Trustworthy Artificial Intelligence: The Impact of Certification Labels on End-Users' Trust and Intention to Use

Joana Glaum

Dissertation written under the supervision of Dr. Filipa de Almeida

Dissertation submitted in partial fulfilment of requirements for the M.Sc. in Management with Specialization in Strategic Marketing, at the Universidade Católica Portuguesa, 02.01.2024

Abstract

Title: Trustworthy Artificial Intelligence: The Impact of Certification Labels on End-Users' trust and Intention to Use

Author: Joana Marie Glaum

With the growing influence of Artificial Intelligence (AI) systems, concerns about their trustworthiness have emerged. Therefore, this dissertation explores the critical issue of trust in AI and introduces certification labels as a method to enhance end users' trustworthiness perceptions of AI. To uncover whether a certification label for trustworthy AI (TAI) has the potential to increase end-users' trust in and the acceptance of AI systems, two experimental studies were conducted in the scope of this research.

Study 1 found that a certification label for TAI can have a positive impact on end-users' trust in and the acceptance of AI systems. Study 2 identified that transparently communicating the requirements used for the certification on the certification label translates into higher levels of trust in the AI system. Moreover, it was examined that the requirement *Transparency* was perceived as most important regarding AI trustworthiness.

The results from these studies have clear implications for policymakers, developers, and organizations that seek to enhance the trustworthiness of AI, suggesting that certification labels for TAI are an effective method to communicate trustworthiness of AI systems to end-users and thereby to increase the acceptance of their AI-based products and services. Considering the potential competitive advantage of embedding TAI in products and services, the dissertation underscores the relevance of TAI in shaping the future landscape of AI technologies. Further, the findings also complement the knowledge on labels (in general) and the effectiveness of different label designs.

Keywords: Artificial Intelligence, Trustworthy AI, Certification, Trust, Cognitive and Affective Trust, Requirements for TAI, AI Acceptance, Ethical Guidelines, AI Auditing, Labels

Sumário

Título: Inteligência Artificial de Confiança: O impacto dos rótulos de certificação na confiança e na intenção de utilização dos utilizadores finais

Autor: Joana Marie Glaum

Com a crescente influência da Inteligência Artificial (IA), surgiram preocupações quanto à sua fiabilidade. Por conseguinte, esta dissertação explora a questão crítica da confiança na IA e introduz rótulos de certificação como um método para melhorar as percepções de fiabilidade da IA por parte dos utilizadores finais. Para descobrir se um rótulo de certificação para uma IA fiável (IAF) tem potencial para aumentar a confiança dos utilizadores finais e a aceitação da IA, foram realizados dois estudos experimentais.

O estudo 1 concluiu que um rótulo de certificação para uma IAF pode ter um impacto positivo na confiança dos utilizadores finais e na aceitação da IA. O estudo 2 identificou que a comunicação transparente dos requisitos utilizados para a certificação no rótulo de certificação se traduz em níveis mais elevados de confiança na IA. Além disso, verificou-se que o requisito Transparência foi considerado o mais importante no que respeita à fiabilidade da IA.

Os resultados destes estudos têm implicações claras para os decisores políticos, os criadores de IA, e as organizações, sugerindo que os rótulos de certificação para as IAF são um método eficaz para comunicar a fiabilidade da IA aos utilizadores finais e, assim, aumentar a aceitação da IA. Tendo em conta a potencial vantagem competitiva da incorporação de IAF em produtos e serviços, a dissertação sublinha a relevância das IAF na definição do futuro panorama da IA. Além disso, os resultados também complementam o conhecimento sobre os rótulos (em geral) e a eficácia das diferentes concepções de rótulos.

Palavras-chave: Inteligência artificial, IA fiável, certificação, confiança, confiança cognitiva e afectiva, requisitos para a TAI, aceitação da IA, orientações éticas, auditoria da IA, rótulos

Acknowledgments

This dissertation marks the last step of an exciting chapter in my personal life and educational career, which has been an amazing journey of continuous learning and personal growth. The experience of restarting in a foreign country, building new friendships, and overcoming various challenges has been instrumental in shaping my preparation for the forthcoming phase in my life. I am deeply grateful to all those who supported me along the way, without whom this achievement would not have been possible.

First and foremost, I would like to thank my supervisor Filipa de Almeida, who guided me through this thesis by sharing her expertise, ideas and experiences. I thank you for always taking the time to answer my questions, to clarify my doubts, and to provide insightful feedback. By always encouraging me to proceed and to challenge my work, you have not only brought my thesis to a greater level, but also made this time an exciting new experience.

I would also like to thank my friends for always being there for me and making this journey unforgettable. You made it possible for me to make Lisbon feel like a second home away from Germany. Finally, I could not have completed this academic degree without the infinite support of my family who always believed in me.

Thanks to all of you!

Table of Contents

List of Figures

List of Abbreviations

&	And
AI	Artificial Intelligence
ANOVA	Analysis of Variance
b	Regression coefficient
df	Degrees of freedom
F	F-statistic
H1	Hypothesis 1 (2-5 respectively)
M	Sample mean
N	Total number of cases
P	p-value
R^2	Multiple correlation squared; measure of strength of association
SD	Standard Deviation
SE	Standard Error
t	t-statistic

Model 14 of Hayes PROCESS macro for SPSS

M1	Mediator 1 (2 respectively)
X	Independent variable
Y	Dependent variable
CI	Confidence Intervall

1. Introduction

"Don't trust AI until we build systems that earn trust."

Gary Marcus, coauthor of "Rebooting AI" (The Economist, 2019).

1.1 Topic presentation

It is widely acknowledged that Artificial intelligence (AI) is more than just a technology. AI has emerged as a transformative force, having a significant impact on society. It has proven its potential to transform businesses, reshape entire industries, augment human capabilities, and revolutionize the way we interact with technology (Lukyanenko et al., 2022). However, as AI systems become increasingly integrated into our daily lives, concerns about their trustworthiness have taken center stage. While some of these arise from the AI's autonomous nature, the system's black-box character adds to uncertainty (Choung et al., 2023). Further concerns of AI revolve around safety, accountability, and responsibility (Morik et al., 2021), as well as potential impacts on employment and privacy (Choi, 2023). These risks have caused a growing sense of distrust among AI users, adopters, and policymakers, holding back AI's development and acceptance, and limiting its full potential (Crockett et al., 2021). The European Commission's High Level Expert Group on AI (AI HLEG; 2019) emphasizes that it is precisely this (dis)trust that plays a crucial role in harnessing the potentials of AI, while simultaneously reducing or potentially eliminating its risks. Thus, to achieve its full positive potential and gain widespread acceptance, AI systems must be trustworthy (Kaur et al., 2022).

1.2 Relevance of the topic

In response to the growing importance of AI, the concept of Trustworthy Artificial Intelligence (TAI) has received heightened interest within research. TAI underlines the importance of establishing trust in the development, deployment, and utilization of AI (AI HLEG, 2019; Kaur et al., 2022; Thiebes et al., 2021). Thus, ensuring that AI systems are trustworthy, as well as building trust in AI stands as a prominent priority on current political, business, social, and legal agendas (AI HLEG, 2019). In recent years, several frameworks and (ethical) guidelines have been proposed by researchers, organizations, industry, and policymakers striving towards making AI trustworthy (Hagendorff 2020; Jobin et al., 2019). The importance of developing AI systems within a trustworthy framework (Kaur et al., 2022) is also reflected in current statistics, indicating that by

2025, approximately 30% of AI-driven digital products will demand the adoption of such a framework (Burke et al., 2021), and 86% of users are projected to place their trust in and stay loyal to companies that adhere to ethical AI principles (Edelman, 2019). However, the question of how these principles can be implemented in practice is now increasingly attracting attention from researchers (Brundage et al., 2020; Mittelstadt, 2019; Morley et al., 2020). More importantly, the communication of an AI's trustworthiness to various stakeholders is key in building trust in it (Liao & Sundar, 2022). In this context, AI auditing has captured interest in literature as it plays a crucial role in enabling trust in AI by ensuring that the principles of TAI are met (Avin et al., 2021; Knowles & Richards, 2021; Toreini et al., 2020). However, it seems that end-users do not have the required expertise and understanding to assess various trustworthiness principles, resulting in information asymmetries (Morik et al., 2021). This leaves open an important question: How can trustworthiness of AI systems be communicated to end-users (Morik et al., 2021)?

1.3 Problem Statement and Research Objective

Motivated by this challenge, this thesis aims to contribute to the present discussion around the significance of TAI and presents certification as a concrete method to address the previous question. Specifically, the aim of this research is to study the effects of certification labels on end-users' perceived trustworthiness of AI systems. While some researchers (e.g. Holland et al., 2020; Scharowski et al., 2023; Seifert et al., 2019; Stuurmann & Lachaud, 2022) and entities such as the European Commission's AI HLEG (2019), suggested certification as a non-technical method to communicate trustworthiness of AI, empirical evidence and comparative studies on the effectiveness of labels remain limited until now. In light of the identified research gap, the central research problem addressed in this thesis lies in empirically determining the role of certification labels in communicating system trustworthiness and thereby enhancing end-users' trust in AI systems. Likewise, and in line with the literature (Scharowski et al., 2023; Stuurman & Lachaud, 2022), it is crucial to conduct research that further looks at the effectiveness of different label designs in helping consumers determine an AI system's trustworthiness. In particular, the amount of information displayed on a label is a topic of discussion. Thus, this dissertation attempts to examine how communicating information about trust requirements on labels influences the perceived trust in the AI. Furthermore, the importance of different requirements for TAI will be assessed from the end-users' perspective to understand whether some requirements are more important than others and to inform future studies about which requirements to communicate on a

label. Consequently, this dissertation intends to provide an answer to the following research question:

Research question (RQ): Does a certification label for TAI have the potential to increase end-users' trust in AI technologies?

To address the identified research gap, the central research question was divided into the following four sub-questions:

RQ1: Are certification labels an effective way to communicate trustworthiness of AI systems to end-users?
RQ2: How do certification labels impact end-users' trust in the AI system?
RQ3: Are AI systems with a certification label for trustworthiness more accepted by end-users?
RQ4: What influence, if any, does displaying ethical requirements for TAI on certification labels have on communicating trustworthiness of AI systems to end-users?

