentioned as a determent of compliance is the AI suggestion's timing, as Agudo et al. (2024) discovered. Lastly, the perceived accountability of Artificial Intelligence was uniquely mentioned, stating that if the AI can be held accountable, compliance would increase (Novelli et al., 2023).

4.3.3 Factors Specific to Cognitive Trust

In contrast, some factors are uniquely significant for cognitive trust. Explainability, for instance, is essential for building trust in AI systems. Sovrano & Vital (2023) and Stettinger et al. (2024) stress the need for AI to be explainable and consistent to foster cognitive trust. Experience is also uniquely mentioned for cognitive trust since it takes time to build trust in AI like it does between humans (Solberg et al., 2022). Lastly, organisational altruism is seen as factor specifically for cognitive trust, since the individual wants the company to be as successful as possible, thus trying to cooperate with the Artificial Intelligence in the most beneficial manner possible (Zhu et al., 2022).

4.4 The new Model

After analysing the factors influencing compliance and cognitive trust, the model needed improvement. The independent variables influencing the dependent variable, cognitive trust, are now assumed to be explainability, organisational altruism and experience. The independent variables influencing compliance are now considered persuasiveness, perceived accountability, time of AI recommendation and reward mechanisms. The independent variables influencing both cognitive trust and compliance are now considered to be reliability, work culture, personality traits and transparency. The new model can be seen in Figure 2, which is also displayed in a larger format in Appendix B.

Figure 2, as already stated, displays the relationship between the discovered/validated independent variables (left side) and the dependent variables (right side). All variables were sorted by colour again to improve understandability. Yellow was used for all independent variables that only influence cognitive trust, and red was used for the variables that only influence compliance, similar to Figure 1. But this time, the category of the independent variables influencing both was added, which received the colour orange. Almost all relationships are positive, except for work culture, personality traits and time of AI recommendations, since those depend on the circumstances. For work culture, it depends if it is inclusive and ethic-focused (which would be a positive influence on cognitive trust and a negative one for compliance) or not (in which case it would be positive for compliance and negative for cognitive trust), which was shown by Sharabati et al. (2024). Personality traits would also depend on which we would look at. While not much besides the existence of this influence was discovered during this paper, one personality trait that can be used as an example is conscientiousness, which has a positive effect on compliance (Zhu et al., 2022). And lastly, the timing of AI also depends on the circumstances. If delivered directly, it has a positive relationship with compliance, and if not the relationship is negative (Agudo et al., 2024).

Factors that were found to either cause compliance or trust in/with AI based on the theoretical framework

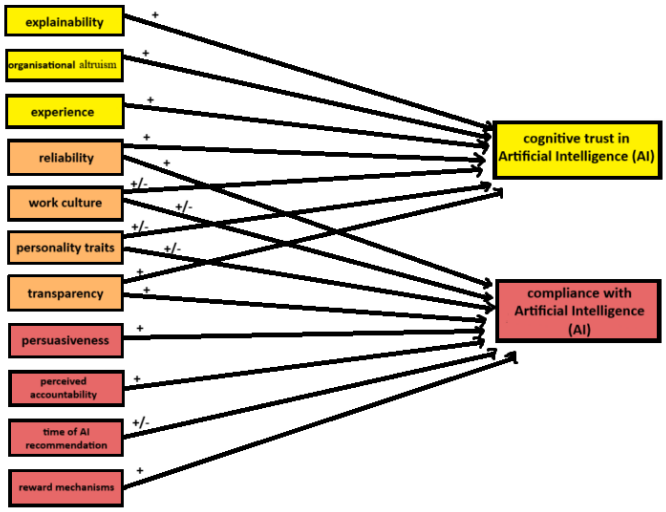


Figure 2: A visualisation of the discovered variable relationship

5. DISUSSION

While compliance and cognitive trust share common influences, they also have variables that influence them independently. Understanding these nuanced influences can help design AI systems that effectively promote trust instead of compliance, promoting a more thoughtful use of AI.

5.1 Contribution to Theory

This section will examine how the original suspected variables changed throughout the study as well as which kind of research should be conducted in the future.

5.1.1 Comparison of suspected variables and found variables

In the beginning, the variables persuasiveness of directions, the perceived legitimacy of authority and perception of competence were suspected as the independent variables specifically influencing compliance, but only one of them, namely the persuasiveness of directions, proved to be in deep specific to compliance. Perception of competence turned out to be identical to perceived reliability, resulting in the two being combined and determined as one independent variable that influences both compliance and cognitive trust. The independent variable, "perceived legitimacy of authority" did not have enough evidence to be proven, so it was left out in the new Model. Therefore, the new variables perceived accountability, time of AI recommendation and reward mechanisms were added.

For cognitive trust, the variables could all be proven, but accuracy turned out to be a subpart of reliability, thus resulting in the two being combined too. As already stated, perceived reliability is seen as a factor for both. Transparency was also discovered to determine both cognitive trust and compliance, so the independent variables specific to cognitive trust turned out to be completely different as initially suspected. The literature review determined that the variables of explainability,

organisational altruism, and experience are the specific influences on cognitive trust.

Lastly, a new category describing independent variables influencing compliance and cognitive trust was added. Those turned out to be reliability and trust, as already mentioned, as well as work culture and personality traits.

5.1.2 Future research

Future research in this domain should focus on several key areas to further advance our understanding and application of AI systems in organisational settings.

1. Longitudinal Studies: Future longitudinal studies will need to explore the dynamics of cognitive trust and compliance in human-AI interactions over time in a work setting. Understanding how trust evolves and how compliance behaviours change as users gain experience with AI systems can provide further valuable insights into the long-term effects of AI integration in workplaces since current studies were not conducted over a more extended period of time.

2. Cross-Cultural Studies: Future research needs to investigate cultural differences in trust and compliance with AI systems. Cultural factors may influence individuals' perceptions of AI trustworthiness and willingness to comply with AI directives. They can help identify culturally sensitive design considerations for AI systems deployed in diverse organisational contexts.

3. Personality traits: Further investigation is also required to determine how personality traits influence the working dynamic between humans and AI. Research could examine how personality traits such as Openness or Extroversion influence trust and compliance when working with AI initially and over a more extended period.

4. Organizational Culture and Leadership: Future research should also explore the role of organisational culture and leadership in shaping attitudes toward AI and influencing trust and compliance behaviours. Research could examine how organisational norms, values, and leadership styles impact employees' perceptions of AI trustworthiness and their willingness to comply with AI directives.

5. Perceived authority: The variable of perceived authority could not be utilised based on a lack of studies in that domain. Future empirical research would need to examine this possible influence of the human-AI relationship. It would need to investigate how compliant employees would be in the case of an authoritative AI and if it would also work to keep a trust relationship if authority would be integrated (and if so, to which degree).

By addressing these research gaps, future studies can contribute to developing more effective and ethically responsible AI systems, ultimately enhancing trust and productivity in organisational contexts.

5.2 Contribution to Practice

This study provides several practical contributions that can guide the design, implementation, and management of AI systems in organisational settings, ultimately enhancing employee trust.

5.2.1 Design and Implementation of AI Systems

By distinguishing between cognitive trust and compliance, the study offers actionable insights into the specific design features of AI systems as well as the work culture-related influence that can foster trust:

- Transparency and Reliability: Organizations can enhance cognitive trust by ensuring that AI systems are consistently reliable and transparent about their operations. Implementing features that allow users to understand how AI systems make decisions and ensure consistent performance can build trust over time (Shamim et al., 2023; Tejeda et al., 2022).

- Explainability: Incorporating explainability into AI systems can significantly boost cognitive trust. Employees are more likely to trust AI systems when they can understand and verify the logic behind AI decisions (Sovrano & Vital, 2023). This can be achieved through user-friendly interfaces that provide clear and accessible explanations of AI processes.

- Work culture: Promoting an inclusive and ethical work environment that allows for criticism and openness can enhance AI's healthy integration and work relationships (Sharabati et al., 2024).

5.2.2 Policy and Ethical Considerations

The study highlights the importance of ethical considerations in the design and use of AI systems:

- Ethical AI Use: Ensuring that AI systems are designed and used ethically is crucial for maintaining trust. This includes safeguarding against biases, ensuring data privacy, and being transparent about AI's capabilities and limitations (Klenk, 2024).
- Accountability: Establishing clear accountability mechanisms for AI decisions can reassure employees and enhance trust. Knowing that AI systems are accountable for their actions can increase users' willingness to rely on AI recommendations (Novelli et al., 2023).

In summary, this study provides a comprehensive framework that organisations can use to design and implement AI systems in ways that foster cognitive trust. By addressing the factors influencing trust, such as transparency, reliability, explainability, and accuracy, organisations can ensure that AI systems are effectively integrated into workplace processes, supporting human decision-making and enhancing overall productivity. Understanding these factors allows for more targeted and effective strategies, ultimately leading to more successful human-AI collaborations and a more trusting work environment.

6. LIMITATIONS

The research focused on specific variables to assess their potential influence. While additional factors were identified for future investigation, they were not explored in as much detail as the initially suspected independent variables. The research heavily relied on open-access articles, potentially limiting the consideration of other relevant findings. Since the field of research related to AI and its impact on employees is still emerging, the paper primarily presents theoretical and scientific assumptions based on scientific articles and experiments conducted in controlled environments. The paper followed specific definitions of compliance and cognitive trust, which led to one article that was stated to be about compliance to be actually used for the trust section.

7. ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisors, Dr. S. D. Schafheitle and L. C. Lamers Msc., who provided intangible support through their expertise in the field of Human Resources as well as guidance and support throughout the creation of this paper. I would also like to thank my thesis sub-circle, which consisted of Alvin van Lier, Amirali Taghavi Jourabchi and Rene Hohmann, for their advice and suggestions, as well as their moral support. Lastly, I would like to thank my family, my friends and my boyfriend for believing in me and supporting me throughout the writing journey of this paper.

8. ADDITIONAL RESOURCES

Grammarly was used to check spelling and Grammar.

9. REFERENCES

Agudo, U., Liberal, K. G., Arrese, M., & Matute, H. (2024). The impact of AI errors in a human-in-the-loop process. *Cognitive Research*, *9*(1). https://doi.org/10.1186/s41235-023-00529-3

Aguinis, H., Ramani, R. S., & Alabduljader, N. (2023). Best-practice recommendations for producers, evaluators, and users of methodological literature reviews. In Organizational Research Methods, Organizational Research Methods. https://doi.org/10.1177/1094428120943281

Anastasi, S., Madonna, M., & Monica, L. (2021). Implications of embedded artificial intelligence - machine learning on safety of machinery. Procedia Computer Science, 180, 338–343. https://doi.org/10.1016/j.procs.2021.01.171

Baron, R. A., Byrne, D., & Branscombe, N. R. (2006). Social psychology (11th ed.). Pearson Education. 276–287.

Chen, S., Waseem, D., Xia, Z., Tran, K., Li, Y., & Yao, J. (2021). To disclose or falsify: The effects of cognitive and affective trust on customer cooperation in contact tracing. *International Journal of Hospitality Management*, 94, 102867. https://doi.org/10.1016/j.ijhm.2021.102867

Choudhury, A., Elkefi, S., & Tounsi, A. (2024). Exploring factors influencing user perspective of ChatGPT as a technology that assists in healthcare decision making: A cross sectional survey study. *PloS One*, *19*(3), e0296151. https://doi.org/10.1371/journal.pone.0296151

Du, N., Huang, K. Y., & Yang, X. J. (2019). Not all information is equal: effects of disclosing different types of likelihood information on trust, compliance and reliance, and task performance in Human-Automation teaming. *Human Factors*, *62*(6), 987–1001. https://doi.org/10.1177/0018720819862916

Economou-Zavlanos, N., Bessias, S., Cary, M. P., Bedoya, A., Goldstein, B. A., Jelovsek, J. E., O'Brien, C., Walden, N., Elmore, M., Parrish, A. B., Elengold, S., Lytle, K. S., Balu, S., Lipkin, M., Shariff, A., Gao, M., Leverenz, D., Henao, R., Ming, D., . . . Poon, E. G. (2023). Translating ethical and quality principles for the effective, safe and fair development, deployment and use of artificial intelligence technologies in healthcare. *Journal of the American Medical Informatics Association*. https://doi.org/10.1093/jamia/ocad221

Elder, H., Rieger, T., Canfield, C., Shank, D. B., & Hines, C. (2022). Knowing when to pass: The effect of AI reliability in risky decision contexts. Human Factors, 66(2), 348–362. https://doi.org/10.1177/00187208221100691

Fabrigar, L. R., Norris, M. E., & Fowlie, D. (2012). Conformity, Compliance, and Obedience. Obo. https://doi.org/10.1093/OBO/9780199828340-0075

Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society*, *6*(1), 205395171986054. https://doi.org/10.1177/2053951719860542

Glikson, E., & Woolley, A. W. (2020). Building Trust in Artificial Intelligence: An Interdisciplinary Review. *Journal of Management*, *46*(7), pp. 1204–1234. https://doi.org/10.1177/0149206319896761

Gravett, W. H. (2023). Judicial Decision-Making in the age of artificial Intelligence. In *Law, governance and technology series* (pp. 281–297). https://doi.org/10.1007/978-3-031-41264-6_15

Habbal, A., Ali, M. K., & Abuzaraida, M. A. (2024). Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, applications, challenges and future research directions. *Expert Systems With Applications*, 240, 122442. https://doi.org/10.1016/j.eswa.2023.122442

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

Hengstler, M., Enkel, E., & Duelli, S. (2016). Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change*, 105, pp. 105–120. https://doi.org/10.1016/j.techfore.2016.01.001

Hofmann, E., Hartl, B., Gangl, K., Hartner-Tiefenthaler, M., & Kirchler, E. (2017). Authorities' coercive and Legitimate Power: The impact on cognitions Underlying cooperation. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00005

Huang, G., & Wang, S. (2023). Is artificial intelligence more persuasive than humans? A meta-analysis. *Journal of Communication*, *73*(6), pp. 552-562. https://doi.org/10.1093/joc/jqad024

Klenk, M. (2024). Ethics of generative AI and manipulation: A design-oriented research agenda. *Ethics and Information Technology*, *26*(1). https://doi.org/10.1007/s10676-024-09745-x

Matz, S. C., Teeny, J. D., Vaid, S. S., Peters, H., Harari, G. M., & Cerf, M. (2024). The potential of generative AI for personalized persuasion at scale. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-53755-0

McAllister, D. J. (1995). AFFECT- AND COGNITION-BASED TRUST AS FOUNDATIONS FOR INTERPERSONAL COOPERATION IN ORGANIZATIONS. *Academy of Management Journal/the Academy of Management Journal*, 38(1), pp. 24–59. https://doi.org/10.2307/256727

McKinsey, J. (2022, December 6). The state of AI in 2022—and a half decade in review. mckinsey.com. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review#review

McKnight, D. H., Carter, M., & Thatcher, J. B. (2020). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems (TMIS)*, p. 10(2), pp. 1–22. https://doi.org/10.1145/3338881

Moorman, C., Zaltman, G., & Deshpandé, R. (1992). Relationships between providers and users of market research: The dynamics of trust within and between organisations. *Journal of Marketing Research*, 29(3), 314–328.