To answer these questions two experimental studies were conducted. While Study 1 tests the effect of certification labels on trust in and the acceptance of AI technologies, Study 2 expands on the first study by examining whether displaying information regarding the requirements of TAI on the certification label impacts the perceived trustworthiness of an AI system. The findings of these studies aim to achieve several objectives: Firstly, to contribute to ongoing discussions regarding the significance of TAI, and trust in AI, and secondly, to contribute to the literature on labels' impact on consumer behavior in a broader sense. Moreover, the results will also provide practical insights for policymakers, developers, and organizations seeking to enhance the trustworthiness of AI systems. This is becoming increasingly relevant as embedding TAI in products and services could potentially represent a competitive advantage for producers of AI systems in the future (AI HLEG, 2019).

1.4 Structure of the Dissertation

Following this introduction, Chapter 2 lays the theoretical foundation for this thesis by reviewing and discussing existing literature on the concepts of TAI and certification. In this thesis, two empirical studies were conducted to gather data to answer the research questions. Therefore, the methodology that was used in each study is described in Chapters 3 and 4. In Chapter 5, the main findings of the studies are analyzed and discussed in light of existing literature, leading to the

identification of theoretical and practical implications. In the second part of this chapter, limitations are pointed out and an outlook for further research is provided. Finally, Chapter 6 provides a conclusion for the thesis.

2. Theoretical Background

The following chapter elaborates on the concept of TAI as well as certification, both providing essential theoretical background knowledge for this thesis. By combining both concepts, this chapter further outlines the relevance of certification for communicating trustworthiness of AI systems to end-users and thereby offers insights into the potential impact of certification labels on trust and acceptance of AI systems.

2.1 Trustworthy Artificial Intelligence (TAI)

2.1.1 AI and users' concerns

AI is a broad field encompassing computer systems capable of performing tasks that previously demanded human intelligence, including visual perception, speech recognition, and decision making under uncertainty (Rossi, 2018; Russell & Norvig, 2010). Thus, what makes AI unique is its human-like capabilities (Krafft et al., 2020). This perspective is also consistent with Gillath et al.'s (2021) definition of AI, referring to it as the imitation of human intelligence functions like learning, reasoning, and self-correction by machines, specifically computer systems. Nowadays, examples for AI range from personal assistants such as Siri to medical diagnostic tools and self-driving cars (Gillath et al., 2021), as well as generative AI tools, like ChatGPT (Baidoo-Anu & Ansah, 2023). This thesis focuses on the definition of AI formulated by the Organization for Economic Cooperation and Development (OECD) as this represents the basis for the EU's proposal for the establishment of a legal definition of 'AI system' in EU law (EU, 2023). According to the OECD (2019, p. 7), an 'artificial intelligence system' is a machine-based system designed to produce outputs such as content, predictions, recommendations, or decisions, based on a given set of human-specified objectives, thereby impacting the environments it engages with.

By automating tasks that previously required human intelligence, AI systems have significantly disrupted our lives (Lewis et al., 2020). With the growing influence of AI systems, concerns about their trustworthiness have emerged among AI users, adopters, and policymakers (Crockett et al., 2021). These are mainly caused by the AI's capability for autonomous functioning whereby,

compared to traditional technologies, users do not have complete control over their operating (Choung et al., 2023). Moreover, the reliance of AI on machine learning algorithms causes unpredictability and uncertainty, as they have been characterized as black boxes. This is due to the complexity of the algorithm's inner workings and thereby the lack of the user's understanding regarding the learning and reasoning behind the decision-making of AI systems (Choung et al., 2023). This complexity not only causes skepticism and distrust, but also hinders the development and deployment of AI and thereby the realization of its full potential and continued success (Lukyanenko et al., 2022). Moreover, many are concerned about AI's possible impact on job security and privacy, highlighting the necessity for transparent and responsible development and application of AI systems (Choi, 2023). Long-term consequences of progress in AI are also a reason for concerns and distrust because by possibly outperforming human abilities and thereby enabling possibilities that are beyond current comprehension, AI is feared to emerge from being just another technology to a precursor to superintelligence (Bostrom, 1998; Harari, 2016; Yampolskiy, 2015).

2.1.2 The concept of TAI

As a result of the growing recognition of AI-caused challenges, an expanding body of literature recognizes the concept of TAI. According to the European Commission's AI HLEG (2019), TAI emphasizes the relevance of establishing trust in the development, deployment, and utilization of AI for it to gain widespread acceptance and achieve its full positive potential. The perceived trustworthiness of AI systems among its users, including consumers, organizations, and society, can be achieved by ensuring that it is developed, deployed, and employed in manners that guarantee legal compliance and robustness. But more importantly, AI systems should demonstrate adherence to universal ethical principles (AI HLEG, 2019).

When it comes to the term TAI, Lee and See (2004) found that it is crucial to acknowledge a fundamental distinction between actual trustworthiness and perceived trustworthiness, primarily due to the time-relative property inherent in AI systems. The levels of actual trustworthiness in AI systems are often unknown to users because they lack the capability to comprehensively assess it. Instead, users rely on the perceived trustworthiness, which they assess through their interactions with the system. Since these perceived trustworthiness levels often deviate from the actual levels of trustworthiness, the challenge lies in matching the end-user's trust to the degree of actual trustworthiness of an AI and enabling appropriate trust in the system (Lee & See, 2004).

This thesis, while centered on "TAI", aligns with other concepts in recent literature that refer essentially to the same goal of maximizing the advantages of AI while mitigating or preventing its risk. Notably, there is a vast body of research on explainable AI, focusing on transparency as the crucial element for better understanding AI's processes and outputs (Adadi & Berrada, 2018). It emphasizes the effort of researchers to turn the black-box nature of an AI into a transparent system by focusing on explaining the reasoning behind an AI's decision or recommendation (Xu et al., 2019).

This thesis focuses on TAI as a framework designed to assure the trustworthiness of a system based on the adherence to its specified requirements. TAI operates to confirm that the expectations of users and stakeholders are fulfilled in a verifiable manner (International Organization for Standardization ((ISO), 2020). While actual trustworthiness is clearly important, perceived trustworthiness shall not be ignored, as it has been shown to have serious consequences in human-AI interaction (Lee & See, 2004).

2.1.3 Trust in AI and its role in the acceptance of AI

The uncertainties and possible risks connected to AI's decision-making highlight that there is a crucial factor influencing the acceptance and usage of AI systems. This factor is trust (Choung et al., 2023). Researchers like Nordheim et al. (2019) emphasized the importance of trust in reducing doubts and perceived risks when adopting new technologies. A similar effect was found by Yang and Wibowo (2020) who argue that trust is a crucial factor in helping users conquer their uncertainties and concerns, contributing to the acceptance of emerging technologies like AI. Moreover, several researchers have extended the widely used technology acceptance model (TAM), proposed by Davis (1985), by adding trust as a predictor for the acceptance of AI systems (Alalwan et al., 2018; Egea & González, 2011; Salloum & Al-Emran, 2018).

In various research areas, trust is thought of as a belief formed through the evaluation of certain attributes of an object (Colquitt & Rodell, 2011; McKnight et al., 2002). This view is also shared by Lee and See (2004), according to whom peoples' trust in AI is shaped by their perception of information about the system's trustworthiness characteristics and their preexisting views. This definition of trust makes it clear that trustworthiness is not automatically present, but rather must be communicated and recognized as such by the user. A possibility to do so is through the use of transparency cues. Humans then use heuristics, which are mental shortcuts, to assess these cues to

form judgments about the AI's trustworthiness. In this context, information display and the user's perception of it play a crucial role in forming appropriate trust judgements, highlighting the importance of effective communication of system trustworthiness (Lee & See, 2004; Liao & Sundar, 2022). The relevance of an AI's trustworthiness is also emphasized in Jacovi et al.'s (2021) definition of human-AI trust, according to whom trustworthiness is the source of trust.

In the context of AI, both trust in people (Mayer et al., 1995) and trust in technology (Mcknight et al., 2011) are relevant due to the combination of AI's functionality with human-like competencies (Krafft et al., 2020). Consequently, trustworthiness can be evaluated by using the three dimensions, as modified from the interpersonal trusting beliefs: (1) Ability – having the skills and competence to perform a task; (2) Benevolence – displaying a caring and thoughtful manner characterized by goodwill; (3) Integrity – adhering to a set of mutually acceptable principles for behavior (Mayer et al., 1995; Liao & Sundar, 2022). The ability, benevolence and integrity (ABI) model has demonstrated its versatility across a range of disciplines (Bhattacherjee, 2002; Gefen & Straub, 2004; McKnight et al., 2002), and is frequently employed to determine the trustworthiness of technologies (Mazey, 2018; McKnight et al., 2011; Xu et al., 2014). Drawing from findings of previous research (Schumann et al., 2012), ABI can be mapped into the cognitive (ability) and affective (benevolence, integrity) trust dimensions which were first introduced by Johnson and Grayson (2005). The former dimension, cognitive trust, is knowledge-driven and emerges when the AI system demonstrates its competence and reliability. It is based on the user's rational evaluation of the AI's performance and features. The latter dimension, affective trust, refers to the emotional bond between the user and the system (Johnson & Grayson, 2005). It relates to the social and cultural values of the algorithms, including the ethical criteria on which the design of AI systems is based (Chen & Sundar, 2023). The dimension "behavioral intention to use" that is used in several models to assess the acceptance of AI (e.g. TAM (Davis, 1985); AI device use acceptance (AIDUA; Gursoy, 2019); Unified Theory of Acceptance and Use of Technology (UTAUT; Venkatesh et al., 2003)) completes the framework used for this thesis (See Figure 1). Since behavioral intention has widely been characterized as an individual's subjective probability to engage in a particular behavior (e.g., Fishbein & Ajzen, 1975), in this case to use the AI, it will serve as measurement for assessing the acceptance of AI in this thesis.

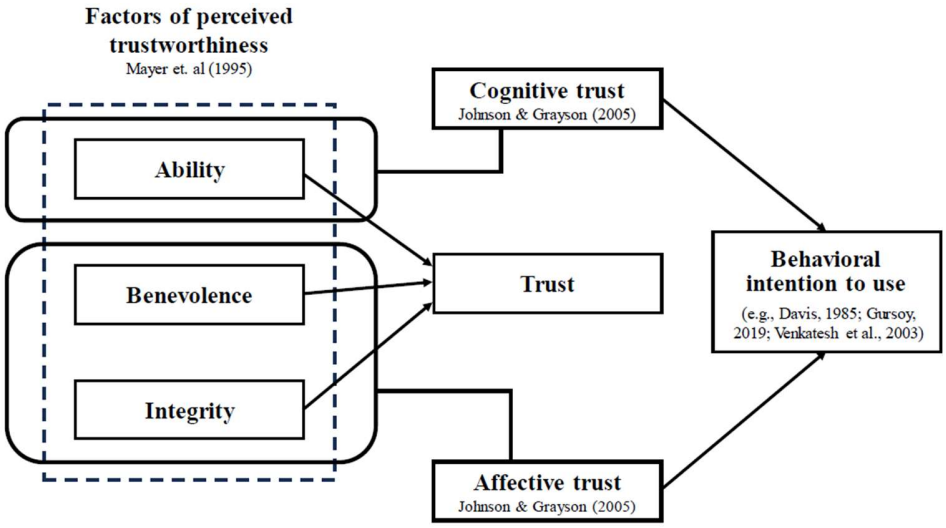

Figure 1: Trust & acceptance measurement framework

While the key role of trust in enhancing the acceptance and usage of AI systems has been attested by several researchers (Ghazizadeh et al., 2012; Hoff & Bashir, 2015; Lee & See, 2004; Pavlou, 2003), the question of how trust in AI can be established remains a topic of debate. Ensuring the trustworthiness of AI, which refers to the elements necessary for people to trust it (Varshney, 2019), is seen as a practical aspect of implementing ethical AI principles (Toreini et al., 2020). Therefore, the following subchapter addresses requirements for building trust in AI systems by introducing the TAI framework, proposed by the European Commission's AI HLEG (2019).

2.1.4 Requirements for TAI

Given its human-like manifestations and its growing influence on people's lives, AI systems now demand a broader socio-technical understanding that goes beyond mere functionality. This understanding should rely on human trust and must comply with ethical values (Choung et al., 2023). Along with this emerging understanding of AI, also the requirements that make AI trustworthy have developed (Thiebes et al., 2021). In recent years, the growing influence of AI and the associated concerns have led to the proposal of various frameworks and principles for developing and deploying TAI. These guidelines, published by researchers, tech giants like Google, Microsoft, and IBM, and organizations like the OECD and the Institute of Electrical and Electronics Engineers (IEEE), as well as policymakers and governments such as the EU, exemplify the broad interest in TAI (Hagendorff 2020; Jobin et al., 2019).

A succinct framework of requirements for TAI was proposed by the European Commission's AI HLEG in 2019. Considering the scope of this thesis, this framework and the related requirements shall serve as criteria for TAI in the context of this work, as it covers the most commonly proposed aspects necessary for building trust in AI systems in the literature and is one of the first frameworks from a governmental organization with practical relevance for EU cititzens. According to the AI HLEG's (2019) TAI framework (Figure 1), three components of TAI, which must be met throughout the entire duration of the system's life cycle, can be defined. Firstly, TAI should be lawful, meaning it should operate within the bounds of all relevant laws and regulations. Secondly, compliance to ethical principles and values is essential. Lastly, TAI should demonstrate technical and social robustness, acknowledging that AI systems, despite good intentions, can unintentionally cause harm. Derived from essential human rights, including the recognition of human autonomy, harm prevention, fairness, and explicability, the framework by the European Commission's AI HLEG's (2019) offers seven key requirements for the development, deployment and use of AI systems in order to be considered trustworthy: (1) *Human agency and oversight,* emphasizing the adherence to fundamental rights; (2) *Technical robustness and safety,* containing resilience to potential attacks and security, as well as general safety, accuracy, reliability, and reproducibility; (3) *Privacy and data governance*, involving respect for privacy, quality and integrity of as well as access to data; (4) *Transparency*, with the key aspects of traceability, explainability, and communication; (5) *Diversity, non-discrimination and fairness*, highlighting the importance of avoiding unfair and biased outputs, accessible and universal design, and stakeholder participation; (6) *Environmental and societal well-being*, entailing aspects such as sustainability, environmental friendliness, social impact, and their implications for society and democracy; and (7) *Accountability*, emphasizing auditability, minimization and reporting of negative impact, trade-offs and remediation (AI HLEG, 2019).

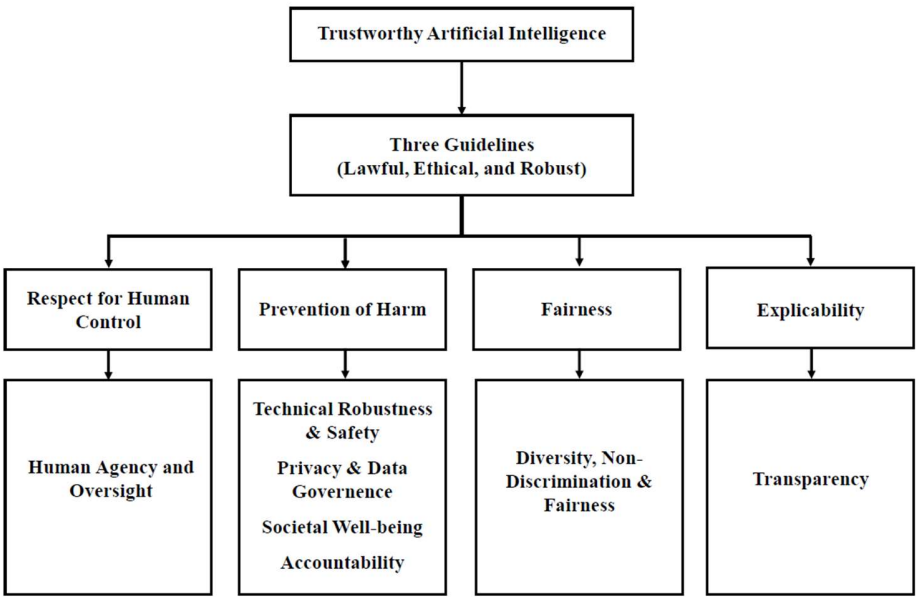

Figure 2: TAI framework (Based on Kaur et al., 2020; Kumar et al., 2020)

While the principles and corresponding requirements outlined for TAI hold significant value, it is essential to acknowledge a fundamental limitation associated with them. According to the AI HLEG (2019), the proposed framework and guidelines provide insufficient guidance as they do not explain how to translate the outlined TAI principles into practical implementation. Furthermore, they are not complemented by instructions on how they can shape future research concerning technical and non-technical methods, aimed at promoting the realization of TAI (Thiebes et al., 2021). Thus, in the following, a non-technical method to practically implement the proposed guidelines (AI HLEG, 2019) will be presented: certification.

2.2 Certification

2.2.1 The impact of labels on trust

Labels play a key role in helping consumers make well-informed decisions by featuring specific product or service characteristics (Cihon et al., 2021). They have gained recognition across diverse products and industries, including agriculture (Gorton et al., 2021), food (Jones et al., 2019), energy (Stadelmann & Schubert, 2018), and e-commerce (Thompson et al., 2019). By providing information about a product or service, labels can function as an extrinsic cue for consumers, reducing the uncertainty regarding a product's quality (Binninger & Robert, 2005). Therefore, labels act as a source of transparency and, thereby, trust (Mazzù et al., 2022).

While various types of labels exist, and multiple classification systems have been suggested, this dissertation focuses on certification labels because they are mainly directed at the consumer (Jahn et al., 2005), which is the focus of this thesis: the communication of TAI to end-users.

2.2.2 The process of certification and certification labels

Certification labels certify that a product, procedure, individual, or organization conforms to one or more predefined standards or criteria (ISO, 2004). Well-known and widely used examples for certification labels include for example “fairtrade”, "low-fat" or "organic" labels (Jahn et al., 2005). In their paper, Cihon et al. (2021) decomposed certification into its components. In brief, the process of certification involves an evaluation of the *object of certification*, typically conducted by a trusted third-party, which, upon confirming alignment with the specified criteria, grants the entity official certification accordingly. Those predefined standards or criteria could range from voluntary principles to mandatory regulatory requirements. While the *certifier* is the one who issues the actual certificate, the *assessor* evaluates the conformation of the object to the specified criteria in the first place. The process of determining whether the object of certification meets the predefined standards refers to *assessment* or *conformity assessment* and typically involves an audit of systems or processes (Cihon et al., 2021). Thus, only products that have successfully passed an auditing process, are granted a certification label (King et al., 2005). Following the definition of the IEEE (2008, p. 30), auditing refers to "an independent evaluation of conformance of software products and processes to applicable regulations, standards, guidelines, plans, specifications, and procedures." This process is essential to verify and validate the claims made by these labels regarding the trustworthiness and adherence to specific standards. Consequently, auditing is a pivotal process in the context of certification labels and serves as a critical means of ensuring the reliability and credibility of labels in consumer policy (Taufique et al., 2022). The last component of the certification process is the *audience*, referring to the individuals or group who will receive the information about the certification (Cihon et al., 2021).

2.2.3 Purpose and effects of certification labels

The primary objective of certification is to mitigate information asymmetry in the market by providing insights into the internal operations of the object of certification and sharing these with external stakeholders (King et al., 2005). Past research of Tonkin et al. (2015) has shown the effectiveness of certification labels in communicating the results of audits and thus, enhancing trust in the labeled products. Therefore, the process of certification is used by governments to promote

transparency and to motivate those within the entity to meet established standards (Cihon et al., 2021). In this context, labels are also critical for reducing users' uncertainty towards a product's quality for two reasons: First, they reveal information about the development of the product and secondly, they provide security for end-users (Binninger & Robert, 2005). However, this purpose can only be achieved if the authority that issues the label effectively ensures the quality of certification and, by extension, the reliability of the outcome of the auditing process (Jahn et al., 2005). Therefore, labels are typically assigned by reputable third-party organizations as they provide an institutional assurance of trustworthiness (Tonkin et al., 2015). By overseeing and verifying adherence to standards, the impartial entity ensures that certification labels can promote a product's claims about meeting specific standards and criteria (Taufique et al., 2022). This supports quality and, thereby builds trust with consumers, ultimately shaping the acceptance and credibility of labels (Blair, 2008; Brundage et al., 2020). A further benefit of certification labels is that they are understandable by various stakeholders. Through the use of simple wording, icons and colour coding, information is provided in a clear and accessible way, helping consumers processing it appropriately (Goodman et al., 2018; Grunert et al., 2014). Similar to the effect of brands, labels can impact the consumer's purchase intention. Research on existing label schemes shows that consumers often strongly favor products which are labelled over those which are not (Larceneux, 2004).