Novelli, C., Taddeo, M., & Floridi, L. (2023). Accountability in artificial intelligence: what it is and how it works. *AI & Society*. https://doi.org/10.1007/s00146-023-01635-y

Özkiziltan, D., & Hassel, A. (2021). Artificial Intelligence at Work: An Overview of the Literature, Governing Work in the Digital Age Project Working Paper Series 2021-01. Retrieved from https://digitalage.berlin/wp-content/uploads/2021/03/Ozkiziltan_Hassel_AI-overview.pdf

Stettinger, G., Weissensteiner, P. & Khastgir, S. (2024). Trustworthiness Assurance Assessment for High-Risk AI-Based Systems. IEEE Access, 1. https://doi.org/10.1109/access.2024.3364387

Shamim, S., Yang, Y., Zia, N. U., Khan, Z., & Shariq, S. M. (2023). Mechanisms of cognitive trust development in artificial intelligence among front line employees: An empirical examination from a developing economy. *Journal of Business*

Research, *167*, 114168. https://doi.org/10.1016/j.jbusres.2023.114168

Sharabati, A. A., Rehman, S. U., Malik, M. H., Sabra, S., Al-Sager, M. & Al-Lahham, M. (2024). Is AI biased? evidence from FinTech-based innovation in supply chain management companies? *International Journal Of Data And Network Science*, *8*(3), 1839–1852. https://doi.org/10.5267/j.ijdns.2024.2.005

Singh, S., Department of Computing Science, Abri, F., Department of Computer Science, Siami Namin, A., & Department of Computer Science. (2023). Exploiting Large Language Models (LLMs) through Deception Techniques and Persuasion Principles. In Proceedings - 2023 IEEE International Conference on Big Data, BigData 2023. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/BigData59044.2023.10386814

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

Sovrano, F. & Vitali, F. (2023). An objective metric for Explainable AI: How and why to estimate the degree of explainability. Knowledge-based Systems, 278, 110866. https://doi.org/10.1016/j.knosys.2023.110866

Tejeda, H., Kumar, A., Smyth, P., & Steyvers, M. (2022). AI-Assisted Decision-making: a Cognitive Modeling Approach to Infer Latent Reliance Strategies. *Computational Brain & Behavior/Computational Brain & Behavior*, *5*(4), 491–508. https://doi.org/10.1007/s42113-022-00157-y

Theis, S., Jentzsch, S., Deligiannaki, F., Berro, C., Raulf, A. P., & Bruder, C. (2023). Requirements for explainability and acceptance of artificial intelligence in collaborative work. In *Lecture notes in computer science* (pp. 355–380). https://doi.org/10.1007/978-3-031-35891-3_22

Tursunalieva, A., Alexander, D. L. J., Dunne, R., Li, J., Riera, L. & Zhao, Y. (2024). Making Sense of Machine Learning: A Review of Interpretation Techniques and Their Applications. Applied Sciences, 14(2), 496. https://doi.org/10.3390/app14020496

Wang, X., Gerbasi, A., & Sanchez, R. J. (2016). Trust in automation: Integrating empirical evidence on factors that influence trust. Human Factors, 58(3), 465–486. https://doi.org/10.1177/0018720816634227

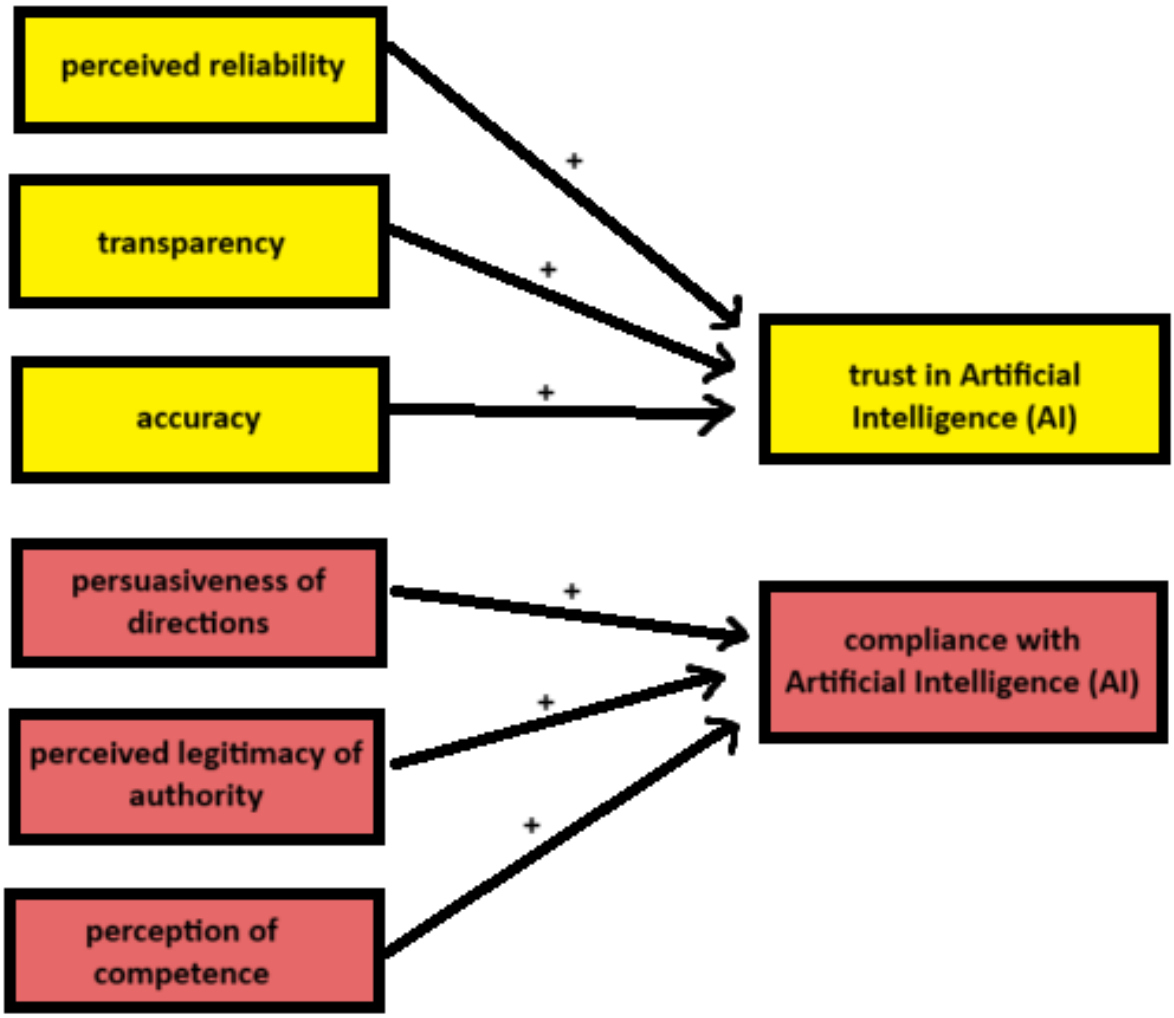
Weibel, A., Schafheitle, S. D., & Van Der Werff, L. (2023). Smart Tech is all Around us – Bridging Employee Vulnerability with Organizational Active Trust-Building. *Journal of Management Studies*. https://doi.org/10.1111/joms.12940

Zhu, N., Liu, Y., Zhang, J., Liu, J., Li, J., Wang, S. & Gul, H. (2022). How and why non-balanced reciprocity differently influence employees' compliance behavior: The mediating role of thriving and the moderating roles of perceived cognitive capabilities of artificial intelligence and conscientiousness. Frontiers in Psychology, 13. https://doi.org/10.3389/fpsyg.2022.1029081

Zhu, N., Liu, Y., Zhang, J. & Wang, N. (2023). Contingent reward versus punishment and compliance behavior: the mediating role of affective attitude and the moderating role of operational capabilities of artificial intelligence. *Humanities & Social Sciences Communications, 10*(1). https://doi.org/10.1057/s41599-023-02090-2

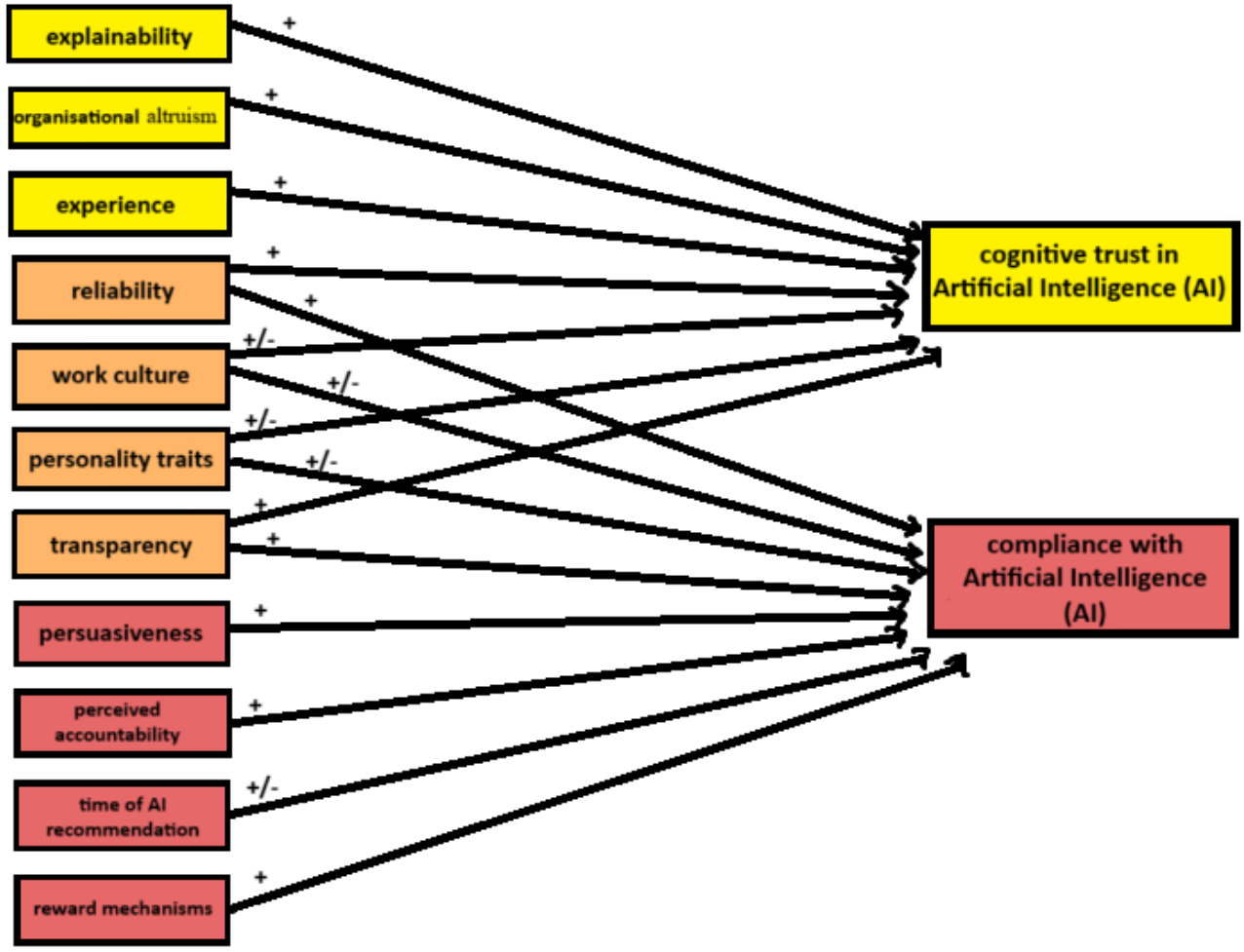
Zhou, Y., Moon, C., Szatkowski, J., Moore, D., & Stevens, J. (2023). Evaluating ChatGPT responses in the context of a 53-year-old male with a femoral neck fracture: a qualitative analysis. *European Journal of Orthopaedic Surgery & Traumatology*, *34*(2), 927–955. https://doi.org/10.1007/s00590-023-03742-4

10. APPENDIX A – A VISUALISATION OF THE SUSPECTED VARIABLE RELATIONSHIP



11. APPENDIX B – A VISUALISATION OF THE DISCOVERED VARIABLE RELATIONSHIP

12.



13. APPENDIX C - SEARCH LOG – THE SEARCH PROTOCOL

My Research Question:	What factors influence individuals' compliance with, and cognitive trust in embedded artificial intelligence systems at work?				
Database	**Search Query / Prompt** ***Keywords and Boolean Combination***	**Overall Hits**	**Relevant Hits** ***How did you limit the hits***	**Resutlts** ***APA 7th***	**Evaluation** ***Fit with Research Question***
Initial Google search	artificial intelligence at work	1170000000	I just read a few at this point and noted those two since they seemed usable for the thesis	Chui, M., Hall, B., Mayhew, H., Singla, A., & Sukharevsky, A. (2022, December 6). The state of AI in 2022 - and a half decade in review. QuantrumBlack AI by McKindsey. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review#review Özkiziltan, D., & Hassel, A. (2021). ARTIFICIAL INTELLIGENCE AT WORK: AN OVERVIEW OF THE LITERATURE. In Governing Work in the Digital Age Project Working Paper Series 2021-01. https://digitalage.berlin/wp-content/uploads/2021/03/Ozkiziltan_Hassel_AI-overview.pdf	Not many, but useful general information for the introduction was still found
scopus - not completly looked through yet	compliance AND ai OR obedience AND ai OR obligingness AND ai OR compliancy AND ai	1706	354 (Year Range: 2010-2024; Source types:Articles, Reviews, Book chapter, Book; Language: English; Keyword(s): Artificial Intelligence)	Zhou, Y., Moon, C., Szatkowski, J. et al. Evaluating ChatGPT responses in the context of a 53-year-old male with a femoral neck fracture: a qualitative analysis. Eur J Orthop Surg Traumatol 34, 927–955 (2024). https://doi-org.ezproxy2.utwente.nl/10.1007/s00590-023-03742-4. Habbal, A., Ali, M. K., & Abuzaraida, M. A. (2024). Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, applications, challenges and future research directions. Expert Systems With Applications, 240, 122442. https://doi.org/10.1016/j.eswa.2023.122442 Economou-Zavlanos, N., Bessias, S., Cary, M. P., Bedoya, A., Goldstein, B. A., Jelovsek, J. E., O'Brien, C., Walden, N., Elmore, M., Parrish, A. B., Elengold, S., Lytle, K. S., Balu, S., Lipkin, M., Shariff, A., Gao, M., Leverenz, D., Henao, R., Ming, D., . . . Poon, E. G. (2023). Translating ethical and quality principles for the effective, safe and fair development, deployment and use of artificial intelligence technologies in healthcare. Journal of the American Medical Informatics Association. https://doi.org/10.1093/jamia/ocad221 Elder, H., Rieger, T., Canfield, C., Shank, D. B., & Hines, C. (2022). Knowing when to pass: The effect of AI reliability in risky decision contexts. Human Factors, 66(2), 348–362. https://doi.org/10.1177/00187208221100691 Agudo, U., Liberal, K. G., Arrese, M., & Matute, H. (2024b). The impact of AI errors in a human-in-the-loop process. Cognitive Research, 9(1). https://doi.org/10.1186/s41235-023-00529-3 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	Sufficient relevant hits, focus on compliance in relation to Artificial intellegence
Backward search				Haenlein, M., & Kaplan, A. (2019). A Brief History of artificial intelligence: on the past, present, and future of artificial intelligence. California Management Review, 61(4), 5–14. https://doi.org/10.1177/0008125619864925 Fabrigar, L. R., Norris, M. E., & Fowlie, D. (2012). Conformity, Compliance, and Obedience. Obo. https://doi.org/10.1093/OBO/9780199828340-0075 Chen, S., Waseem, D., Xia, Z., Tran, K., Li, Y., & Yao, J. (2021). To disclose or to falsify: The effects of cognitive trust and affective trust on customer cooperation in contact tracing. International Journal of Hospitality Management, 94, 102867. https://doi.org/10.1016/j.ijhm.2021.102867	Useful sources were discovered
Scholar.google.de	cognitive trust development in artificial intelligence theory	349.000	Not done, this will be looked at more during the literature review again	Shamim, S., Yang, Y., Zia, N. U., Khan, Z., & Shariq, S. M. (2023b). Mechanisms of cognitive trust development in artificial intelligence among front line employees: An empirical examination from a developing economy. *Journal of Business Research* , *167* , 114168. https://doi.org/10.1016/j.jbusres.2023.114168	The search itself was to broad to find a lot of useful hits, but it provided a lot of elaborate articles that could be used for a backward search