2.3 Realizing TAI through certification labels

2.3.1 The challenge: Communicating trustworthiness of AI systems to end-users

As previously outlined, the effective communication of system trustworthiness and its perception as such by the users is essential for appropriate trust (Liao & Sundar, 2022). In light of the importance of an AI's compliance to ethical standards for enhancing trustworthiness, scholars such as Cihon et al. (2021) have raised the question of how external stakeholders can determine whether organizations and their AI systems are adhering to ethical principles. Given the growing influence of AI on society, various stakeholders are involved with AI systems who require various degrees of explanation and communication of trustworthiness due to their different levels of knowledge and needs (Yurrita et al., 2022). Kaur et al. (2022) emphasized that while domain experts for example need more in-depth explanations regarding the decision-outcomes of certain attributes, policymakers demand information about the system as a whole to evaluate its compliance with current laws. End-users, on the other hand, require accessible communication tailored to their

specific concerns. For them, information regarding the reasoning of an AI's decision is needed to evaluate the trustworthiness of the system (Kaur et al., 2022).

Therefore, a growing body of work has recognized the crucial role of auditing in the context of AI to ensure system trustworthiness (Avin et al., 2021; Knowles & Richards, 2021; Toreini et al., 2020) by strengthening key elements like fairness (Wilson et al., 2021), accountability (Costanza-Chock et al., 2022), and governance (Falco et al., 2021), all of which are instrumental in building trust and fostering acceptance of AI technologies. Another notable strength of AI audits is their capacity to reveal problematic behavior within AI systems, as they can uncover issues such as algorithmic discrimination, distortion, exploitation, and misjudgment (Bandy, 2021). While outcomes of AI audits, such as model cards and training datasets, serve as valuable tools for determining whether key principles of TAI have been met; they are primarily tailored for regulators and experts who have extensive knowledge on the topic of AI (Kaur et al., 2022). Likewise, the various proposed requirements for TAI to increase the user acceptance of those systems (Kaur et al., 2022) remain at a high-level, lacking specific metrics and measurement procedures to certify AI-based solutions (Lakkaraju et al., 2020). However, end-users typically lack the necessary expertise, knowledge or measurement mechanism to quantitatively assess the diverse trustworthiness principles (Morik et al., 2021) or do not understand the technical information that AI documentation provides (Arnold et al., 2019). Consequently, such documentations prove ineffective in helping end-users make well-informed judgments about trusting or using AI (Knowles & Richards, 2021). Morik et al. (2021) emphasize that the gap between technical knowledge and user understanding can further prevent trust building, thus impacting the acceptance of AI systems. Although end-users are just as important as developers and providers of AI systems, they are often ignored as a stakeholder group (Morik et al., 2019). Therefore, a prominent challenge lies in bridging this gap by communicating a system's trustworthiness to end-users.

2.3.2 Bridging the gap: Certification labels for communicating TAI to end-users

Given the complexity and impact of AI systems, and hence the challenge of communicating the workings and effects of AI systems to end-users, organizations, auditors or regulators, capable of certifying AI systems as transparent, accountable and fair, can be considered (Morik et al., 2021). By verifying and ensuring the trustworthiness of AI, they can help end-users with assessing system trustworthiness, and guiding their decision-making (Floridi et al., 2022; Zicari et al., 2021). In line with this proposal, several researchers have emphasized the introduction of labels as a form of

certification for AI (Cihon et al., 2021; Holland et al., 2020; Scharowski et al., 2023; Seifert et al., 2019; Stuurmann & Lachaud, 2022). Likewise, the European Commission's (2020) "white paper on AI" proposed the establishment of a voluntary label based on the AI HLEG guidelines for TAI.

Certification labels are particularly relevant in the context of realizing TAI for several reasons. First, they are designed to be easily understandable by various stakeholder groups. This is of particular importance in the case of communicating trustworthiness of AI systems to end-users as they usually lack the knowledge and time to evaluate the trustworthiness criteria (Grunert et al., 2014). By signaling that AI-enabled products and services in question comply to certain objective and standardized requirements, labels can address the previous outlined challenges of end-users regarding the assessment of an AI's trustworthiness. Consequently, this may help enhance the trust of end-users in AI systems (European Commission, 2020). Furthermore, as AI systems are characterized as black-boxes, labels can be beneficial by reducing consumers' uncertainty towards the system's quality (Binninger & Robert, 2005). As labels serve as external sources of validation for assessing system trustworthiness, they can help with matching the end-user's trust to the degree of actual trustworthiness of an AI, thereby enabling appropriate trust in the system (Jacovi et al., 2021; Lee & See, 2004). Consequently, the first hypothesis is as follows:

H1a: *A certification label for TAI can increase end-users' trust in AI systems.*

Moreover, previous studies on the effect of different kind of labels have shown that they can influence both the cognitive and affective dimension of end-users' trust. Research by Aiken and Boush (2006) has found that, in the context of user trust in e-commerce, the presence of a trustmark had an effect on beliefs about privacy and security (affective trust), which in turn influenced more general beliefs about the trustworthiness of the firm (cognitive trust). Furthermore, research on energy efficiency labels for instance has shown that they had a significant effect on the cognitive dimension of trust but a rather low effect on the affective dimension of trust (Schuitema et al., 2020). Drawing from findings of Chen and Sundar (2023), who studied the effect of data credibility on user trust, it can be expected that AI systems with a certification label are likely to be evaluated as competent and reliable (cognitive trust) because they have typically passed an auditing process, conducted by a reputable third-party. Affective trust, on the other hand, might be less affected by the label, as AI auditing is more about evaluation of AI competence and reliability rather than the

system's capacity to convey warmth and caring (Chen & Sundar, 2023). Thus, the following hypothesis is proposed:

H1b: *A certification label for TAI has a greater influence on cognitive than affective trust in AI systems.*

By enhancing the user's assessment of an AI's trustworthiness, certification labels have two main purposes: First, to enhance end-users' trust in the AI and secondly, to advance a more widespread acceptance of it (European Commission, 2020). If a label communicates that the AI is considered trustworthy by a third-party, end-users are likely to associate the AI with lower risks. In turn, the subjective evaluation of the risks and benefits of the expected outcome is likely to influence the behavioral intention to use a system (Ajzen, 1985). This leads to the following hypothesis:

H2*: A certification label for TAI increases end-users' acceptance of (behavioral intention to use) AI systems.*

Since trust is expected to be a predictor of AI acceptance, a further hypothesis can be put forward:

H3a*: A certification label for TAI increases end-users' acceptance of (behavioral intention to use) AI systems though increased cognitive and affective trust in AI.*

As the effect of the certification label is expected to be stronger on cognitive trust than affective trust, this is also predicted to apply to the mediation, leading to a further hypothesis:

H3b: *A certification label for TAI increases end-users' acceptance of (behavioral intention to use) AI systems mainly through increased cognitive trust in AI.*

Since the label is only granted to AI systems that meet certain criteria for TAI, it indicates that the system is trustworthy. This is because each principle establishes specific criteria that collectively contribute to shaping an overall sense of trust in AI (Jacovi et al., 2021; Liao & Sundar, 2004). Therefore, transparently communicating the criteria used in the underlying AI auditing process can enhance the trust-building in the AI system among end-users (Stuurman & Lachaud, 2022). Based on the previous findings, the following hypothesis can be derived:

H4*: When a certification label is based on requirements for TAI, end-users will have higher levels of trust in the respective AI system than in the same AI with a "simple" certification label.*

Therefore, the first conceptual model looks as shown in Figure 2.

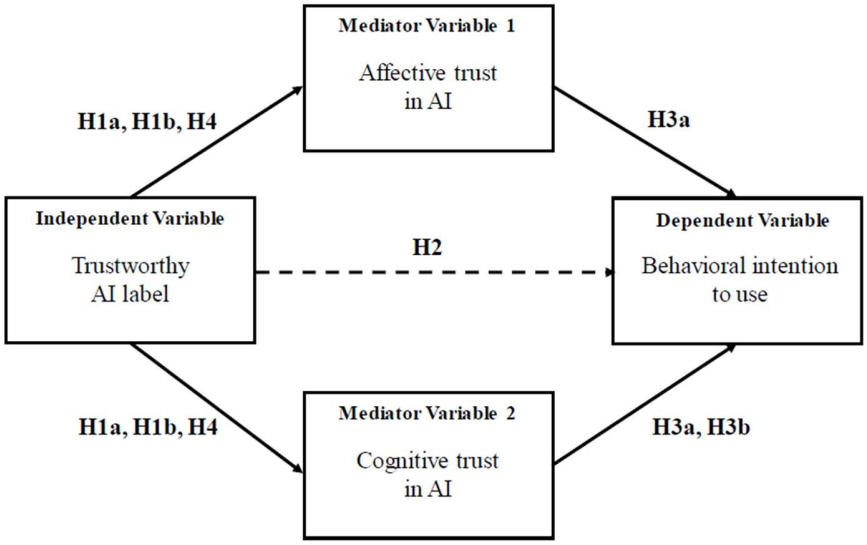

Figure 3: Conceptual Model

While the requirements for TAI hold significant value, Li et al. (2023) claim that demonstrating them equally to end-users when AI systems are adopted is not possible. According to the authors, this applies in particular to the requirements of transparency or privacy preservation. Since in the context of AI, transparency involves disclosing information regarding its entire lifecycle, this might be especially important for end-users as AI systems are characterized as black boxes (Li et al., 2023). Furthermore, a previous study by Toreini et al. (2020) suggests that it is especially important to cover end-users' data-related concerns in AI as privacy violations and unauthorized use of personal data are major reasons for distrust in AI. Therefore, these requirements are expected to be especially important for end-users regarding their interaction and the trustworthiness of an AI system, which leads to the final hypothesis:

H5: *Transparency as well as Privacy and Data Governance are the most important criteria for end-users in their interaction with AI systems.*

To test the outlined hypotheses, two experiments were conducted. The methodological approach and the results for each of them will be described in the following two chapters.

3. Study 1

3.1 Research Design

To test whether and how a label for TAI can increase end-users' trust in AI systems and thus, the acceptance of AI systems, an experimental approach was chosen for Study 1. According to Malhotra et al. (2017), experimentation is an appropriate research approach to test for causality in hypothetical situations. In such an experiment the independent variable (IV), is manipulated by the researcher and the effect on a dependent variable (DV) is measured. The random allocation to the different levels of the IV allows to neutralize the effect of potential extraneous variables that could impact the outcome of the DV and thereby ensures internal validity (Bell et al., 2022).

In line with previous research that studied the presence of AI labels on end-users' trust (Scharowski et al., 2023), a quantitative study in the form of an online experiment, designed with Qualtrics, was used to answer the research questions and to test the stated hypotheses. Among others, advantages of an online survey are low administration costs, as well as ease of data collection (Evans & Mathur, 2005). The random assignment of participants to different scenarios is crucial for ensuring a true experimental design and prevent it from turning into a quasi-experimental approach (Bhattacherjee, 2012). Inspired by previous research, a scenario-based approach was established to examine the effect of certification labels on end-users' trust in AI (Binns et al., 2018; Jakesch et al., 2022; Kapania et al., 2022). To do so, people were presented with a real-word example of an AI system in a hypothetical scenario, adapted from Kapania et al. (2022). In this case, a hiring procedure with an AI system was chosen as labels were found to be specifically effective in such high-risk scenarios (Scharowski et al., 2023). Using hypothetical scenarios instead of actual consumer applications offers the advantage of allowing control over variations in participants' prior experience with the applications. Moreover, participants' actions in scenario-based experiments have been found to mirror their real-life behaviour (Kapania et al., 2022; Woods et al., 2006). The experimental and the control scenario was the same in all aspects, except that in the former, participants were presented an AI with a certification label for trustworthiness and in the control scenario, the AI was not accompanied by any label. A between-subject design enabled the comparison of participants' responses from different treatment conditions because the manipulated IV led to the random assignment of half of the participants to the treatment group (label presence) but not to the other half (the control group, where the label was absent). Along with the benefit of

comparability, such between-subjects designs also prevent knowledge and spillover effects (Charness et al., 2012).

3.2 Procedure

The survey starts with the informed consent form, which includes information about the general procedure, the approximate duration, and a brief overview of the research topic "TAI", without providing concrete information about the aim of the study. After agreeing to take part in the study, participants were then asked to rate their understanding of AI systems. Following, participants were randomly and evenly exposed to one of the two scenarios and were advised to imagine themselves in the scenario presented, while answering the subsequent quantitative questions as realistically as possible. In the scenarios, participants of all groups were asked to take on the role of applicants for a new job at a company which is using an AI system for evaluating job applications. Therefore, they read information regarding the applicant the AI system assessed to determine whether or not they will be invited for an interview. At the end of the description of both scenarios, participants were given a short definition of AI to make sure participants are aligned and thereby reduce variability in the trust variable not related to the manipulation. Participants in the label condition were further given a brief explanation and a graphic representation of the certification logo. Following the scenario briefing, participants were asked to indicate their agreement with 11 different statements referring to the level of trust they had in the presented AI system. In between these questions, a question aimed at testing participants' attention was included. Furthermore, the participants' acceptance of the presented AI system was assessed by asking them to rate their agreement with statements referring to their behavioral intention to use the AI. Before debriefing and thanking participants for their participation in this study, they were asked general demographic questions concerning their age, gender, nationality, education, and employment status, as well as their English language skills. In addition, participants were queried on how much attention they paid when answering the questions and through which channel they accessed the survey. See Appendix 1 for the full survey questionnaire.

3.3 Participants and Data Cleaning

The required sample size was determined beforehand by running a power analysis in G Power (at 0.8 power) for the mixed ANOVA and a Monte Carlo simulation for the parallel mediation. The effects were assumed to be medium sized, which resulted in a minimum required sample size of

128. To enhance statistical power, the required sample size was further increased to 150. Data was collected using a nonprobability sampling technique; convenience sampling took place with voluntary participants from my personal and professional network that were recruited through Social Media (Instagram, Facebook, LinkedIn), WhatsApp and via mail. Participants could choose on a voluntary basis whether they wanted to take part in this study and no extrinsic incentives or rewards were set for successful participation.

A total of 178 responses was collected. The process of data cleaning and subsequent analyses was carried out using the statistical software IBM SPSS. All participants that indicated "1 = *Strongly disagree*" or "2 = *Disagree*" regarding the control question "*I have never used a smartphone"* were eliminated by a listwise deletion. This approach was chosen to ensure that the used data referred to attentive participants. Although this approach decreases statistical power, it increases data quality (Tsikriktsis, 2005).

After the elimination, the total valid sample size included 129 participants, of which 58.1% were females and 39.5% males. Their age ranged from 21 to 60 years (M = 27.64, SD = 5.86) and the majority of participants had a German nationality (N = 87; 64.4%). Furthermore, most respondents were either employed (N = 59; 45.7 %) or students (N = 41; 31.8%) at the time of the survey and had a Bachelor's (N = 64; 49.6%) or Master's degree (N = 51; 39.5%). On average, participants rated themselves as having a moderately good understanding of AI (M = 5.30, SD = 1.06). For more details on the sample descriptive statistics, see Appendix 3.

3.4 Variable measurement

3.4.1 Independent variable

Certification label for TAI: The IV is the certification label for TAI which has two conditions: presence of the label vs. absence of the label. In each of the scenarios, participants where either presented the situation in which the AI system was certified with the certification label for trustworthiness (treatment group) or the AI scenario without any label (control group). The AI system was described as performing the same tasks across conditions. In the treatment group, this information was preceded by information regarding the AI label. The information provided regarding the certification label was based on relevant findings from prior research on a similar topic (Scharowski et al., 2023; Stuurman & Lachaud, 2022).

The certification label used for this research purposes (see Figure 3) was designed based on key success factors for the effective implementation of labels in AI regulation, particularly those related to end-users (Stuurman & Lachaud, 2022). In alignment with the suggestions of researchers and academic literature on labeling, an explanation of the meaning of the label was added to ensure its clarity and to prevent the potential for misinterpretation (Schebesta, 2019). This is essential, as the extensive use of labels on a wide range of products has led to concerns related to consumer confusion and uncertainty regarding the labels' significance and meaning (Stuurman & Lachaud, 2022; Velčovská & Del Chiappa, 2015).

Figure 4: Certification label for TAI (Certification label 1)

3.4.2 Dependent variable

Behavioral Intention to Use: The DV was measured by using three statements referring to the "behavioral intention to use" and will thereby serve to assess the acceptance of AI systems in this thesis. The statements were taken from relevant previous research that studied similar effects (Johnson & Grayson, 2005; Castelo et al., 2019), with only minor changes in wording to relate it to the described AI in the scenario. Participants indicated their agreement with the three statements using a seven-point Likert scale (1 = *Strongly disagree* and 7 = *Strongly agree*).

3.4.3 Mediator

Trust in AI: Trust in AI was measured based on the three pillars of the construct used for trust in people (Mayer et al., 1995) and trust in technology (Mcknight et al., 2011): ability, benevolence, and integrity. Since these three pillars of the ABI model were mapped into the cognitive (ability) and affective (benevolence, integrity) trust dimension, trust was measured using a two-factor structure, consisting of cognitive trust in AI (five items) and affective trust in AI (six items). All statements about the AI systems were to be rated on a seven-point Likert-scale, ranging from 1 (*Strongly disagree*) to 7 (*Strongly agree*). The statements were drawn from prior research exploring

comparable effects (Johnson & Grayson, 2005; Castelo et al., 2019), with slight modifications in language to align with the specific AI scenario under consideration.

3.4.4 Covariate

Familiarity with AI: As familiarity has shown to be a predictor for trust (Gefen, 2000), to control for its effects in the context of AI, this variable was included as a covariate in the model. Therefore, participants were asked to rate their understanding of AI on a seven-point Likert scale (1 = *Extremely bad*; 7 = *Extremely good*).

3.5 Scale reliability

Although the scales used in this experiment have proved reliable in past studies (Mazey, 2018; McKnight et al., 2011; Xu et al., 2014), a reliability analysis was conducted using Cronbach's alpha (α). Reliability is the measure of internal consistency of the constructs used in the study. A construct is considered reliable if the α value is greater than .70 (Hair et al., 2013). Before running the reliability analysis, the data was checked for reverse-coded items and any missing data points. In addition, the statements referring to each trust dimension were grouped together by taking the mean and were labelled accordingly. The results of the reliability analysis revealed that the affective trust scale (α = .86) and the cognitive trust scale (α = .94) both showed high internal consistency between the items and are reliable enough to compose the respective variables. Similarly, the behavioral intention to use scale was also found dependable (α = .89). For more details see Appendix 4.

3.6 Hypotheses testing

To test the hypothesis whether a certification label for TAI had an impact on affective and cognitive trust in and the acceptance of the AI system (**H1a, H1b, H2**), a mixed analyses of variance (ANOVA) was performed, using the statistics software IBM SPSS.[1]

The certification label increased user trust in AI, such that the certification label for trustworthy AI led to a greater effect on both affective ($M = 4.71$, $SD = 1.00$, $t(127) = 4.81$, $p < .001$) and cognitive

[1] The mixed ANOVA was also run on the full sample, without elimination of participants, and showed similar effects, which supports the reliability of the results. See Appendix 5 for more details.

trust (M = 5.24, SD = 1.16, $t(127) = 4.90$, $p < .001$) in the AI system compared to the AI system in the control scenario without the label (Affective: M = 3.80, SD = 1.15; Cognitive: M = 4.18, SD = 1.24). Thus, the findings of the mixed ANOVA suggest that an AI that received a certification label for trustworthiness leads to higher levels of end-users' trust, compared to the same AI system without any certification label. **H1a** is thereby supported. Furthermore, the results of the mixed ANOVA showed that the effect of certification labels on the behavioral intention to use the AI (M = 4.60, SD = 1.36) was also higher compared to the effect of the AI without a certification label on behavioral intention to use (M = 3.38, SD = 1.34, $t(127) = 5.08$, $p < .001$). This finding strenghtens **H2**. Moreover, the mixed ANOVA revealed that the effect of the certification label on cognitive trust (M = 5.24) was higher than the effect on affective trust ((M = 4.71), $F(1, 126) = 31.42$, $p < .001$), thus supporting the hypothesis **H1b**.