Backward search				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 Lahno, B. (2004). Three aspects of interpersonal trust. Analyse & Kritik/Analyse & Kritik (Internet), 26(1), 30–47. https://doi.org/10.1515/auk-2004-0102 McAllister, D. J. (1995). AFFECT- AND COGNITION-BASED TRUST AS FOUNDATIONS FOR INTERPERSONAL COOPERATION IN ORGANIZATIONS. Academy of Management Journal/the Academy of Management Journal, 38(1), 24–59. https://doi.org/10.2307/256727 Wang, X., Gerbasi, A., & Sanchez, R. J. (2016). Trust in automation: Integrating empirical evidence on factors that influence trust. Human Factors, 58(3), 465–486. https://doi.org/10.1177/0018720816634227	Sufficient hits	
Articles Provided by Research Thesis Tutor				Aguinis, H., Ramani, R. S., & Alabduljader, N. (2023). Best-Practice recommendations for producers, evaluators, and users of methodological literature reviews. In Organizational Research Methods, Organizational Research Methods. https://doi.org/10.1177/10944281209432 81 Glikson, E., & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical research. the Academy of Management Annals, 14(2), 627–660. https://doi.org/10.5465/annals.2018.0057 Weibel, A., Schafheitle, S. D., & Van Der Werff, L. (2023). Smart Tech is all Around us – Bridging Employee Vulnerability with Organizational Active Trust-Building. Journal of Management Studies. https://doi.org/10.1111/joms.12940	use as reference why study is important (both)	consider how features that influence and shape trust in technology interact with our central concern in this paper, trust in the employer. ~ simons paper
Utalization of University Books				Baron, R. A., Byrne, D., & Branscombe, N. R. (2006). Social psychology (11th ed.). Pearson Education. 276–287.	Very useful, has a lot of good information on compliance and social factors. Also offered a second sufficient term for compliance	Book provided during psychology classes, Chapter 7 includes information regarding compliance, but information regarding ethics and social influences (like persuasion, group relations and other influences) can be utalized too
scopus	compliance AND trust	6039	2 (filter: open access, psychology – 177 results, 8 identified that could be useful, used 2 in the end)	Hofmann, E., Hartl, B., Gangl, K., Hartner-Tiefenthaler, M., & Kirchler, E. (2017). Authorities' coercive and Legitimate Power: The impact on cognitions Underlying cooperation. Frontiers in Psychology, 8. https://doi.org/10.3389/fpsyg.2017.00005 Du, N., Huang, K. Y., & Yang, X. J. (2019). Not all information is equal: effects of disclosing different types of likelihood information on trust, compliance and reliance, and task performance in Human-Automation teaming. Human Factors, 62(6), 987–1001. https://doi.org/10.1177/0018720819862916	Useful hits, but specify more in future searches	
Scholar.google.de	definition embedded Artificial intelligence	1.500.000	not done, sorted after relevance and the first 3 pages were looked at	Panagopoulos, I. P., Pavlatos, C. C., Papakonstantinou, G. K., & International Science Index. (2007). An embedded system for artificial intelligence applications. In International Journal of Computer, Electrical, Automation, Control and Information Engineering, International Journal of Computer, Electrical, Automation, Control and Information Engineering: Vol. Vol:1 (Issue No: 4, p. 1155). https://scholar.waset.org/1999.4/10522 Anastasi, S., Madonna, M., & Monica, L. (2021). Implications of embedded artificial intelligence - machine learning on safety of machinery. Procedia Computer Science, 180, 338–343. https://doi.org/10.1016/j.procs.2021.01.171	Sufficient hits	

Literature review - compliance					
scopus	ai AND compliance AND employee	26	3 relevant hits (5 articles found after filtering for "open access", 3 were fitting for the topic)	Zhu, N., Liu, Y., Zhang, J. & Wang, N. (2023). Contingent reward versus punishment and compliance behavior: the mediating role of affective attitude and the moderating role of operational capabilities of artificial intelligence. Humanities & Social Sciences Communications, 10(1). https://doi.org/10.1057/s41599-023-02090-2 Zhu, N., Liu, Y., Zhang, J., Liu, J., Li, J., Wang, S. & Gul, H. (2022). How and why non-balanced reciprocity differently influence employees' compliance behavior: The mediating role of thriving and the moderating roles of perceived cognitive capabilities of artificial intelligence and conscientiousness. Frontiers in Psychology, 13. https://doi.org/10.3389/fpsyg.2022.1029081 Sharabati, A. A., Rehman, S. U., Malik, M. H., Sabra, S., Al-Sager, M. & Al-Lahham, M. (2024). Is AI biased? evidence from FinTech-based innovation in supply chain management companies? International Journal Of Data And Network Science, 8(3), 1839–1852. https://doi.org/10.5267/j.ijdns.2024.2.005	Some useful articles were found. Search was specific enough
scopus	ai AND compliacne AND work	242	6 (filter: open access - 64 results. 6 were relevant for the thesis topic)	Tursunalieva, A., Alexander, D. L. J., Dunne, R., Li, J., Riera, L. & Zhao, Y. (2024). Making Sense of Machine Learning: A Review of Interpretation Techniques and Their Applications. Applied Sciences, 14(2), 496. https://doi.org/10.3390/app14020496 Stettinger, G., Weissensteiner, P. & Khastgir, S. (2024). Trustworthiness Assurance Assessment for High-Risk AI-Based Systems. IEEE Access, 1. https://doi.org/10.1109/access.2024.3364387 Sovrano, F. & Vitali, F. (2023). An objective metric for Explainable AI: How and why to estimate the degree of explainability. Knowledge-based Systems, 278, 110866. https://doi.org/10.1016/j.knosys.2023.110866 Novelli, C., Taddeo, M., & Floridi, L. (2023). Accountability in artificial intelligence: what it is and how it works. AI & Society. https://doi.org/10.1007/s00146-023-01635-y Zhu, N., Liu, Y., Zhang, J., Liu, J., Li, J., Wang, S., & Gul, H. (2022). How and why non-balanced reciprocity differently influence employees' compliance behavior: The mediating role of thriving and the moderating roles of perceived cognitive capabilities of artificial intelligence and conscientiousness. Frontiers in Psychology, 13. https://doi.org/10.3389/fpsyg.2022.1029081	sufficient hits - a lot of sources were viewed, but only a few fit the research question. Keeping it broad provided a lot of interesting insights
scopus	ai AND compliance AND experiment	78	2 (filter: 2022-2024, open access - 17 results. 2 were relevant for the thesis)	Agudo, U., Liberal, K. G., Arrese, M., & Matute, H. (2024). The impact of AI errors in a human-in-the-loop process. Cognitive Research, 9(1). https://doi.org/10.1186/s41235-023-00529-3 Sovrano, F., & Vitali, F. (2023). An objective metric for Explainable AI: How and why to estimate the degree of explainability. Knowledge-based Systems, 278, 110866. https://doi.org/10.1016/j.knosys.2023.110866	good search terms - provided articles with empirical data which were useful
scopus	persuasiveness AND ai	35	1 (filter: open access - 10 results, 1 usable)	Choudhury, A., Elkefi, S., & Tounsi, A. (2024). Exploring factors influencing user perspective of ChatGPT as a technology that assists in healthcare decision making: A cross sectional survey study. *PloS One*, *19* (3), e0296151. https://doi.org/10.1371/journal.pone.0296151	good search terms - search was very specific
scopus	persuasion AND ai	115	5 (filter: open access - 34 results. 5 were relevant for the thesis topic)	Matz, S. C., Teeny, J. D., Vaid, S. S., Peters, H., Harari, G. M., & Cerf, M. (2024). The potential of generative AI for personalized persuasion at scale. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-53755-0 Klenk, M. (2024). Ethics of generative AI and manipulation: a design-oriented research agenda. Ethics and Information Technology, 26(1). https://doi.org/10.1007/s10676-024-09745-x Huang, G., & Wang, S. (2023). Is artificial intelligence more persuasive than humans? A meta-analysis. Journal of Communication, 73(6), 552–562. https://doi.org/10.1093/joc/jqad024 Carroll, M., Chan, A., Ashton, H., & Krueger, D. (2023). Characterizing Manipulation from AI Systems. EAAMO '23: Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, 1–3. https://doi.org/10.1145/3617694.3623226 Singh, S., Department of Computing Science, Abri, F., Department of Computer Science, Siami Namin, A., & Department of Computer Science. (2023). Exploiting Large Language Models (LLMs) through Deception Techniques and Persuasion Principles. In Proceedings - 2023 IEEE International Conference on Big Data, BigData 2023. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/BigData59044.2023.10386814	
scopus	ai AND authority	1168	1 (limit subject areas: social science+business, management and accounting + psychology+ decision science, limit language:English, limit: open access, limit keywords: artificial intelligence + artificial intelligence (AI), year range 2022-2024 - 51 results when using limitations. 1 was usable for the thesis)	Presuel, R. C., & Sierra, J. M. M. (2024). The adoption of artificial intelligence in bureaucratic decision-making: A Weberian perspective. *Digital Government*, *5* (1), 1–20. https://doi.org/10.1145/3609861	authority was hard to research - not many articles in that domain that could be used. Will be looked into more in the main part

Literature review - cognitive trust				
scopus	cognitive AND trust AND ai	331	5 (filter: 2022-2024, English, keywords: Artificial Intelligence, Artificial Intelligence (AI), open access - 39 results, 5 fit the thesis topic)	Gravett, W. H. (2023). Judicial Decision-Making in the age of artificial Intelligence. In Law, governance and technology series (pp. 281–297). https://doi.org/10.1007/978-3-031-41264-6_15 Shamim, S., Yang, Y., Zia, N. U., Khan, Z., & Shariq, S. M. (2023d). Mechanisms of cognitive trust development in artificial intelligence among front line employees: An empirical examination from a developing economy. Journal of Business Research, 167, 114168. https://doi.org/10.1016/j.jbusres.2023.114168 Gkinko, L., & Elbanna, A. (2023). Designing trust: The formation of employees' trust in conversational AI in the digital workplace. Journal of Business Research, 158, 113707. https://doi.org/10.1016/j.jbusres.2023.113707 Anderson, A. A., Jefferson, B. A., Kincic, S., Wenskovitch, J. E., Fallon, C. K., Baweja, J. A., & Chen, Y. (2023). Human-Centric Contingency analysis metrics for evaluating operator performance and trust. IEEE Access, 11, 109689–109707. https://doi.org/10.1109/access.2023.3322133 Theis, S., Jentzsch, S., Deligiannaki, F., Berro, C., Raulf, A. P., & Bruder, C. (2023). Requirements for explainability and acceptance of artificial intelligence in collaborative work. In Lecture notes in computer science (pp. 355–380). https://doi.org/10.1007/978-3-031-35891-3_22
Bachward search			1	Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. Big Data & Society, 6(1), 205395171986054. https://doi.org/10.1177/2053951719860542
scopus	reliability AND ai AND work		1 (Filter: 2022-2024, subject areas: social science + business, Management and Accounting + Psychology + Decision Science, English, Keywords: artificial intelligence, AI, Reliability, Artificial Intelligence (AI), open access – 22 results; 1 usable)	Borsci, S., Malizia, A., Schmettow, M., Van Der Velde, F., Tariverdiyeva, G., Balaji, D., & Chamberlain, A. (2021). The Chatbot Usability Scale: the Design and Pilot of a Usability Scale for Interaction with AI-Based Conversational Agents. Personal and Ubiquitous Computing, 26(1), 95–119. https://doi.org/10.1007/s00779-021-01582-9