To test for the parallel mediation of affective and cognitive trust on behavioral intention to use, a statistical analysis was conducted using Hayes' PROCESS macro for SPSS. The macro is based on regression-path analyses to reveal moderation and mediation effects using a bootstrapping approach (Hayes, 2018). As hypotheses **H3a** and **H3b** involved testing for the indirect effect of the label on behavioral intention to use via the trust dimensions, model number 4 was run. For the following hypotheses tests, a 5% significance level with 5,000 bootstrap replications was chosen. The chosen model involves testing for the indirect effect of X on Y via the proposed mediators M1 (affective trust) and M2 (cognitive trust), making it a parallel mediation. Furthermore, the participant's understanding of AI was included as covariate.[2] To test whether the effects are significant, two regression sub-models were conducted.

The first sub-model entailed regressing the M1 onto X and showed a positive and significant effect of the label condition (absence vs. presence) on the affective trust dimension, $t(2, 124) = 4.69$, b = .91, $p < .001$. Thus, the relationship between the IV and M1, affective trust, is positive and significant. The overall sub-model is significant and explains 15.23% of the variance, R^2 = .1523, $F(2, 124) = 11.14$, $p < .001$. Moreover, the simple regression of M2 onto X shows a positive and significant relationship between the label condition and the cognitive trust dimension, $t(2, 124)$ =

[2] Results show that the covariate does not have a significant effect within the model. The PROCESS was also run on the full sample without the covariate. The outputs are presented in Appendix 6. However, the results are unchanged.

5.04, $b = 1.07$, $p < .001$. The overall sub-model is significant and explains 18.98% of the variance, $R^2 = .1898$, $F(2,124) = 14.51$, $p < .001$.

The second sub-model represents the regression of Y onto X, M1, M2, as well as the indirect effect of X on Y via M1 and M2, which captured the mediating effect of cognitive and affective trust. To understand the relationship between both mediators and the DV (behavioral intention to use), the PROCESS's output data for the b-paths is analyzed. The regression sub-model analyzing the effect of affective trust on behavioral intention to use shows a positive and significant effect, $t(4, 122) = 4.12$, $b = 0.43$, $p < .001$. The same effect can also be found for the relationship between cognitive trust and behavioral intention to use, $t(4, 122) = 5.38$, $b = .51$, $p < .001$. The overall sub-model is significant and explains 60.58% of the variance, $F(4, 122) = 46.86$, $p < .001$. Further, the second sub-model entailed an indirect effect of the label scenario on behavioral intention to use AI through both cognitive and affective trust. Since 0 falls outside the lower and upper interval bounds (Affective CI [.16, .69]; Cognitive CI [.28, .88]), these variables mediate the main effect, which supports **H3a**. As the contrast indicates that the difference between the mediation effect is significant (C1 CI [.11, .33]), and according to the b-paths stronger for cognitive (b1 = .51) than affective trust (b2 = .43), **H3b** can also be supported. Figure 4 summarizes the results of the PROCESS model.

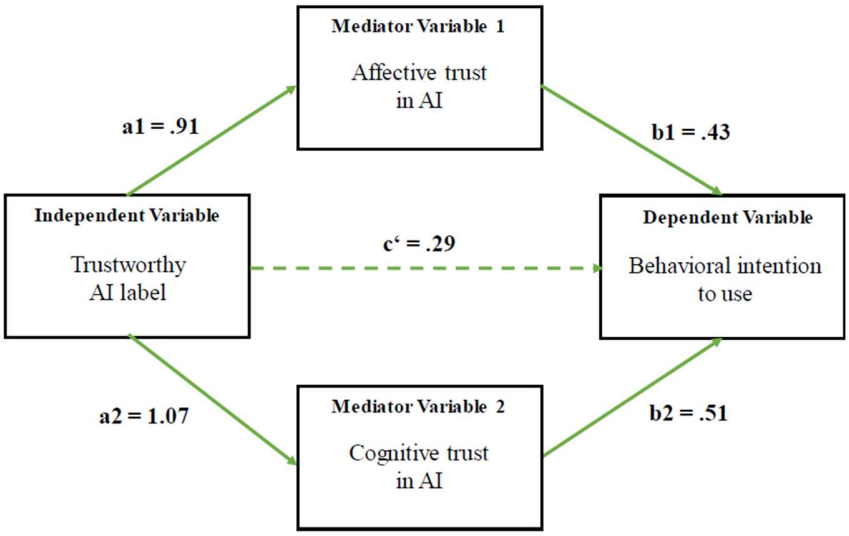

Figure 5: Parallel Mediation Model

4. Study 2

4.1 Research Design

The aim of Study 2 was twofold: to test the robustness of the findings from Study 1, and to uncover which influence the seven criteria for TAI have on the previously examined effect. To test whether displaying this information on the certification label influences the perceived trustworthiness of an AI system, an experimental approach was found to be a useful method. Therefore, a second experiment was conducted, in which participants were randomly and evenly assigned to one of the three groups (no label vs. certification label vs. certification label with requirements). To ensure comparability and replicate the effect of the first study, the same scenario of the hiring procedure with the AI system was chosen for Study 2. Furthermore, the participant's perceptions of the importance of the seven requirements for TAI was investigated to test whether displaying them on the label influences end-users' trust in the AI system.

4.2 Procedure

The procedure followed a similar structure as Study 1, with the following exceptions: To examine whether the effect of the first experiment also holds true when the seven requirements are displayed in the label, participants were now randomly and evenly exposed to one of three different scenarios. Moreover, participants had to rate the seven ethical requirements proposed by AI HLEG (2019) regarding their importance in the design of AI systems that interact with them. Each of these requirements was briefly explained to ensure alignment. This question order was chosen to prevent an influential effect of the requirement rating on the trust-related questions since the importance of the requirements was also assessed for future research purposes. See Appendix 2 for the full experiment.

4.3 Participants and Data Cleaning

The required sample size was pre-determined by running a power analysis in G Power (at 0.8 power) for the mixed ANOVA and the repeated measures ANOVA. For the parallel mediation, a Monte Carlo simulation was conducted, which resulted in the larger sample size of 128. Considering the structure of the study, which includes three experimental groups, the required sample size was further increased to 200 to increase statistical power.

The required sample size was determined beforehand by running a power analysis in G Power (at 0.8 power) for the mixed ANOVA and the repeated measures ANOVA. For the parallel mediation, a Monte Carlo simulation was conducted. Considering the structure of the study, which includes three experimental groups, the required sample size was further increased to 200.

In total, 234 survey responses were completed, from which some were eliminated following the data cleaning approach used in Study 1. After the elimination the total valid sample size included 221 participants, with 65.2% being female and 33.5% male. Their age ranged from 19 to 60 years (M = 28.02, SD = 6.71) and most participants had a German nationality (N = 139; 62.9%). Furthermore, the majority of respondents were either employed (N = 108; 48.9%) or a student (N = 78; 35.3%) at the time of the survey. Regarding their education, most participants indicated to either have a Bachelor's (N = 107; 48.4%) or Master's degree (N = 89; 40.3%). On average, participants rated themselves as having a moderately good understanding of AI (M = 5.48, SD = 0.91). For more details on the population statistics, see Appendix 7.

4.4 Variable measurement

4.4.1 Independent variable

Certification label for TAI: The IV is the TAI certification label which has three levels. In each of the scenarios, participants where either presented the situation in which the AI system was certified with the label for trustworthiness (treatment groups 1 & 2) or the scenario where the AI system did not receive any label (control group) – with the task the AI system had to perform in all conditions being the same. While the label in the first treatment group received the same certification label as in Study 1 (certification label 1), the label in the second treatment group (certification label 2[3]) further displayed the seven requirements for TAI to indicate that they served as criteria for the certification of the AI system (see Figure 5). In the control group, the AI system did not receive any label for trustworthiness.

[3] For simplicity, the label without the TAI criteria is now referred to as "certification label 1" and the label displaying the requirements for TAI as "certification label 2".

Figure 6: Certification label with requirements for TAI (Certification label 2)

4.4.2 Dependent variable

Behavioral intention to use: As in Study 1, the DV, behavioral intention to use AI, was measured with three items using a seven-point Likert scale (1 = *Strongly disagree* and 7 = *Strongly agree*).

4.4.3 Mediator

Trust in AI: The mediator is the participants' levels of trust in the AI system by measuring the cognitive and affective dimension of trust. Using a seven-point Likert scale (1 = *Strongly disagree* and 7 = *Strongly agree*), participants had to rate their agreement to eleven statements referring to these dimensions. This format was chosen for consistency with Study 1.

4.4.4 Covariate

Familiarity of AI: Similar to Study 1, the participants' familiarity with AI was included as covariate in the model. For consistency, it was measured on a seven-point Likert scale (1 = *Extremely bad*; 7 = *Extremely good*).

4.4.5 Importance Analysis

Requirements for TAI: The importance of each of the seven requirements for TAI, proposed by the AI HLEG (2019), was judged on a Likert scale, with the anchors being 1 = *Not important at all* and 7 = *Extremely important*. The description of each requirement was taken from a previous study (Choung et al., 2023). In this analysis, the importance rating is the DV, and the requirement is the IV.

4.5 Scale reliability

As in Study 1, a reliability analysis was conducted using Cronbach's α. Both the affective trust scale (α = .94) and the cognitive trust scale (α = .97) were reliable. Moreover, the behavioral

intention to use scale was also found to be reliable (α = .92). For more details on the Cronbach's α, see Appendix 8.

4.6 Hypotheses testing

To test **H4** a mixed ANOVA was performed in SPSS, comparing the effect of the two different label conditions on end-users' trust. A planned contrast indicated that the effect of label 2 on end-users' trust was above and beyond the effect of certification label 1. Specifically, the results of the mixed ANOVA demonstrate that certification label 2 had a greater influence on both cognitive (M = 6.29, SD = .82, $t(144) = 4.09$, $p < .001$) and affective trust (M = 5.28, SD = .75, $t(137) = 3.99$, $p < .001$) in the AI system, compared to certification label 1 (Cognitive: M = 5.72, SD = .86; Affective: M = 4.85, SD = .55). Thus, the findings suggest that displaying trustworthiness requirements on the label translates into higher trust in the AI system. Therefore, **H4** is affirmed. The effect of the different AI scenarios on end-users' trust is illustrated in Figure 6.[4]

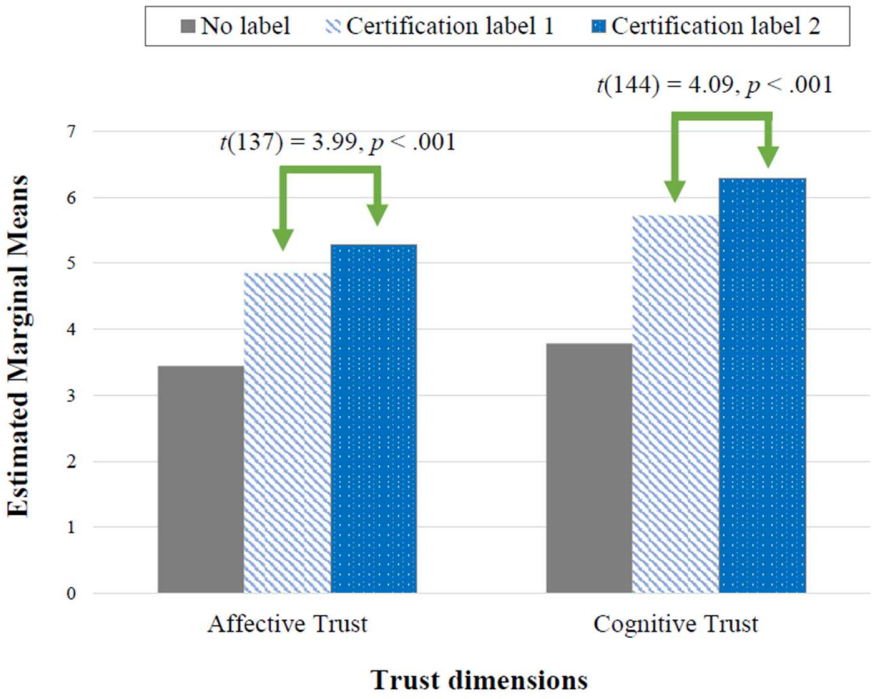

Figure 7: Effect of different certification labels for TAI on end-users' trust

[4] The mixed ANOVA was also performed on the complete sample, without the elimination of participants. However, the results do not change and further suggest the robustness of the findings of Study 1 (See Appendix 9).

To test for the indirect effect of the label scenarios on behavioral intention to use the AI via the two trust dimensions, a Hayes' PROCESS macro model 4 was applied. As in Study 1, a 5% significance level with 5,000 bootstrap replications was chosen. Since the IV is multicategorial, consisting of three levels, indicator coding was applied with the control group as reference group (Hayes & Preacher, 2014). Therefore, X1 refers to the comparison between the certification label 1 and the control group, and X2 compares the certification label 2 and the control group. Furthermore, the model controls for "understanding of AI" by including it as covariate. As the model aimed to test whether a label for TAI increases end-users' behavioral intention to use the AI systems though increased cognitive and affective trust, the focus is on the output for the indirect effects of M1 and M2. Since 0 falls outside the lower and upper interval bounds for both affective (X1: 95% CI [.11, .73]; X2: 95% CI [.15, .94]) and cognitive trust (X1: 95% CI [.24, .82]; X2: 95% CI [.28, .98]), trust has a significant indirect effect on behavioral intention to use. Thus, the test revealed that a label for TAI increases end-users' acceptance (behavioral intention to use) of AI systems though increased cognitive and affective trust in AI, which strengthens **H3a**. As the mediation effect was found to be stronger for cognitive trust (C1 CI [.23, .55]), b1 = .57), **H3b** is also supported. Since the results of the PROCESS macro are consistent with the findings of the first study, the confidence in those findings is increased.[5] In contrast to Study 1, a positive and significant effect of the covariate on both trust dimensions was found in Study 2. Thus, a greater understanding of AI is associated with higher levels of affective trust, $t(3, 217) = 3.11$, $b = .17$, $p = .002$, as well as cognitive trust, $t(3, 217) = 4.24$, $b = .27$, $p < .001$.

To test **H5**, which suggested that both *Privacy and Data Governance* as well as *Transparency* are the most important criteria for end-users in their interaction with AI, a repeated measures ANOVA was performed on the data of the participants in the control group[6]. This approach was chosen because in the control group an influence of the AI labels on the importance perception of the requirements could be prevented. From the results of the repeated measures ANOVA, it becomes evident that *Transparency* is of greatest importance for end-users regarding their interaction with AI systems ($M = 6.32$, $SD = .67$). Moreover, the importance of this requirement was significantly

[5] The PROCESS model was also performed on the complete sample, without the elimination of participants and the covariate. The results do not change (See Appendix 10).

[6] The repeated measures ANOVA was also performed on the data of all treatment groups; however, the results did not change.

different to those of the other requirements for TAI (*MD* to second highest requirement = .78, $p <$ .001). *Privacy and Data Governance* was rated as the second most important requirement for TAI ($M = 5.55$, $SD = .71$). However, compared to *Technical Robustness and Safety*, which was ranked as third most important ($M = 5.25$, $SD = .75$), *Privacy and Data Governance* was not significantly more important ($MD = .30$, $p = .114$). Thus, **H5** can only partially be supported. The remaining criteria were ranked in descending order of importance as follows: *Diversity, non-discrimination and fairness*; *Accountability*; *Human agency and oversight;* and *Societal and environmental well-being*. However, it is important to mention that the means of all requirements where above 4.49, which indicates that overall, all requirements were ranked as rather important.

5. Discussion

5.1 Research Findings

The theoretical part of this thesis unveiled a research gap which led to several hypotheses that were examined in two experiments. First, the experiments tested the effect of a certification label for TAI on end-users' trust in the respective AI system. Both experiments' results support **H1a,** which predicted that a certification label for TAI can increase end-users' trust in the AI. These results are consistent with the findings of previous studies suggesting that compliance with requirements for TAI can be communicated through certification labels, potentially increasing end-users' trust in the AI system (Scharowski et al., 2023; Stuurman & Lachaud, 2022). Since the AI system was the same in all scenarios, with only the certification label for trustworthiness bearing the difference, it can be interpreted that the label serves as external source of validation for assessing system trustworthiness (Jacovi et al., 2021; Lee & See, 2004). This also supports previous research stating that labels can function as "trustworthiness cues" for end-users to build trust in the AI system (Lee & See, 2004; Liao & Sundar, 2022). Furthermore, both experiments revealed how a label for TAI influences different trust dimensions by examining the effect on cognitive and affective trust in AI. From the results it became evident that a certification label for TAI positively impacts both end-users' cognitive and affective trust in the AI system. In particular, the results indicate a greater influence of the certification label on cognitive than affective trust, which supports **H1b**. This outcome may be attributed to the likelihood that certified AI systems commonly undergo an audit procedure, which focuses primarily on assessing AI competency and reliability rather than evoking an emotional bond between the user and the system (Chen & Sundar, 2023). By applying the cognitive and affective trust framework in this context, this study complements existing research (Cihon et al., 2021; Scharowski et al., 2023; Seifert et al., 2019) on the effectiveness of certification labels in building end-users' trust in AI systems.

Moreover, in both experiments the label for TAI was also found to increase end-users' behavioral intention to use the AI, which supports **H2**. Therefore, it can be interpreted that a label for TAI can effectively communicate system trustworthiness and thereby, positively contributes to their acceptance. In fact, the data suggests that the certification label for TAI promotes the behavioral intention to use the AI through increased cognitive and affective trust in AI. Since, both cognitive and affective trust were found to have a significant mediation effect on the users' behavioral

intention to use the AI, **H3a** was supported. This emphasizes the importance of trust for a successful adoption and deployment of AI. In fact, the mediation effect was found to be stronger for cognitive trust, which supports **H3b** and further reinforces the idea that cognitive trust plays a greater role in the relationship between the certification label and AI acceptance. Nevertheless, as both trust dimensions were found to have a significant indirect effect, it is important to assure that the perceptions of both are high.

Additionally, the importance of familiarity with AI for the studied relationship differed across studies, being important in Study 2 but not in Study 1. This is in line with literature, as some studies have found it to play a role (Gefen, 2000), but others have shown that this is not always the case (Gillath et al., 2021). This further suggests that the familiarity with AI does not have a reliable relationship with AI outcomes.

Furthermore, the experiments compared the effect of different types of certification labels on end-users' trust in the AI system. The results revealed that the label with the requirements for TAI led to higher levels of affective and cognitive trust, compared to the scenario where the AI system was certified with the simple certification label that did not provide any information about the requirements that were used to certify the AI system. These findings support **H4**. The results further allow to derive aspects to consider for effective certification labels for AI applications. Although researchers suggest that displaying a lot of information on a label can be overwhelming for consumers and thereby reducing its effectiveness (Tonkin et al., 2015), this could not be supported in the case of this experiment. By contrast, participants displayed higher levels of trust in the AI system that received the certification level showing all seven trustworthiness requirements than in the AI that was certified with the simple label for trustworthiness. Examining how end-users perceive the importance of different requirements for TAI revealed that overall, all seven requirements were perceived as rather important by end-users regarding their interaction with AI, underlining the importance of an AI system meeting the TAI requirements for trust building. More specifically, communicating the criteria used in the underlying AI auditing process enhances users' perception of the AI system's trustworthiness, as they can see which certification criteria have been met by the AI (Stuurman & Lachaud, 2022). Each criterion reassures end-users about the AI's competence, ethical behavior, and overall reliability, allowing them to form an overall sense of trust in the AI (Jacovi et al., 2021; Liao & Sundar, 2004). However, the findings also revealed that *Transparency* was found to be significantly more important than the other requirements. This result

can be explained by the fact that transparency is particularly important for users to understand the reasoning behind an AI system's decision. Since many AI-related concerns arise from its black-box character, transparently disclosing information regarding the system's entire lifecycle can ease end-users' uncertainties. Although *Privacy and Data Governance* was ranked as second most important, it was not significantly more important than all the other requirements. Thus, **H5** can only partially be supported. This result might also be attributed to the hiring scenario chosen for the experiment as users might be more concerned about the decision-making of the AI in this case, rather than their data. Since the results might differ with a different scenario, I will conduct future research that compares the effect in different AI scenarios.