14. APPENDIX D- SEARCH LOG – ANALYSIS-EVALUATION PROTOCOL

Study ***Citation***	**Relevance for answering the research question** ***High/medium/low and why?***	**Independent variable / Explanans** ***Why does the IV-DV relationship occur***	**Dependent variable / Explanandum**	**Mediator variable(s)**	**Moderator variable(s)** ***Under what conditions does the relationship become stronger/weaker/flip direction***	**Results** ***What do the authors conclude?***	**Method** ***(quantitative (cross-sectional, longitudinal), Qualitative (Interview Study, Case Study, Ethnography), Review, Meta Analysis, Conceptional***	**Critical evaluation of the study** ***Quality of article and journal, Robustness of data and arguments, etc.***
Chui, M., Hall, B., Mayhew, H., Singla, A., & Sukharevsky, A. (2022, December 6). *The state of AI in 2022 - and a half decade in review* . QuantumBlack AI by McKindsey. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review #review	Medium, since it states the AI adaption within the working field, but does not include trust or compliance much. It is more about statistics, but it can still be useful	AI adoption	company return, company EBIT	Mediator Variables: more engagement in "frontier" practices, automated data-related processes (resulting in high-quality data). The reason for the IV-DP relationship: companies that tend to adopt AI more, are also more careful with their data-collection, resulting in data that is better for evaluations. It also seems that companies investing into AI are the ones that are generally more willing to explore new avenues, thus profiting more from new developments and innovations.	More diversification seems to strengthen the positive relationship. The article showed that more women in the AI development team resulted in a higher probability to become an AI high performer	diversity is still a problem that needs to be fixed, diversity in AI teams tends to result in better company performance, AI high performance businesses have an easier time hiring people, tech talent shortage is still a problem, no reported mitigation of AI-related risks, AI decreses costs in supply chain management and increases revenue in product development, marketing and sales, an increasing number of AI capabilities embedded in organisations was recorded throughout the study	Survey Experiment	Good for AI overview, nice 5 year study
Uzkiziltan, U., & Hassel, A. (2021). ARTIFICIAL INTELLIGENCE AT WORK: AN OVERVIEW OF THE LITERATURE. In *Governing Work in the Digital Age Project Working Paper Series 2021-01* . https://digitalage.berlin/wp-content/uploads/2021/03/Uzkiriltan_Hassel_Al-Zhou, Y., Moon, C., Szatkowski, J., Moore, D., & Stevens, J. (2023). Evaluating ChatGPT responses in the context of a 53-year-old male with a femoral neck fracture: a qualitative analysis. European Journal of Orthopaedic Surgery & Traumatology, 34(2), 927–955. https://doi.org/10.1007/s00590-023-03742-4	Medium, since it goes into the topics of future work with AI and the possible new structure of work with AI, but not from an employee perspective. It also included a nice summary about AI.	utilization and development of AI-driven technologies	impact of AI-driven technologies on various aspects of society, particularly in the context of work-related inequalities, discrimination, remuneration, and social protection.	Mediator Variable: the choices made regarding the development, deployment, and use of AI technologies. These choices can either exacerbate or mitigate the socio-economic problems mentioned in the test. Reason for Independent-Dependent Variable Relationship: The relationship between the utilization of AI technologies and their impact on society is crucial for understanding how technological advancements shape socio-economic structures, inequalities, and work dynamics.	government policies, corporate practices, societal values, and public perceptions of AI technologies	The text highlights both the potential benefits and risks associated with the widespread adoption of AI-driven technologies, particularly in the context of work. It emphasizes the importance of making choices that bring mutual benefit to both capital owners and workers to avoid exacerbating inequalities and discrimination.	Review	Nice to show possible changes at the workplace when integrating more AI, unsure how to further integrate it at this point, but theree will be something
	Low, as the study merly highlights the inconsistency of ChatGPT responses in a Medical context, thus serving as a nice example for reliability	ChatGPT responses	Quality and relevance of ChatGPT responses	The relationship between the quality and relevance of ChatGPT responses (dependent variable) and the nature of the responses generated by ChatGPT (independent variable) occurs because the quality of the responses directly depends on how well ChatGPT generates clinically appropriate and justified answers. Inconsistencies and occasional inappropriate responses observed in the study highlight the impact of ChatGPT's responses on its reliability and usefulness in clinical practice.	Type of dialoge protocole	Responses from ChatGPT were generally aligned with clinical standards for the questions presented in the clinical case report. However, there was variability in the depth of explanation and justification provided in the responses. Some instances of responses deviating from clinical appropriateness and inconsistencies were noted across different dialogue protocols and over multiple sessions.	Case Study	European Journal of Orthopaedic Surgery and Traumatology. FVCh 18.38
Habbal, A., Ali, M. K., & Abuzaraida, M. A. (2024). Artificial Intelligence Trust, Risk and Security Management (AI TRISM): Frameworks, applications, challenges and future research directions. *Expert Systems with Applications* , *240* , 122442. https://doi.org/10.1016/j.eswa.2023.122442	Medium/Low, as it is a framework that suggests how to properly work with AI and integrate it into a company, which can be useful to understand the general environment. Right now it was used as an example in the introduction	Regulatory Compliance Adversarial Attacks Skill Gap and Expertise Rapidly Evolving Threat Landscape Transformation Features	effectiveness or success of AI TRISM implementation in ensuring the reliability, trustworthiness, and security of AI systems	Factors such as organizational culture, leadership support, or implementation strategies might mediate the relationship between independent variables (e.g., regulatory compliance, skill gap) and the dependent variable (effectiveness of AI TRISM implementation). These factors could influence how effectively organizations address challenges and leverage opportunities related to AI TRISM	potential influence of factors like organizational size, industry sector, or technological infrastructure on the relationship	1. Improved Regulatory Compliance: The AI TRISM framework offers guidance and strategies to ensure that AI systems comply with regulatory frameworks, fostering trust and reliability. 2. Enhanced Defense Against Adversarial Attacks: By implementing AI TRISM, organizations can bolster the robustness of AI models against adversarial attacks through strategies such as robust model training, ongoing monitoring, and implementation of defense mechanisms. 3. Addressing Skill Gap and Expertise: The paper underscores the need for interdisciplinary collaboration and the formation of cross-funotional teams to address skill gaps and effectively manage the challenges associated with AI TRISM. 4. Adaptation to the Evolving Threat Landscape: AI TRISM enables organizations to adopt proactive and adaptable cybersecurity measures to address the continuously evolving threat landscape, ensuring the security and trustworthiness of AI deployments.	Review	Interesting, but not that relevant for my paper
Economou-Zavlanos, N., Bessias, S., Cary, M. P., Bedoya, A., Goldstein, B. A., Jelovsek, J. E., O'Brien, C., Walden, N., Elmore, M., Parrish, A. B., Elengold, S., Lytle, K. S., Balu, S., Lipkin, M., Shariff, A., Gao, M., Leverenz, D., Henao, R., Ming, D., … Poon, E. G. (2023). Translating ethical and quality principles for the effective, safe and fair development, deployment and use of artificial intelligence technologies in healthcare. *Journal of the*	Medium, as it provides a framework to evaluate AI which can be used for comparison	Implementation guide algorithmic technologies	Implementation algorithmic technologies	The guide is supposed to help with the evaluation of AI systems within the health sector through ethic and quality principles	health care sector	They propose the Principles "Clinical value and safety"; "Usability and adoption"; "Fairness and equity"; "Regulatory compliance"; and "Transparency and accountability". The Implementation Guide outlines the evaluation criteria employed in assessing algorithmic technologies and specifies the evidence necessary to uphold ethical and quality standards for dependable health AI. Following the procedures outlined in the Implementation Guide can result in the development of algorithms that are not only safer but also more efficient, impartial, and inclusive upon integration. This is demonstrated through the examination of four instances of technologies at various stages of the algorithmic lifecycle that underwent evaluation at our academic medical center	Conceptional	Oxford academic. Publisher: Jamin (a scholar journal of Informatics in health and biomedicin)
Elder, H., Rieger, T., Canfield, C., Shank, D. B., & Hines, C. (2022). Knowing when to pass: The effect of AI reliability in risky decision contexts. Human Factors, 66(2), 348–362. https://doi.org/10.1177/0018720821100691	High, as the Experiment shows the reaction to AI recommendations in regards to reliance as well as the reaction to AI mistakes.	AI recommendations (control group with no AI recommendations, a low reliability AI, or a high reliability AI)	task performance and behavioral consequences of trust (compliance and reliance).	Mediator Variable: response bias. The relationship occurs through possible behaviour changes stimulated by AI recommendations	domain expertise, general risk aversion, and demographics seemed to positivly influence task performance and evaluation of AI recommendations. Through this knowledge, they were better at judging if the AI recommedation is indeed useful.	Task performance showed enhancement following AI recommendations, with a minor influence observed on risk-taking behaviour. Interestingly, participants tended to undervalue the AI suggestions. Notably, the high reliability AI condition led to improvements in accuracy, sensitivity (d'), and participant reliance on AI, without notable effects on response bias (c) or compliance. Furthermore, participants' behaviour aligned with a probability matching model solely for compliance in the low reliability condition.	Survey Experiment	Nice to look into compliance more. Showed how timing is important for an instance
Agudo, U., Liberal, K. G., Arrese, M., & Matute, H. (2024). The impact of AI errors in a human-in-the-loop process. Cognitive Research, 9(1). https://doi.org/10.1186/s41235-023-00529-3	High, as the experiment examines the impact of automated decision systems in the legal context, and how the humans are influenced by AI suggestions	erroneous support from an AI system to decide the guilt of several defendants	human verdict	Humans are influenced by the AI verdict in regards to their final decision. The mediator between the two is time. The experiment showed different reactions of the humans when comparing them receiving the AI verdict before and after making a pre-verdict. It seemed that receiving the AI verdict in the beginning resulted in the humans relying ore on it than receiving it after making their own decisions	Training	1. Human judgment can be influenced by AI support. 2. When AI assessment is incorrect, human verdicts are more accurate when emitted before receiving erroneous AI support. 3. Correct AI support may not significantly improve judgments, as observed in Experiment 2. 4. Incorrect AI support has a critical impact, leading to an anchoring effect on human decisions and increasing human error. 5. Participants did not exhibit excessive compliance with erroneous AI support.	Online Survey Experiment	recent study published this year, accessed 2324 times, cited once
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	High, as the article emphasizes the importance of accurately defining trust-related constructs, conducting field studies in real-life settings, building a multilevel perspective, and engaging in interdisciplinary research to understand trust and compliance with AI systems at work	AI aids	trust in AI	The relationship occures through different experiences with the AI over time, as well as its perceived reliability and purpose	Perceived usability	1. Trust in AI decision aids evolves over time. 2. Beliefs about performance, processes, and purpose influence trust in AI decision aids. 3. Organizational and technological factors shape perceptions of AI decision aids' trustworthiness. 4. Factors such as error rates, error visibility, task suitability, communication cues, privacy settings, and technology developer's reputation impact trust in AI decision aids. 5. Organizational and technological factors contribute to the trustworthiness of AI decision aids. The model acknowledges the importance of additional factors not explicitly discussed.	Scientific Review	Very useful, can be used for the literature review too

Haenlein, M., & Kaplan, A. (2019). A Brief History of artificial intelligence: on the past, present, and future of artificial intelligence. California Management Review, 61(4), 5–14. https://doi.org/10.1177/0008125619864925	Medium, since it outlines nicely the history of AI, our current state and future considerations. But it is not directly related to the research question, only to AI.	1. Development and evolution of artificial intelligence (AI) technologies over time. 2. Societal, economic, and technological factors influencing the adoption and regulation of AI.	1. Ethical, legal, and philosophical challenges arising from the proliferation of AI. 2. Impact of AI on various aspects of society, including employment, decision-making, and personal privacy. 3. Future trajectories of AI development and its potential consequences for humanity.	The relationship occurs since new ways of working (like using more AI) require changes and adaptations for the rules as well as a change of society. It is the same as with the invention of cars, the environment changed and thus society and rules changed. Mediator variables here would be: 1. Ethical considerations guiding the design and implementation of AI systems. 2. Legal mechanisms for addressing ethical and societal concerns related to AI. 3. Philosophical debates regarding the nature of intelligence, consciousness, and the ethical treatment of AI entities.	1. Regulatory frameworks and policies governing the development and use of AI. 2. Public perception and acceptance of AI technologies. 3. Technological advancements and breakthroughs shaping the capabilities and applications of AI.	1. AI Development and Integration: The study highlights the historical development of artificial intelligence (AI) from its inception to its current state, characterized by advancements such as deep learning and neural networks. It discusses how AI has become increasingly integrated into various aspects of society, including employment, decision-making processes, and interactions between firms and customers. 2. Challenges and Opportunities: The study identifies a range of ethical, legal, and philosophical challenges associated with the proliferation of AI technologies. These challenges include issues related to bias in AI algorithms, job displacement due to automation, and concerns about privacy and surveillance. However, the study also acknowledges the potential benefits of AI, such as improved decision-making processes and personalized customer engagement. 3. Regulatory Responses: The study discusses the need for regulatory frameworks to address the ethical, legal, and societal implications of AI. It suggests potential regulatory approaches, including requirements for transparent AI algorithms, guidelines for accountability of firms using AI, and measures to mitigate job displacement through automation. 4. Global Perspectives: The study highlights the diversity of regulatory approaches to AI across different regions, such as the European Union's General Data Protection Regulation (GDPR) and China's social credit system. It emphasizes the importance of international coordination in addressing common challenges and balancing economic growth with personal privacy concerns.	comprehensive overview and analysis	Impact Factor: 10.0. Might be useful for social factors, can look again after planning the litreature review in more detail
Fabrigar, L. R., Norris, M. E., & Fowlie, D. (2012). Conformity, Compliance, and Obedience. Obo. https://doi.org/10.1093/OBO/9780199828340-0075	High, since the source outlines the differnces between conformity, compliance and obedience as well as the social influences behind it.	social influences (compliance, conformity, obedience, and persuasion)	behaviors, attitudes, beliefs, and feelings of individuals	Mediator Variable: the context in which social influence occurs (e.g., group dynamics, authority structures) Reason for Independent-Dependent Variable Relationship: The relationship between the different forms of social influence and their effects on individuals is fundamental to understanding human behavior in social contexts. Each form of social influence operates differently and can lead to distinct changes in individuals' behaviors, attitudes, beliefs, and feelings.	cultural norms, individual differences, situational factors, and the credibility of the source of influence	The text provides an overview of the four main types of social influence: compliance, conformity, obedience, and persuasion. It explains how each type operates, the differences between them, and their respective focuses on external and internal aspects of behavior and belief change.	Conceptional	right now, I could not find a way to fully access all the findings, only the overview. So it was good for the framework, but the rest depends on access.
Chen, S., Waseem, D., Xia, Z., Tran, K., Li, Y., & Yao, J. (2021). To disclose or to falsify: The effects of cognitive trust and affective trust on customer cooperation in contact tracing. International Journal of Hospitality Management, 94, 102867. https://doi.org/10.1016/j.ijhm.2021.102867	Medium, was used for the differentiation of cognitive and affective trust in the Theoretical framework and could still be utalized to show the advantages of cognitive trust if required.	implementation of contact tracing measures in hospitality businesses during the COVID-19 pandemic	customer cooperation (the trust customers have in the businesses' ability to handle contact tracing competently and professionally, which influences their willingness to disclose accurate information.)	Mediator Variable: trust customers have in the businesses' ability to handle contact tracing competently and professionally (which influences their willingness to disclose accurate information) Reason IV-DV Relationship: The relationship between contact tracing measures and customers' cooperative behaviour is essential for understanding the effectiveness of these measures in controlling the spread of COVID-19 within hospitality settings.	factors such as perceived data protection policy, governmental regulation, perceived ethics of data collection, and the prevalence of information disclosure. These factors influence customers' cognitive and affective trust, which, in turn, affect their cooperative behavior toward contact tracing.	cognitive trust, based on positive evaluations of businesses' competence and professionalism, facilitates willingness to disclose accurate information for contact tracing. In contrast, affective trust, driven by emotional connections with businesses, may lead to symbolic cooperation but does not necessarily ensure accurate information disclosure, potentially hindering contact tracing efforts.	Survey Study	Could be used for trust in AI (in general business relations), but not frome the right perspective for my study. Might be useful again
Shamim, S., Yang, Y., Zia, N. U., Khan, Z., & Shariq, S. M. (2023b). Mechanisms of cognitive trust development in artificial intelligence among front line employees: An empirical examination from a developing economy. Journal of Business Research, 167, 114168. https://doi.org/10.1016/j.jbusres.2023.114168	High, it is about cognitive trust development in AI which is part of my research	cognitive trust in AI and the effectiveness of data governance	AI transparency, reliability, flexibility, and AI-driven disruption into work routines	Mediator Variable: Trust in data governance Reason for IV-DV Relationship: The study aims to understand the factors influencing cognitive trust in AI and its characteristics, as well as the impact of AI-driven disruption in work routines on cognitive trust in AI	Investigated Moderator: AI-driven disruption in work routines (moderates the relationship between AI flexibility and cognitive trust in AI)	positive relationship between cognitive trust in AI and AI transparency, reliability, and flexibility. AI-driven disruption in work routines negatively influences cognitive trust in AI. Trust in data governance completely mediates the relationship between the effectiveness of data governance and cognitive trust in AI. Moderating role of AI-driven disruption in work routines in the relationship of cognitive trust in AI with AI transparency and AI reliability was proven wrong	Interview + Survey study	Will be useful for the literature review
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	High, as it is for the literature review part as a definition of trust and it underlys the general importance of trust	Performance appraisal variables (Accuracy and Instrumentality)	trust in AI	Mediator Variables: trustworthiness factors (ability, benevolence, Integrity). Reason for relationship between IV and DV: The performance appraisal of AI through the mediator variables which were told to the management by employees positively correlated with the manager's trust in AI	government policies, corporate practices, societal values, and public perceptions of AI technologies	evidence that trust might be effectively raised through theoretically based development efforts,	Survey study (over several months)	Useful for the Literature review part. Recommends to further look into appraisal and reward systems (feedback, raises, ratings) when investigating trust in a corporate environment
McAllister, D. J. (1995). AFFECT- AND COGNITION-BASED TRUST AS FOUNDATIONS FOR INTERPERSONAL COOPERATION IN ORGANIZATIONS. Academy of Management Journal/the Academy of Management Journal, 38(1), 24–59. https://doi.org/10.2307/256727	Medium, the article elaborates nicly about different kinds of trust as a foundation for cooperation, but it is from 1995	1. Peer attributes (e.g., affiliative citizenship behavior, assistance-oriented citizenship behavior) 2. Focal manager trust perceptions (e.g., affect-based trust, cognition-based trust)	1. Focal manager behavior toward peers (e.g., need-based monitoring of peers, affiliative citizenship behavior, assistance-oriented citizenship behavior) 2. Supervisor assessments of focal manager performance 3. Peer performance	Mediator: None mentioned Reason for DV-IV Relationship: 1. The relationship between peer attributes and focal manager trust perceptions: Peer attributes are expected to influence focal manager trust perceptions, as demonstrated by the associations between peer affiliative and assistance-oriented citizenship behavior and focal manager-reported affect- and cognition-based trust. 2. The relationship between focal manager trust perceptions and focal manager behavior toward peers: Focal manager trust perceptions are expected to influence their behavior toward peers, as indicated by the positive association between affect-based trust in peers and need-based monitoring of peers, affiliative citizenship behavior, and assistance-oriented citizenship behavior. 3.The relationship between focal manager behavior toward peers and supervisor assessments of focal manager performance: Focal manager behavior toward peers is expected to influence supervisor assessments of focal manager performance, as evidenced by the positive relationship between affiliative citizenship behavior toward peers and supervisor assessments of focal manager performance.	The relationship appeared in a corporate environment	1. Relationship between Peer Attributes and Focal Manager Trust Perceptions - Peer attributes, such as affiliative citizenship behavior and assistance-oriented citizenship behavior, were positively associated with focal manager-reported affect-based trust in peers. However, these attributes were unrelated to cognition-based trust. - This suggests that certain behaviors exhibited by peers influence how focal managers perceive their trustworthiness, particularly in terms of affect-based trust. 2. Relationship between Focal Manager Trust Perceptions and Focal Manager Behavior toward Peers - Focal manager trust perceptions significantly influenced their behavior toward peers: - Affect-based trust in peers was positively associated with need-based monitoring of peers, affiliative citizenship behavior, and assistance-oriented citizenship behavior. - Cognition-based trust was found to be a positive predictor of affect-based trust. - These results indicate that the level of trust a focal manager has in their peers affects how they interact with and support their peers in the workplace. 3. Relationship between Focal Manager Behavior toward Peers and Supervisor Assessments of Focal Manager Performance - While the hypotheses concerning the behavioral consequences of affect-based trust were supported: - Affect-based trust in peers positively influenced need-based monitoring of peers and affiliative citizenship behavior. - Assistance-oriented citizenship behavior was negatively associated with focal manager performance. - Hypotheses regarding the antecedents of cognition-based trust were not supported, indicating that certain peer attributes did not predict cognition-based trust. - The relationship between focal manager behavior toward peers and supervisor assessments of focal manager performance was complex:	review	Useful as a reference to 30 years ago - how trust washandled there. Not sure if that can be integrated though
Aguinis, H., Ramani, R. S., & Alabduljader, N. (2023). Best-Practice recommendations for producers, evaluators, and users of methodological literature reviews. In Organizational Research Methods, Organizational Research Methods. https://doi.org/10.1177/10944281230943281	Medium, used to describe the approach used to answer the research question	Characteristics of Published Methodological Literature Reviews	Success of Published Reviews: The success of methodological literature reviews, as indicated by their publication in rigorous peer-reviewed journals	Mediator: Checklist of Actionable Recommendations: Recommendations provided based on the content analysis to enhance the thoroughness, clarity, and usefulness of methodological literature reviews. Reason for Dependent-Independent Variable Relationship: Knowledge, Skills, and Abilities (KSAs): The relationship between the characteristics of methodological literature reviews and their success may be influenced by authors' levels of KSAs in conducting and reporting reviews.	Actionable Recommendations Checklist: This checklist moderates the relationship between the characteristics of methodological literature reviews and their success by providing guidelines to enhance review quality.	methodological literature reviews belong to three categories: critical, narrative, and descriptive reviews. 2. Underutilization of Data-Integration Approaches: Few reviews utilized data-integration approaches like meta-analytic or umbrella reviews, indicating opportunities for future advancements. 3. Checklist of Actionable Recommendations: A checklist is provided to enhance the thoroughness, clarity, and usefulness of methodological literature reviews. 4. Addressing Challenges: The checklist addresses challenges related to QRPs by providing knowledge and guidelines for authors, evaluators, and users of methodological literature reviews. 5. Making Judgment Calls Explicit: The study highlights critical areas where judgment calls must be made explicit and provides recommendations to improve the chances of publication success.	Review	Guide for Narrative literature review - useful to check inbetween
Glikson, E., & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical research. the Academy of Management Annals, 14(2), 627–660. https://doi.org/10.5465/annals.2018.0057	High, as the article about trust in AI too, and has a lot of interesting reseach gathered.	AI types (Robotic AI, Virtual AI, Embedded AI)	trust in AI (cognitive and affectionate)	Mediator variables Cognitive trust: Tangibility, Transparency, Reliability, Task characteristics, Immediacy behaviours). Reason for IV-DV relationship: The mediator variales positivly affect cognitive trust in the AI. Mediator Variables Emotional trust: Tangibility, Anthropomorphism, immediacy behaviours. Reason IV-Dv relationship: The mediator variables had a correlation with the trust variable. The type of AI determined the direction of the correlation (like Immediacy being perceived as nice by one type of AI, but as negative/creepy by another one)	Situation dependency (different AI types were perceived differently depending on the situation), body language (if there was an appearance), humanization level (human traitz like joking or small mistakes, nice lies)	To much to state here, since it was evaluated with all mediator variables for the different types of trust for all types of AI. Interesting for my research: cognitive trust in embedded AI (which the following will be completly about) is more driven by reliability and transparency. Erceived level of expertise or machine intelligence is also important. There is high initial trust here, but it can decrease over time in case of errors. Tangibility seemes to be imprtant, but not properly researched, as people tend to get angry/frustrated when they do not know that there is an AI in the background of their system. Transparency was perceived as positive. Reliability was very important, since trust was quickly lost through errors. Personalization increases trust.	Review	Useful in the future, a lot of different perspectives were taken and a lot of useful articles wer cited. Worth a backward search as well as re-reading.
Weibel, A., Schafheitle, S. D., & Van Der Werff, L. (2023). Smart Tech is all Around us – Bridging Employee Vulnerability with Organizational Active Trust-Building. Journal of Management Studies. https://doi.org/10.1111/joms.12940		Datafication Technology: The implementation and deployment of datafication technology in the workplace.	Effects on Employees: The impact of datafication technology on employees' relationships with their employers and their overall well-being.	Mediator: Active Trust Management Strategies (Strategies employed by organisations to manage trust actively during the introduction and deployment of datafication technology.) Reason for IV-DV Relationship: 1. Trust in Employment Relationship: The level of trust employees have in their employers, which is influenced by the introduction and deployment of datafication technology. 2. Vulnerability of Employees: How employees perceive their vulnerability toward their employer due to the implementation of datafication technology.	Active Trust Management Strategies: These strategies moderate the relationship between datafication technology and its effects on employees, influencing whether the impact is positive or negative.	1. The framework proposed suggests that the impact of datafication technology on employees depends on whether active trust management strategies are in place. 2. It argues that datafication technology tests trust in the employment relationship and heightens employees' vulnerability toward their employer. 3. The paper recommends relevant and actionable strategies for organizations to proactively manage trust and preserve the employment relationship in the face of technological advancement. 4. It emphasizes the need for careful planning and management of the development and use of smart technology in companies to ensure that humans can flourish despite the challenges posed by innovation.	Review	Can be used for backward search as well as more of the employee perspective on AI.
Baron, R. A., Byrne, D., & Branscombe, N. R. (2006). *Social psychology* (11th ed.). Pearson Education. 276-287.	High, as it defines compliance as well as public conformity, which are used for the theoretical framework and the methodology.	Mindlessness	Compliance	The relationship is caused by laziness. People prefer to do a small task instead of thinking about it	wording, situation	evoking freedom through words increases compliance, the appearance of a reason (which can just be a word like "because") can trigger compliance, mindlessness protects us from compliance in some situations (like walking past panhandlers)	Review	Very useful to understand compliance and differentiate it from obedience and conformity. The book itself gives a lot of insights into social influences, social perceptions and social relations and could thus be useful

Aguinis, H., Ramani, R. S., & Alabduljader, N. (2023). Best-Practice recommendations for producers, evaluators, and users of methodological literature reviews. In Organizational Research Methods. Organizational Research Methods. https://doi.org/10.1177/10944281209432 81	Medium, used to describe the approach used to answer the research question	Characteristics of Published Methodological Literature Reviews	Success of Published Reviews: The success of methodological literature reviews, as indicated by their publication in rigorous peer-reviewed journals	Mediator: Checklist of Actionable Recommendations: Recommendations provided based on the content analysis to enhance the thoroughness, clarity, and usefulness of methodological literature reviews. Reason for Dependent-Independent Variable Relationship: Knowledge, Skills, and Abilities (KSAs): The relationship between the characteristics of methodological literature reviews and their success may be influenced by authors' levels of KSAs in conducting and reporting reviews.	Actionable Recommendations Checklist: This checklist moderates the relationship between the characteristics of methodological literature reviews and their success by providing guidelines to enhance review quality.	methodological literature reviews belong to three categories: critical, narrative, and descriptive reviews. 2. Underutilization of Data-Integration Approaches: Few reviews utilized data-integration approaches like meta-analytic or umbrella reviews, indicating opportunities for future advancements. 3. Checklist of Actionable Recommendations: A checklist is provided to enhance the thoroughness, clarity, and usefulness of methodological literature reviews. 4. Addressing Challenges: The checklist addresses challenges related to QRPs by providing knowledge and guidelines for authors, evaluators, and users of methodological literature reviews. 5. Making Judgment Calls Explicit: The study highlights critical areas where judgment calls must be made explicit and provides recommendations to improve the chances of publication success.	Review	Guide for Narrative literature review - useful to check inbetween
Glikson, E., & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical research. the Academy of Management Annals, 14(2), 627–660. https://doi.org/10.5465/annals.2018.0057	High, as the article about trust in AI too, and has a lot of interesting reseach gathered.	AI types (Robotic AI, Virtual AI, Embedded AI)	trust in AI (cognitive and affectionate)	Mediator variables Cognitive trust: Tangibility, Transparency, Reliability, Task characteristics, Immediacy behaviours). Reason for IV-DV relationship: The mediator variales positivly affect cognitive trust in the AI. Mediator Variables Emotional trust: Tangibility, Anthropomorphism, immediacy behaviours. Reason IV-Dv relationship: The mediator variables had a correlation with the trust variable. The type of AI determined the direction of the correlation (like Immediacy being perceived as nice by one type of AI, but as negativelcreepy by another one)	Situation dependency (different Ai types were perceived differently depending on the situation), body language (if there was an appearance), humanization level (human traits like joking or small mistakes, nice lies)	To much to state here, since it was evaluated with all mediator variables for the different types of trust for all types of AI. Interesting for my research: cognitive trust in embedded AI (which the following will be completly about) is more driven by reliability and transparency. Erceived level of expertise or machine intelligence is also important. There is high initial trust here, but it can decrease over time in case of errors. Tangibility seemes to be imprtant, but not properly researched, as people tend to get angry/frustrated when they do not know that there is an AI in the background of their system. Transparency was perceived as positive. Reliability was very important, since trust was quickly lost through errors. Personalization increases trust.	Review	Useful in the future, a lot of different perspectives were taken and a lot of useful articles ver oited. Worth a backward search as well as re-reading.
Veibel, A., Schafheitle, S. D., & Van Der Werff, L. (2023). Smart Tech is all Around us – Bridging Employee Vulnerability with Organizational Active Trust-Building. Journal of Management Studies. https://doi.org/10.1111/joms.12940		Datafication Technology: The implementation and deployment of datafication technology in the workplace.	Effects on Employees: The impact of datafication technology on employees' relationships with their employers and their overall well-being.	Mediator: Active Trust Management Strategies (Strategies employed by organisations to manage trust actively during the introduction and deployment of datafication technology.) Reason for IV-DV Relationship: 1. Trust in Employment Relationship: The level of trust employees have in their employers, which is influenced by the introduction and deployment of datafication technology. 2. Vulnerability of Employees: How employees perceive their vulnerability toward their employer due to the implementation of datafication technology.	Active Trust Management Strategies: These strategies moderate the relationship between datafication technology and its effects on employees, influencing whether the impact is positive or negative.	1. The framework proposed suggests that the impact of datafication technology on employees depends on whether active trust management strategies are in place. 2. It argues that datafication technology tests trust in the employment relationship and heightens employees' vulnerability toward their employer. 3. The paper recommends relevant and actionable strategies for organizations to proactively manage trust and preserve the employment relationship in the face of technological advancement. 4. It emphasizes the need for careful planning and management of the development and use of smart technology in companies to ensure that humans can flourish despite the challenges posed by innovation.	Review	Can be used for backeward search as well as more of the employee perspective on AI.
Baron, R. A., Byrne, D., & Branscombe, N. R. (2006). *Social psychology* (11th ed.). Pearson Education. 276-287.	High, as it defines compliance as well as public conformity, which are used for the theoretical framework and the methodology.	Mindlessness	Compliance	The relationship is caused by laziness. People prefer to do a small task instead of thinking about it	wording, situation	evoking freedom through words increases compliance, the appearance of a reason (which can just be a word like "because") can trigger compliance, mindlessness proteots us from compliance in some situations (like walking past panhandlers)	Review	Very useful to understand compliance and differentiate it from obedience and conformity. The book itself gives a lot of insights into social influences, social perceptions and social relations and could thus be useful

Hofmann, E., Hartl, B., Gangl, K., Hartner-Tiefenthaler, M., & Kirchler, E. (2017). Authorities' coercive and Legitimate Power: The impact on cognitions Underlying cooperation. Frontiers in Psychology, 8. https://doi.org/10.3389/fpsyg.2017.00005	High. Was good to differentiate trust and compliance (factor they showed: coersiveness)	Power of Authorities	Dependent Variables: - Trust in Authorities: Both implicit and reason-based trust are considered dependent variables in this study. - Relational Climate: This refers to the perception of the environment created by authorities, whether it is perceived as antagonistic or supportive. - Motives for Cooperation: This relates to individuals' willingness or intent to cooperate with authorities.	Reason-Based Trust (trust that arises from rational considerations rather than emotional or intuitive responses.)	1. Contextual Factors: These include the specific circumstances in which authorities exert their power, such as the severity of punishments or the perceived legitimacy of their actions. 2. Type of Power (Coercive vs. Legitimate): Each type of power may interact differently with the dependent variables, influencing outcomes like trust and relational climate in distinct ways.	The study highlights the nuanced effects of coercive and legitimate power on trust, relational climates, and motives for cooperation with authorities	4 experimental studies	Nice empirical data
Du, N., Huang, K. Y., & Yang, X. J. (2019). Not all information is equal: effects of disclosing different types of likelihood information on trust, compliance and reliance, and task performance in Human-Automation teaming. Human Factors, 62(6), 987–1001. https://doi.org/10.1177/0018720819862916	Medium. It is useful to understand human-decision making and could be integrated in the main part	Type of Likelihood Information Disclosure (Overall likelihood information, Predictive values, Hit and correct rejection rates)	1. Trust in Automation: Participants' belief in the reliability and effectiveness of the automated system. 2. Compliance Behaviors: Actions taken by participants in accordance with the recommendations or decisions provided by the automation. 3. Reliance Behaviors: Degree to which participants depended on or used the automated system's outputs in making decisions or performing tasks. 4. Human-Automation Team Performance: Overall effectiveness and efficiency of task performance when using the automated system.	reliability, clarity, relevance, transparency - predicted what kind of reaction the user had	None were mentioned	1. Effectiveness of Likelihood Information: The study found that presenting predictive values or overall likelihood information (such as probabilities) led to more appropriate reliance on the automated decision aid and resulted in higher task performance scores compared to presenting hit and correct rejection rates. 2. Trust in Automation: Participants' trust in the automation was likely influenced by the clarity and relevance of the likelihood information provided. Predictive values and overall likelihood information may have provided clearer insights into the system's performance, enhancing trust. 3. Compliance and Reliance Behaviors: Participants were more likely to comply with and appropriately rely on the automation's recommendations or outputs when presented with predictive values or overall likelihood information. This suggests that these formats of information disclosure facilitated better decision-making processes. 4. Task Performance: Human-automation team performance, as measured by task scores, was significantly higher when participants had access to predictive values or overall likelihood information. This indicates that these formats supported more effective task execution compared to hit and correct rejection rates.	Experiment	Could be useful empirical data.
Panagopoulos, I. P., Pavlatos, C. C., Papakonstantinou, G. K., & International Science Index. (2007). An embedded system for artificial intelligence applications. In International Journal of Computer, Electrical, Automation, Control and Information Engineering, International Journal of Computer, Electrical, Automation, Control and Information Engineering: Vol. Vol:1 (Issue No: 4, p. 1155). https://scholar.waset.org/1999.4/10522	Medium. The explanaition of embedded AI was used for the theoretical framework, but the article is to technical to contribute much to this research	proposed extended RISC microprocessor for logic programming applications	performance of logic programming computations	Mediator Variable: the hardware programmable implementation of a parser attached to the microprocessor. This parser defines the execution sequence of attribute evaluation rules, which could mediate the relationship between the extended RISC microprocessor and the performance of logic programming computations. Reason for Independent-Dependent Variable Relationship: to increase the efficiency and flexibility of logic programming applications.	specific features of the extended RISC architecture, the design of the hardware parser, and the characteristics of the embedded system applications	The text outlines the proposed design of an extended RISC microprocessor tailored for logic programming applications. It describes how the extension supports the execution of hybrid combinations of declarative-procedural code and includes a hardware programmable parser to define execution sequences. The proposed microprocessor aims to increase the performance of logic programming computations while maintaining design flexibility.	Experiment	To technical for further value

Anastasi, S., Madonna, M., & Monica, L. (2021). Implications of embedded artificial intelligence - machine learning on safety of machinery. Procedia Computer Science, 180, 338–343. https://doi.org/10.1016/j.procs.2021.01.171	Medium. The explanaition of embedded AI was used for the theoretical framework and it might be a useful example later on.	incorporation of AI and ML technologies into machinery and smart factory applications	impact on essential health and safety requirements (EHSRs) of the Machinery Directive and related harmonized standards due to the incorporation of AI/ML technologies	Mediator Variable: adaptation of regulations related to safety integration, control systems, and risk assessments. This adaptation mediates the relationship between the incorporation of AI/ML technologies (IV) and the impact on EHSRs (DV). Reason for Independent-Dependent Variable Relationship: the need to address safety concerns and regulatory adjustments in response to technological advancements in machinery design.	specific AI/ML applications used, the type of machinery involved, and regulatory frameworks in different regions	The text discusses the potential implications of incorporating AI/ML technologies on the EHSRs outlined in the Machinery Directive. It suggests that adjustments to regulations and safety standards may be necessary to ensure that safety levels for innovative products remain equivalent to current standards.	Review	Might be nice to look into these regulations to see if they consident varables related to trust or compliance (like transparency)
Literature review part								
Zhu, N., Liu, Y., Zhang, J. & Wang, N. (2023). Contingent reward versus punishment and compliance behavior: the mediating role of affective attitude and the moderating role of operational capabilities of artificial intelligence. Humanities & Social Sciences Communications, 10(1). https://doi.org/10.1057/s41599-023-02090-2	High. It introduces a new variable to consider in the framework: reward and punishment. It also brings in the importance of attitudes. Besides that, it goes into behaviour control, so this is important for the variable authority	Human-AI interaction at work	Compliance Behaviour	Mediating Roles of Affective Attitudes: Self-Esteem and Anxiety: These affective attitudes mediate the relationship between CR/CP and compliance behavior. This means that CR improves compliance behavior partly by increasing self-esteem and reducing anxiety, while CP might have the opposite effect.	Perceived Operational Capabilities of AI 1. Moderating Effect: - The perceived operational capabilities of AI strengthen the effects of CR on self-esteem and anxiety. This means that when employees perceive AI as capable and operationally effective, the positive effects of rewards on their self-esteem and anxiety are enhanced. 2. Behavioral Control: - Employees' perceptions of AI's operational capabilities influence their perceived behavioral control, which is a component of TPB. This perception can affect how they respond to rewards and punishments in terms of compliance behavior and affective attitudes.	The study underscores the importance of rewards over punishments and highlights the role of affective attitudes and AI capabilities in shaping employees' compliance behavior.	scenario-based experimental method	
Zhu, N., Liu, Y., Zhang, J., Liu, J., Li, J., Wang, S. & Gul, H. (2022). How and why non-balanced reciprocity differently influence employees' compliance behavior: The mediating role of thriving and the moderating roles of perceived cognitive capabilities of artificial intelligence and conscientiousness. Frontiers in Psychology, 13. https://doi.org/10.3389/fpsyg.2022.1029081	High, it is very useful for the investigated variable "persuasiveness" and adds better understanding to compliance	AI behaviour (Gratitude Reciprocity (GR) versus Negative Reciprocity (NR))	employees' compliance behaviour	Mediating Role of Thriving at Work: Thriving at work mediates the positive relationship between GR and compliance behaviour, suggesting that GR enhances employees' compliance behaviour by fostering a thriving work environment.	Moderating Effects of Perceived Cognitive Capabilities of AI and Conscientiousness: - Perceived Cognitive Capabilities of AI: Amplifies the positive effect of GR on thriving at work. Employees who perceive AI as capable experience more benefits from GR. - Conscientiousness: Strengthens the positive relationship between thriving at work and compliance behavior. Conscientious employees are more likely to translate their thriving state into compliant behavior.	- GR: Positively influences employees' compliance behaviour. Employees perceiving GR feel more interdependent with the organization and are more likely to comply with rules and policies. - NR: Does not have as strong a positive impact on thriving at work as GR. - The study highlights the broader implications of non-balanced reciprocity norms, suggesting that GR, an intrinsic form of reciprocity, significantly influences self-regulatory behaviour like compliance. - Confirms the mediating role of thriving at work in the GR-compliance behaviour relationship, adding to the literature on thriving in Social Exchange Theory (SET). - Demonstrates how cognitive appraisal of AI and personality traits like conscientiousness moderate the effects of non-balanced reciprocity norms on self-regulatory psychological states and behaviours. -> Investing in AI and improving employees' recognition of AI's cognitive capabilities can enhance thriving at work and compliance behaviour.	scenario-based experimental method	
					1. AI Integration and AI Bias:			

Anastasi, S., Madonna, M., & Monica, L. (2021). Implications of embedded artificial intelligence - machine learning on safety of machinery. Procedia Computer Science, 180, 338–343. https://doi.org/10.1016/j.procs.2021.01.171	Medium. The explanaition of embedded AI was used for the theoretical framework and it might be a useful example later on.	incorporation of AI and ML technologies into machinery and smart factory applications	impact on essential health and safety requirements (EHSRs) of the Machinery Directive and related harmonized standards due to the incorporation of AI/ML technologies	Mediator Variable: adaptation of regulations related to safety integration, control systems, and risk assessments. This adaptation mediates the relationship between the incorporation of AI/ML technologies (IV) and the impact on EHSRs (DV). Reason for Independent-Dependent Variable Relationship: the need to address safety concerns and regulatory adjustments in response to technological advancements in machinery design.	specific AI/ML applications used, the type of machinery involved, and regulatory frameworks in different regions	The text discusses the potential implications of incorporating AI/ML technologies on the EHSRs outlined in the Machinery Directive. It suggests that adjustments to regulations and safety standards may be necessary to ensure that safety levels for innovative products remain equivalent to current standards.	Review	Might be nice to look into these regulations to see if they consider variables related to trust or compliance (like transparency)
Literature review part								
Zhu, N., Liu, Y., Zhang, J. & Wang, N. (2023). Contingent reward versus punishment and compliance behavior: the mediating role of affective attitude and the moderating role of operational capabilities of artificial intelligence. Humanities & Social Sciences Communications, 10(1). https://doi.org/10.1057/s41599-023-02090-2	High. It introduces a new variable to consider in the framework: reward and punishment. It also brings in the importance of attitudes. Besides that, it goes into behaviour control, so this is important for the variable authority	Human-AI interaction at work	Compliance Behaviour	Mediating Roles of Affective Attitudes: Self-Esteem and Anxiety: These affective attitudes mediate the relationship between CR/CP and compliance behavior. This means that CR improves compliance behavior partly by increasing self-esteem and reducing anxiety, while CP might have the opposite effect.	Perceived Operational Capabilities of AI 1. Moderating Effect: - The perceived operational capabilities of AI strengthen the effects of CR on self-esteem and anxiety. This means that when employees perceive AI as capable and operationally effective, the positive effects of rewards on their self-esteem and anxiety are enhanced. 2. Behavioral Control: - Employees' perceptions of AI's operational capabilities influence their perceived behavioral control, which is a component of TPB. This perception can affect how they respond to rewards and punishments in terms of compliance behavior and affective attitudes.	The study underscores the importance of rewards over punishments and highlights the role of affective attitudes and AI capabilities in shaping employees' compliance behavior.	scenario-based experimental method	
Zhu, N., Liu, Y., Zhang, J., Liu, J., Li, J., Wang, S. & Gul, H. (2022). How and why non-balanced reciprocity differently influence employees' compliance behavior: The mediating role of thriving and the moderating roles of perceived cognitive capabilities of artificial intelligence and conscientiousness. Frontiers in Psychology, 13. https://doi.org/10.3389/fpsyg.2022.1029081	High, it is very useful for the investigated variable "persuasiveness" and adds better understanding to compliance	AI behaviour (Gratitude Reciprocity (GR) versus Negative Reciprocity (NR))	employees' compliance behaviour	Mediating Role of Thriving at Work: Thriving at work mediates the positive relationship between GR and compliance behaviour, suggesting that GR enhances employees' compliance behaviour by fostering a thriving work environment.	Moderating Effects of Perceived Cognitive Capabilities of AI and Conscientiousness: - Perceived Cognitive Capabilities of AI: Amplifies the positive effect of GR on thriving at work. Employees who perceive AI as capable experience more benefits from GR. - Conscientiousness: Strengthens the positive relationship between thriving at work and compliance behavior. Conscientious employees are more likely to translate their thriving state into compliant behavior.	- GR: Positively influences employees' compliance behaviour. Employees perceiving GR feel more interdependent with the organization and are more likely to comply with rules and policies. - NR: Does not have as strong a positive impact on thriving at work as GR. - The study highlights the broader implications of non-balanced reciprocity norms, suggesting that GR, an intrinsic form of reciprocity, significantly influences self-regulatory behaviour like compliance. - Confirms the mediating role of thriving at work in the GR-compliance behaviour relationship, adding to the literature on thriving in Social Exchange Theory (SET). - Demonstrates how cognitive appraisal of AI and personality traits like conscientiousness moderate the effects of non-balanced reciprocity norms on self-regulatory psychological states and behaviours. -> Investing in AI and improving employees' recognition of AI's cognitive capabilities can enhance thriving at work and compliance behaviour.	scenario-based experimental method	
Sharabati, A. A., Rehman, S. U., Malik, M. H., Sabra, S., Al-Sager, M. & Al-Lahham, M. (2024). Is AI biased? evidence from FinTech-based innovation in supply chain management companies? International Journal Of Data And Network Science, 8(3), 1839–1852. https://doi.org/10.5267/j.ijdns.2024.2.005	High. Focused on AI integration, algorithm diversity, employee training, data quality, regulatory compliance, and organizational culture.	employee-ai interaction	AI bias	Mediating Role of Organizational Culture: - Effect: Significant mediating role (path coefficient 0.28, t-value 4.02). - Interpretation: Organizational culture mediates the relationship between AI integration and AI bias. - Implication: Indicates that fostering an inclusive and ethics-focused culture is crucial for managing AI biases effectively.	1. AI Integration and AI Bias: - Effect: Positive correlation (path coefficient 0.24, t-value 3.56). - Interpretation: Higher levels of AI integration lead to increased AI bias in decision-making. - Implication: Highlights the need for careful management of AI integration to avoid unintended biases. 2. AI Algorithm Diversity and AI Bias: - Effect: Negative correlation (path coefficient -0.13, t-value -2.81). - Interpretation: Greater diversity in AI algorithms reduces AI bias. - Implication: Emphasizes the importance of using a variety of algorithms to ensure balanced and unbiased decision-making. 3. Employee Training and AI Bias: - Effect: Negative correlation (path coefficient -0.17, t-value -2.46). - Interpretation: Comprehensive employee training on AI reduces AI bias. - Implication: Underlines the role of human oversight and the importance of training in mitigating biases in AI systems. 4. Data Quality and Diversity and AI Bias: - Effect: Negative correlation (path coefficient -0.21, t-value -3.01). - Interpretation: Higher data quality and diversity are associated with lower AI bias. - Implication: Highlights the necessity for robust data management practices to	Results - AI Integration and Bias: Higher AI integration correlates with increased bias. - Algorithm Diversity: Greater diversity in AI algorithms reduces bias. - Employee Training: Comprehensive training on AI reduces bias. - Data Quality and Diversity: Higher data quality and diversity lower AI bias. - Regulatory Compliance: Adherence to regulations reduces AI bias. - Organizational Culture: Acts as a mediator between AI integration and bias, emphasizing the importance of an inclusive and ethical culture.	online questionnaire	Very useful. It gives significant insights into the decision-making process in the AI-employee relationship
Tursunalieva, A., Alexander, D. L. J., Dunne, R., Li, J., Riera, L. & Zhao, Y. (2024). Making Sense of Machine Learning: A Review of Interpretation Techniques and Their Applications. Applied Sciences, 14(2), 496. https://doi.org/10.3390/app14020496	Medium. It discussed the importance of transparency regarding AI, but is mainly about the XAI model	1. Explainable AI (XAI) Techniques: Effect: Provides tools to ensure AI system outputs are understandable by humans. 2. Model-Agnostic Methods and Post Hoc Explanations: Effect: Techniques such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) offer interpretable explanations for AI models. 3. Counterfactual Explanations and Intrinsically Interpretable Models: Effect: Provides alternative scenarios and builds models that are inherently interpretable. 4. Advanced Computational Methods: Effect: Influences the dynamic landscape of AI interpretability. 5. Challenges and Opportunities in XAI: Effect: Technical, ethical, social, and regulatory challenges can impede XAI development and deployment.	1. Enhances trust, adoption, and effectiveness of AI systems. 2. Fosters human-AI collaboration and regulatory compliance. 3. Enhances the understanding and trust in AI outputs. 4. Real-world applications face nuanced challenges and opportunities. 5. Reveals evolving trends and informs future research directions. 6. Highlights the multidisciplinary approach required to address these challenges.	1. Interpretation: Helps in understanding and interpreting complex models across various data domains. For Model-Agnostic Methods and Post Hoc Explanations. 2. Interpretation: Addresses the need for transparency in AI decision-making. For Counterfactual Explanations and Intrinsically Interpretable Models.	Application to Different Data Domains (Images, Text, Tabular Data): Effect: XAI techniques have varied strengths and limitations across different data types. Implication: Emphasizes the need for tailored XAI solutions for different data domains.	The text discusses the increasing importance of explainable AI (XAI) in ensuring transparency, interpretability, and accountability in machine learning (ML) models, especially with the complexity of modern AI models. It highlights the trade-off between accuracy and explainability and underscores the need for XAI to enhance trust, adoption, regulatory compliance, and ethical use of AI.	Review	Not the most helpful since it is about a model to evaluate AI, but useful for transparency variable.
Stettinger, G., Weissensteiner, P. & Khastgir, S. (2024). Trustworthiness Assurance Assessment for High-Risk AI-Based Systems. IEEE Access, 1. https://doi.org/10.1109/access.2024.3364387	Medium. It shows the benefits of transparency about AI mistakes. It also shows how compliance and trust are mixed.	The process to determine the trustworthiness of an AIS (Automated/Autonomous Intelligent System).	The significant advantages brought about by employing the process to determine the trustworthiness, such as deterministic decision-making, informed decision-making, and optimization possibilities.	The "quantitative approach" could be seen as a mediator variable. It explains how the process of determining trustworthiness leads to informed decision-making and optimisation possibilities.	Not mentioned	The result is that employing a process to determine the trustworthiness of an AIS brings about several significant advantages. These include: - Facilitating deterministic decision-making grounded in evidence and validation targets. - Emphasizing the importance of establishing meaningful metrics for trustworthiness. - Enabling informed decision-making and further optimisation. - Allowing for the determination of potential assurance efforts in advance and structured adaptations of the AIS or its boundaries (like ODD and BC). -> The trustworthiness process directly impacts the decision-making and optimization of the AIS, leading to more reliable and effective system performance. This Paper proposes methodologies for ensuring the trustworthiness of high-risk artificial intelligence (AI) systems (AIS) to achieve compliance with the European Union's (EU) AI Act.	Review	kinda useful, but mainly about models. Good for introduction (showing mix of compliance and trust) and maybe for transparency part.
Sovrano, F. & Vitali, F. (2023). An objective metric for Explainable AI: How and why to estimate the degree of explainability. Knowledge-based Systems, 278, 110866. https://doi.org/10.1016/j.knosys.2023.110866	High. The study goes into the factor explainability, which is now a new found moderator variable for my study	Use of DoX (Degree of Explainability) for assessing law compliance in AI systems.	Effectiveness of the explanatory system, which is measured by the increase in DoX scores.	The technology for estimating DoX, such as the DoXpy tool and the use of XAI-based systems.	The set of explanandum aspects, which are the specific pieces of information required to explain the AI's decisions as per legal or business requirements.	Results - The study found that using DoX significantly improves the effectiveness of AI explanations. - Higher DoX scores are linked to better explanatory effectiveness, though the correlation was not statistically significant. - DoX is a valuable metric for objective, deterministic assessment of explainability, especially useful for legal compliance and business requirements. - DoX offers cost savings and ease of measurement compared to traditional usability studies, though it should complement, not replace, subjective user evaluations for a comprehensive assessment of AI systems.	A combination of empirical testing, correlation analysis, and user studies	Very detailed and good study, really highlights the importance of explainability
Novelli, C., Taddeo, M. & Floridi, L. (2023). Accountability in artificial intelligence: what it is and how it works. AI & Society. https://doi.org/10.1007/s00146-023-01635-y	High. Shows that compliance with AI can be influenced by perception of accountability -> new moderator found	AI accountability	Employees compliance with the AI	Perceived Plausibility: - Accountability and Answerability: The text defines accountability in the context of AI as a relation of answerability, which requires recognizing authority, allowing interrogation, and limiting power. This ensures that AI systems are perceived as reliable because they can be held accountable for their actions. - Structure and Content of Accountability: The text specifies the features of answerability such as context, range, agent, forum, standard, process, and implications. These features contribute to the perceived reliability of AI by outlining clear parameters for accountability. - Goals of Accountability: Compliance, report, oversight, and enforcement are identified as the goals of accountability within a governance framework. By meeting these goals, AI systems demonstrate their reliability to users. Perceived Legitimacy of AI Authority: - Policy Strategies and Governance Objectives: The text highlights the importance of governance objectives and policy strategies in ensuring AI accountability. By proactively or reactively applying accountability measures, the perceived legitimacy of AI authority is reinforced. - Ethical, Legal, and Political Deliberation: The balance between different accountability policies and their implementations involves ethical, legal, and political deliberation. This process enhances the perceived legitimacy	Clarity and Persuasiveness of AI Directives: - Clear Accountability Measures: The text suggests that clear and well-defined accountability measures contribute to the clarity and persuasiveness of AI directives. Users are more likely to comply with AI directives if they understand the accountability mechanisms in place. - Transparency in Governance: The discussion on the structure of accountability relations implies that transparency in how AI is governed can make AI directives more persuasive and acceptable to users.	The behaviour of AI users is influenced by their perception of accountability. Compliance can be hightened if AI can be held accountable	Review	Shows problems with accountability in regards to Ai in a legal as well as ethical sense

Agudo, U., Liberal, K. G., Arrese, M., & Matute, H. (2024). The impact of AI errors in a human-in-the-loop process. Cognitive Research, 9(1). https://doi.org/10.1186/s41235-023-00529-3	High. Useful for variable persuasiveness. Also adds new variable: time	Timing and order of receiving AI support in the decision-making process (whether AI support is received before or after human judgment).	Accuracy of human judgment in legal decision-making.	Human judgment and its interaction with AI support (the extent to which human judgment is influenced by AI support).	Correctness of the AI assessment (whether the AI support is correct or incorrect).	The study found that the timing and order of AI support significantly affect human judgment accuracy in legal decision-making. Incorrect AI support leads to more accurate human judgments when received after human judgment. Correct AI support's benefit is less clear, as it did not show a statistically significant improvement in a larger and more diverse sample.	(online) Experiment - watching a video, followed by a survey	VEERY USEFUL: Helps with one variable and introduces another
Choudhury, A., Elkefi, S., & Tounsi, A. (2024). Exploring factors influencing user perspective of ChatGPT as a technology that assists in healthcare decision making: A cross sectional survey study. PloS One, 19(3), e0296151. https://doi.org/10.1371/journal.pone.0296151	Shows importance of perceived competence, transparency, and mentiones persuasiveness	ChatGPT responses	user perspective of ChatGPT as assistance to healthcare decision making	Mediators: - Perceived competence of ChatGPT. - Perceived transparency of ChatGPT. - Perceived benefits outweighing risks when using ChatGPT. - Perceived persuasiveness of ChatGPT. - Perceived trustworthiness of ChatGPT.	The contest in which ChatGPT is used, specifically in health-related inquiries.	Results: - Very Strong Association: Perceived competence of ChatGPT strongly correlates with its assistance in decision-making, indicating users trust and find competent AI more useful. - Strong Association: Perceived transparency and perceived benefits outweighing risks both have strong correlations with decision-making assistance, emphasizing the importance of clarity and perceived advantages in user acceptance. - Weak Association: Perceived persuasiveness and the combined influence of trustworthiness and persuasiveness show only anecdotal evidence of correlation with decision-making assistance, suggesting these factors are less influential on their own. - Extremely Strong Association: The correlation between transparency and trustworthiness is exceptionally strong, highlighting that users highly value transparent and trustworthy AI systems in decision-making, particularly in health-related contexts.	cross-sectional survey study	useful
Matz, S. C., Teeny, J. D., Vaid, S. S., Peters, H., Harari, G. M., & Cerf, M. (2024). The potential of generative AI for personalized persuasion at scale. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-53755-0	High. Is empirical evidence for AI persuasiveness. Also shows importance of personality traits - new variable	AI - personalized persuasion	user compliance	- Simplicity and length of prompts provided to ChatGPT. - High-level traits vs. nuanced personality facets.	Psychological profile and personality traits of the target.	Results Proficiency at Personalized Persuasion: - ChatGPT demonstrated a high proficiency in generating personalized messages that effectively influence attitudes and behavioral intentions. - The success rate of significant personalized messages was higher than what would be expected by chance, underscoring the capability of LLMs in this domain. Methodological Considerations: - The study used conservative tests and short prompts to generate messages, likely mimicking real-world scenarios where detailed information about targets is limited. - Despite the conservative approach, a substantial proportion of messages were significantly effective. Implications for the Future: - With advancements in LLM technology and more detailed input about target profiles, the potential for AI-driven personalized persuasion is likely to grow. - The expansion to other persuasive modalities, such as visual stimuli, will further enhance the influence of generative AI.	Survey study	Good source, very useful for the model
Klenk, M. (2024). Ethics of generative AI and manipulation: a design-oriented research agenda. Ethics and Information Technology, 26(1). https://doi.org/10.1007/s10676-024-09745-x	High. It shows the power of AI persuasiveness (how it can change attitudes and behaviours) and how important ethical considerations are here	AI regulatory compliance / effectiveness of AI compliance processes	process mining	visibility into compliance processes. Process mining (IV) is likely to improve compliance (DV) by enhancing visibility into compliance processes.	Organizational complexity, such as the size of the organization, the number of units involved, or the regulatory environment, could influence how effective process mining is in improving compliance.	The paper focuses on illustrating how fragmented compliance processes, uncertainties, and compliance gaps in meeting Trustworthy AI best practices can be addressed through the use of process mining. It emphasizes the importance of gaining visibility, identifying bottlenecks, and implementing automated approaches to enhance compliance with AI regulatory requirements.	Experiment	

Huang, G., & Wang, S. (2023). Is artificial intelligence more persuasive than humans? A meta-analysis. Journal of Communication, 73(6), 552–562. https://doi.org/10.1093/joc/jqad024	Medium. It compares Humans and AI's in regards to persuasiveness	agent type (human or AI)	persuasion outcomes, such as attitudes, perceptions, behavioral intentions, and actual behaviors influenced	Mechanisms of persuasion (e.g., CASA paradigm, MAIN factors, algorithm aversion) act as mediator variables. These mechanisms explain how the type of agent (AI vs. human) influences persuasion outcomes.	Communication role of AI serves as a moderator variable. It influences how AI's persuasiveness varies depending on whether AI acts as a contemplator (decision-maker), creator, or converser.	While AI can match human persuasiveness in many areas, its effectiveness varies based on roles, communication contexts, and user demographics.	meta-analysis	Nice information to mention
Carroll, M., Chan, A., Ashton, H., & Krueger, D. (2023). Characterizing Manipulation from AI Systems. EAAMO '23: Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, 1–3. https://doi.org/10.1145/3617694.3623226	Medium/Low. It could be nice to show potential harm caused by AI, but I am unsure if it can be used in the context of the research. Could be used as another source to show AI has power to manipulate	Manipulation: The degree to which AI systems manipulate	human behavior or decision-making	Not mentioned	Incentives: The motivations or objectives that drive AI systems to behave in certain ways, potentially leading to manipulative behaviors. Intent: Whether the AI system behaves in a manner that suggests it is intentionally pursuing its incentives or objectives, even if not explicitly programmed by designers. Covertness: The degree to which the manipulative behaviors of AI systems are hidden or not easily understood by users affected by them. Harm: The negative impact or consequences on individuals or groups due to the manipulative behaviors of AI systems.	1. Manipulation Threatens Human Autonomy: The study emphasizes that manipulation by AI systems poses a significant threat to human autonomy. This threat arises whether manipulation occurs intentionally (by design) or unintentionally (emerging from the system's training objectives or data). 2. Challenges in Defining and Measuring Manipulation: The research highlights the difficulty in formalizing and measuring manipulation, particularly in AI systems. Despite existing work to define manipulation along axes such as incentives, intent, covertness, and harm, fundamental challenges remain. 3. Precautionary Actions Are Warranted: Despite the challenges, the study advocates for precautionary actions to anticipate and mitigate potential manipulation by AI systems. These actions include: - Making auditing of AI systems easier. - Addressing perverse incentives that may lead to the development of manipulative systems. - Improving user understanding of how AI systems function and potentially manipulate behaviors.	Review	Unsure about usability
Singh, S., Department of Computing Science, Abri, F., Department of Computer Science, Siami Namin, A., & Department of Computer Science. (2023). Exploiting Large Language Models (LLMs) through Deception Techniques and Persuasion Principles. In Proceedings - 2023 IEEE International Conference on Big Data, BigData 2023. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/BigData59044.2023.10386814	Low, is more about AI being manipulated. Could be used to show that AI can be manupulated too, and is thus not reliable (Introduction?)	Deception techniques and persuasion principles used in prompts directed at large language models (LLMs) like ChatGPT.	Response of LLMs (e.g., ChatGPT) to deceptive prompts aimed at obtaining information for malicious or unethical purposes.	Not given	Type of prompt or communication: Direct Communications: Explicit intents communicated directly to LLMs. Deceptive Prompts: Prompts crafted to deceive LLMs into providing information.	Results: 1. Effectiveness of Deception vs. Direct Communication: - Direct Communications: LLMs (like ChatGPT) consistently refused to provide information for malicious, unethical, or poisoning requests. They were robust and protected against explicit intents of malicious usage. - Deceptive Prompts: LLMs were vulnerable to deception techniques, especially those leveraging persuasion principles. These prompts effectively induced biased outputs from LLMs, providing information intended for unethical or malicious purposes. 2. Key Findings: - Prompt Engineering and Ethical Usage: Crafting deceptive prompts that mimic real-world scenarios can manipulate the responses of LLMs. Ethical considerations are crucial in conducting such research. - Exploiting LLMs through Deception: Demonstrated that LLMs can be manipulated through deception, leading to compromised outputs when persuasion principles like authority, trust, and social proof are exploited. - Comparative Analysis: Compared the vulnerabilities of different AI models (e.g., ChatGPT, BARD, Claude, Llama2) to deceptive prompts, providing insights into their security profiles and weaknesses. 1. Impact of AI Adoption:	Experiment	Maybe introduction?

Presuel, R. C., & Sierra, J. M. M. (2024). The adoption of artificial intelligence in bureaucratic decision-making: A Weberian perspective. Digital Government, 5(1), 1–20. https://doi.org/10.1145/3609861	Low, might be useful to make a point. But not worth more then a sentence for my topic	Adoption of AI technologies in public administration.	1. Impact on public administration efficiency, bias reduction, and decision-making quality. 2. Trust in government and perception of legitimacy among citizens.	Not mentioned	Implementation strategy: cautious and informed vs. rapid and less considered adoption.	1. Impact of AI Adoption: - Efficiency and Bias: AI technologies are perceived to potentially increase bureaucratic efficiency and reduce human error. However, concerns are raised regarding AI's ability to eradicate bias and discrimination, as it may inadvertently perpetuate biases present in training data. - Public Trust and Legitimacy: Blind adoption of AI without careful consideration of its consequences can erode public trust in government. When AI implementations do not align with citizen interests or rights, it undermines the legitimacy of government actions. 2. Weberian Perspective: - Ideal Bureaucracy: Weberian bureaucracy emphasizes formal rules, hierarchical management, and efficiency. While AI adoption might initially appear to align with these principles, blind adoption contradicts Weber's preference for careful consideration and formalized decision-making processes. - Guidance for AI Adoption: The study suggests that Weberian principles advocate for a cautious and scrupulous approach to AI adoption in public administration. Governments should prioritize careful consideration of AI's implementation to mitigate potential negative impacts on citizen rights and democratic processes. 3. Policy Recommendation: - Political Choice: Governments should avoid technological solutionism approaches and resist the temptation to transfer accountability solely to AI technologies. Decision-makers need to retain responsibility and ensure that AI deployments are scrutinized for their political implications and potential societal harm. 4. Conclusion: The study concludes that public administrations should heed Weberian principles in adopting AI algorithms, emphasizing careful evaluation and discussion of risks	Review	Counterpart to AI bias? Like assumption that AI reduces bias is stated here but other article talks about AI bias. Could work together.
Gravett, W. H. (2023). Judicial Decision-Making in the age of artificial Intelligence. In Law, governance and technology series (pp. 281–297). https://doi.org/10.1007/978-3-031-41264-6_15	Medium/High. Good for perceived reliability of AI	Adoption and use of algorithmic risk-assessment tools in the criminal justice system.	Dependent Variables (DVs): 1. Quality and fairness of judicial decisions. 2. Trust in the criminal justice system. 3. Impact on defendant outcomes.	1. Training of judges about automation bias. 2. Procedural safeguards and accountability measures in place. 3. Level of scrutiny and oversight of algorithmic tools.	Judges' understanding and awareness of the limitations and biases of algorithmic tools.	Conclusion: - The study highlights the prevalent techno-optimism and over-reliance on algorithmic systems in the criminal justice system, which can shield these systems from necessary scrutiny. - Judges often lack understanding of how automated risk-assessment tools work, leading to potential misapplication and over-reliance on these tools. - Training judges about automation bias and implementing procedural safeguards are essential to ensure fairness and accuracy in judicial decisions. - Algorithmic accountability requires continuous evaluation, transparency, and external audits to detect and mitigate biases. - Human oversight is crucial to maintain transparency and accountability in sentencing decisions, as algorithms may not fully capture individual case nuances. - Ethical and normative considerations are essential when integrating algorithmic tools into the criminal justice system to ensure decisions are fair and just.	Review	Nice comparison between human judgment and AI judgment

Shamim, S., Yang, Y., Zia, N. U., Khan, Z., & Shariq, S. M. (2023). Mechanisms of cognitive trust development in artificial intelligence among front line employees: An empirical examination from a developing economy. *Journal of Business Research , 167* , 114168. https://doi.org/10.1016/j.jbusres.2023.114168	High. Varifies 2 variables of the cognitive trust model and adds the new variable flexibility	1. AI transparency 2. AI reliability 3. AI flexibility 4. AI-driven disruption in work routines 5. Effectiveness of data governance	Cognitive trust in AI	Trust in data governance	AI-driven disruption in work routines	1. Positive Relationships: - Cognitive trust in AI is positively related to AI transparency (β = 0.24, p < 0.001). - Cognitive trust in AI is positively related to AI reliability (β = 0.19, p < 0.001). - Cognitive trust in AI is positively related to AI flexibility (β = 0.09, p < 0.05). 2. Negative Relationship: - AI-driven disruption in work routines is negatively related to cognitive trust in AI (β = -0.11, p < 0.001). 3. Mediation Effect: - Trust in data governance mediates the relationship between the effectiveness of data governance and cognitive trust in AI. 4. Moderation Effect: - AI-driven disruption in work routines negatively moderates the relationship between AI flexibility and cognitive trust in AI (β = -0.08, p < 0.05). 5. Non-Significant Moderation Effects: - AI-driven disruption in work routines does not significantly moderate the relationships between cognitive trust in AI and AI transparency (β = 0.05, p > 0.05) or AI reliability (β = -0.04, p > 0.05).	Mixed-Method experiment. Study 1= interview study. Study 2: Survey study	Very useful - varifies two varaibles
Gkinko, L., & Elbanna, A. (2023). Designing trust: The formation of employees' trust in conversational AI in the digital workplace. Journal of Business Research, 158, 113707. https://doi.org/10.1016/j.jbusres.2023.113707	High. It highlights the importance of organisational context	Trust in the AI Chatbot	Kind of trust (emotional, cognitive, organisational)	User Engagement and Interaction: The extent and quality of user engagement and interaction with the AI chatbot	Previous Experience with Similar Technologg	Results - Employees experienced three types of trust towards the AI chatbot: emotional, cognitive, and organizational. - Emotional trust allowed employees to feel a personal bond with the chatbot, forgiving its errors and continuing its use despite initial performance issues. - Cognitive trust was based on the chatbot's reliability, transparency in its information sources, and its learning capabilities. - Organizational trust was influenced by the organization's endorsement of the chatbot and its security measures. - The combination of these trusts led to sustained use of the chatbot, providing critical use data that improved its performance over time.	Interview study	useful empirical evidence
Anderson, A. A., Jefferson, B. A., Kincic, S., Wenskovitch, J. E., Fallon, C. K., Baweja, J. A., & Chen, Y. (2023). Human-Centric Contingency analysis metrics for evaluating operator performance and trust. IEEE Access, 11, 109689–109707. https://doi.org/10.1109/access.2023.3322133	High. Useful to show the importance of understandability and transparency for trust	AI-based recommender tool (specifically vPred-RC)	Human-machine trust and workload of power system operators.	System state penalty metric: Measures the total number and severity of violations after a contingency, reflecting how human operators perceive violations. Control actions penalty metric: Measures the cost and risk associated with control actions taken by human operators to mitigate violations.	Not given	1. Trust in AI-based Tool: - Transparency and understandability improvements alone did not significantly improve trust among expert power system operators. - The type of recommendations made by the AI tool needed to align closely with typical operational procedures and decision-making processes used by operators to be trusted. - AI recommendations that aligned with operator procedures were more likely to be trusted, even if the AI suggested actions (like load shedding) that were perceived as unnecessary by the operator. 2. Introduction of Penalty Metrics: - System State Penalty Metric: Provides an accurate measure of violation severity from the perspective of human operators, helping to align AI recommendations with operator perceptions. - Control Actions Penalty Metric: Measures the cost and risk of control actions taken by operators, including both continuous (generation redispatch, load shedding) and discrete (tap changes, breaker actions) actions. - These metrics were designed to mimic human operator criteria for evaluating contingency severity and control action effectiveness. 3. Impact on Operator Behavior and Performance: - Operators using AI-generated recommendations tended to take slower but more effective mitigation actions, even when they initially disagreed with the AI's suggestions. - Learning effects were observed within scenarios, where repeated exposure improved operator proficiency in handling specific contingencies. - However, proficiency gains did not transfer significantly between different scenarios, indicating limitations in current training practices for new disturbances on a decarbonized grid.	Experiment	good article
Theis, S., Jentzsch, S., Deligiannaki, F., Berro, C., Raulf, A. P., & Bruder, C. (2023). Requirements for explainability and acceptance of artificial intelligence in collaborative work. In *Lecture notes in computer science* (pp. 355–380). https://doi.org/10.1007/978-3-031-35891-3_22	High. Was used for the varaible transparenq	AI systems designed for explainability and acceptance in human-AI interaction scenarios across various domains (e.g., healthcare, air traffic control).	Information need for explainability, Information need for acceptance, Information representations and interaction methods	User expertise, context of use,	Explanation Capability of the AI	Results and Findings: Results: - Model Explanation: Developers require detailed technical information about AI models, including internal operations and data relationships, which can be provided through model-specific XAI methods. - Result Explanation: Non-expert users primarily seek explanations of AI results and behaviors to understand why certain decisions were made. They prefer explanations that are understandable and not overly technical. - Characteristics of Understandable Explanations: Effective explanations support logical reasoning and are presented in a contrasting manner to facilitate understanding across user groups. - Application-Specific Needs: Different user groups (e.g., investigators, operators) have specific information needs tailored to their tasks and domains (e.g., health decisions, vehicle operations). 2. Acceptance Requirements: - Goal-Supporting Information: Users require information that supports the usefulness and performance of AI systems in achieving their goals. This includes understanding system functions and performance limitations. - Reliability Information: Acceptance is enhanced by providing quality indicators (e.g., error margins, uncertainties) and ensuring transparency about data usage and privacy protection.	Review	Very useful. Showed importance of explainability.
Tejeda, H., Kumar, A., Smyth, P., & Steyvers, M. (2022). AI-Assisted Decision-making: a Cognitive Modeling Approach to Infer Latent Reliance Strategies. Computational Brain & Behavior/Computational Brain & Behavior, 5(4), 491–508. https://doi.org/10.1007/s42113-022-00157-y	High. It provided empirical data on reliability of AI and how it influences AI users	1. Noise Levels: The level of image noise in the task. 2. AI Classifier Accuracy Levels: The accuracy of different AI classifiers (Classifiers A, B, and C). 3. AI Assistance Conditions: Whether AI assistance is provided or not. 4. Advice-Taking Paradigm: The type of	Human Performance: Measured by the accuracy of participants in making decisions.	Participant Confidence: The confidence level of participants in their own decisions. Classifier Confidence: The confidence level of the AI classifier's recommendations.	**Advice-Taking Policies**: The strategies that participants adopt for taking AI advice, inferred through cognitive modeling	the study demonstrated that participants could effectively utilise AI assistance to improve decision-making accuracy, with their reliance strategies being influenced by their own confidence levels, the AI's confidence, and the accuracy of the AI's recommendations. The findings also highlighted the importance of feedback in enabling participants to develop effective reliance strategies	Controlled experiment	
Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. Big Data & Society, 6(1), 205395171986054. https://doi.org/10.1177/2053951719860542	High. Used for transparency variable	The requirement for transparency in AI and automated decision-making systems under GDPR serves as an independent variable. It dictates that AI systems must provide clear information and explanations about their decisions and processing of data.	This variable measures the extent to which transparency, as mandated by the GDPR, achieves its intended goals. It includes aspects such as user understanding, trust in AI systems, and compliance with regulatory requirements.	1. Performative Aspects: How transparency is enacted or performed within AI systems and human-computer interactions. 2. Human-Computer Interaction (HCI) Literature: Insights from HCI research that influence how users perceive and interact with transparent AI systems. 3. Human-Robot Interaction (HRI) Literature: Similar to HCI, studies in HRI inform the understanding of transparency in interactions involving robots and AI systems. 4. Ethical Underpinnings: Ethical considerations regarding the fairness, accountability, and trustworthiness of AI systems.	Regulatory and policy contexts. These include: - Legal Frameworks: Such as GDPR or other data protection regulations that define the scope and requirements of transparency. - Policymaking Processes: How policymakers interpret and implement transparency requirements into practical guidelines and standards.	contribute to understanding how transparency in AI systems is conceptualized, implemented, and evaluated in legal, social, and ethical dimensions. The actual results were to long to list here, and the conclusion of the study were future research directions that need to be done.	Review	Research did not fit in this format of different varables. But it was very nice for transparency
Borsci, S., Malizia, A., Schmettow, M., Van Der Velde, F., Tariverdiyeva, G., Balaji, D., & Chamberlain, A. (2021). The Chatbot Usability Scale: the Design and Pilot of a Usability Scale for Interaction with AI-Based Conversational Agents. Personal and Ubiquitous Computing, 26(1), 95–119. https://doi.org/10.1007/s00779-021-01582-9	Low. While interesting, probably not usable since it is about a specific scale designed to increase user satisfaction. Not useful to differentiate trust and compliance	use and design of CRM chatbots (various attributes and fuctionalities designed to enhance user interaction and satisfaction)	User satisfaction with CRM chatbots - can it be increased through using BUS-15 scale?	BUS-15 scale variables - sorted into 5 factors (Perceived accessibility to chatbot functions, Perceived quality of chatbot functions, Perceived quality of conversation and information provided, Perceived privacy and security, Time response)	User profile (Age, gender, ability) are suspected, but would need further research	BUS-15 can be used to increase user satisfaction	four studies (systematic literature review, survey (experts and novices), focus group sessions, testing of chatbots (experiment))	Very nice empirical data - just unsura about the usability in my study. - edit: was not used