5.2 Academic and Managerial Implications

5.2.1 Academic Implications

The growing body of literature on TAI in recent years illustrates the emerging relevance of the topic. However, there has been limited research focusing on how to implement TAI into practice and communicating an AI's trustworthiness to end-users. Building on the previous literature and the research gaps it reveals, certification communication was examined more closely as non-technical method to enhance end-users' trust in AI, offering several theoretical implications.

First, the research findings contribute significantly to the academic discourse surrounding trust in AI, emphasizing its essential role in system acceptance. By establishing a crucial connection between the literature on TAI and the broader discussions on certification, this dissertation was able to support the statement that certification labels can have a positive influence of end-users' trust in and thereby, the acceptance of AI. In particular, the findings underpin the results of past studies, suggesting that certification labels have a greater effect on cognitive than affective trust (Chen & Sundar, 2023; Schuitema et al., 2020). Additionally, the dissertation supported the effectiveness of labels in high-stake scenarios. The successful replication of effects identified by previous researchers enhances confidence in the original findings which contributes to the robustness of the literature (Cihon et al., 2021; Scharowski et al., 2023; Seifert et al., 2019).

Second, by investigating different label designs, this work adds to the theoretical knowledge on labels, offering insights applicable for the specific context of AI. The findings suggest that end-users display higher trust levels in AI systems with certification labels featuring the trustworthiness requirements used in the audit. This challenges previous recommendations (Stuurman & Lachaud,

2022) and suggests that, in the context of AI, comprehensive information on labels contributes positively to user trust. This provides a foundation for further research and potential replication by scholars.

Finally, this is the first empirical study that tests the importance of TAI requirements and how they translate into trust when embedded in labels. The findings of the second study support the importance of requirements for TAI in trust building. Moreover, the identification of *Transparency* as significantly important criteria for end-users suggests that certain elements hold more weight in communicating a system's trustworthiness.

5.2.2 Managerial Implications

In addition to the theoretical contributions, this research also offers managerial implications. With the growing influence of AI and its disruption of the workplace, ensuring the trustworthiness of AI systems stands out as a prominent challenge these days. With respect to organizations that provide AI-based products and services, it is highly recommended to ensure the trustworthiness of their AI systems, given the demonstrated importance of trust in user acceptance of AI. The study further emphasizes that embedding TAI principles could become a competitive advantage in the evolving landscape (AI HLEG, 2019).

The shown effectiveness of certification labels for TAI in enhancing trust in AI suggests that organizations should leverage certification labels as a strategic tool for communicating AI system trustworthiness to end-users. Managers should view certification labels not only as symbolic endorsements, but as practical tools that can effectively communicate trustworthiness and, consequently, reduce perceived risks associated with AI systems. Since trust was found as a predictor of AI acceptance, certification labels should be leveraged to not only communicate trustworthiness, but also actively promote users' behavioral intention to use AI systems.

Moreover, the findings recommend the development of certification labels that transparently communicate the trustworthiness requirements met by the AI system. Therefore, it is advisable for institutions, issuing labels, to follow reputable trustworthiness requirements, aligning with the quantitative insights derived from the study's results. Furthermore, since the TAI requirement *Transparency* was found to be especially important in communicating system trustworthiness, organizations are advised to clearly reassure end-users that the AI system is committed to transparency.

While there is no official label for TAI yet, political bodies, like the EU, should strive towards establishing a certification for TAI. Therefore, it remains to be seen which certification programs will be implemented, their level of effectiveness, the complexity of their implementation and how durable they will be in the face of changes in AI (Cihon et al., 2021).

5.3 Limitations and future research

While this research carries important theoretical and practical implications, it is crucial to acknowledge certain limitations particularly with respect to the generalization of the results.

The first constraint lies in the utilization of a non-probability sampling technique for data collection, which given the time and resource constraints was a reasonable technique, however it led to a non-representative sample (Vehovar et al., 2016). Specifically, most of the participants were contacted personally via WhatsApp or they became aware of the anonymous Qualtrics link via Social Media, which led to the result that most of the participants originated from Europe, and particularly from Germany. To enhance the reliability of the outcomes, it is recommended that future research replicates this study with a larger and more diverse sample, addressing the existing limitations and ensuring broader generalizability of the findings.

Second, although certification labels have shown to be effective in communicating an AI system's trustworthiness to end-users, they have a limited capacity to indicate untrustworthiness, as they offer only binary information by either being present or absent. Consequently, it becomes challenging to distinguish whether a product lacking a certification label is considered untrustworthy due to its failure to meet the label's criteria or if it simply hasn't undergone an audit yet (Scharowski et al., 2023). Therefore, future research could explore different types of labels and their effect on end-users' trust. Nutrition labels for example could display a layer of interpretation about how the AI system scored in each requirement for TAI (e.g., Andrews et al., 2011).

Another limitation of this study lies in its exclusive focus on a high-stake AI scenario in the experiments, which was chosen due to the proven effectiveness of labels in those specific scenarios (Scharowski et al., 2023). However, the trustworthiness for a high-risk AI system, such as the hiring procedure, may differ from the one of a low-risk system (e.g. music recommendation; Ferrario, 2023), thereby restricting the possibility to make comprehensive assertions about the broader efficacy of certification labels in enhancing end-users' trust across diverse AI applications. Similarly, many certification frameworks lack modification for each of the various sectors where

AI finds application (Cihon et al., 2021). Thus, future research could compare the effect of certification labels on end-users' trust dimensions with AI systems that are used in both high- and low-risk contexts as well as different sectors. Additionally, the trust in the AI system was only assessed from the applicant's point of view in this thesis, rather than the company's view that implements the AI system. Given the purpose of testing end-users' trust in AI in such high-risk scenarios, this approach was reasonable, however, future studies should also investigate the company's trust levels.

The two experiments in this thesis examined the mediating effect of both affective and cognitive trust on the relationship between the AI label and the acceptance of the AI. While this effect has shown to be significant, future studies could expand the model by adding further moderators suggested by literature, such as the perceived credibility of the institution that issues the label (Schuitema et al., 2020). To do so, they could investigate how different issuer of the label impact the trust levels generated by the AI with the label.

Another limitation regarding this dissertation's model refers to the scale that was used to assess the acceptance of the AI. Although the behavioral intention to use is applied in several acceptance models (e.g. TAM (Davis, 1985); AIDUA (Gursoy, 2019); UTAUT (Venkatesh et al., 2003)), it only represents one dimension of it and could further be expanded by other dimensions referring to the acceptance of AI, such as the attitude toward using and the actual system usage.

Another limitation of this study pertains to the dynamic nature and continuous evolution of AI, making certification programs tailored for present-day AI systems potentially inadequate for the AI of tomorrow (Cihon et al., 2021). Therefore, continuous research is needed to evaluate and update relevant trustworthiness requirements. While this study focused on the seven requirements for TAI, proposed by the European Commission's AI HLEG, further studies could consider the effect of different trustworthy frameworks.

The second experiment revealed that the certification label with the information regarding the requirements for TAI led to higher trust and acceptance of the AI. Further research should investigate whether this is due to the quantity of requirements or other factors, e.g. the order. Likewise, researchers suggest highlighting key label criteria to prevent user overwhelm and enhance understanding of the label's meaning (Stuurman & Lachaud, 2022). While *Transparency* was identified as the most important requirement, this theoretical insight requires a more detailed

examination, especially regarding the importance of the requirements in terms of trust-building. Therefore, I am conducting further research to explore if displaying this requirement on labels leads to higher trust than others. By doing so, the amount of information presented on certification labels will be tested and redefined to find the right balance that effectively communicates trustworthiness.

6. Conclusion

As AI is being increasingly integrated in our lives, concerns about its trustworthiness have arisen. Ensuring that AI systems are trustworthy and especially communicating system trustworthiness to end-users has become a challenge. This dissertation has contributed to the important research field of TAI by introducing certification labels as an effective non-technical method for communicating an AI system's trustworthiness to end-users. The results found in the two studies presented demonstrate that certification labels have the potential to effectively communicate the outcome of AI audits to end-users, enhancing both trust in and the acceptance of AI. Furthermore, the findings emphasize that an AI's adherence to trustworthy requirements plays a crucial role in end-users' perceptions regarding system trustworthiness.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6,52138 – 52160. https://doi.org/10.1109/ACCESS.2018.2870052.

Aiken, K. D., & Boush, D. M. (2006). Trustmarks, objective-source ratings, and implied investments in advertising: investigating online trust and the context-specific nature of internet signals. *Journal of the academy of marketing science*, *34*(3), 308-323. https://doi.org/10.1177/0092070304271004.

Ajzen, I. (1985). From intentions to actions: A theory of planned behavior. In *Action control: From cognition to behavior* (pp. 11-39). Berlin, Heidelberg: Springer Berlin Heidelberg.

Ajzen, I. & Fishbein, M. (1972). Attitudes and normative beliefs as factors influencing behavioral intentions. *Journal of Personality and Social Psychology*, 21, 1-9. https://doi.org/10.1037/h0031930.

Alalwan, A. A., Baabdullah, A. M., Rana, N. P., Tamilmani, K., & Dwivedi, Y. K. (2018). Examining adoption of mobile internet in Saudi Arabia: Extending TAM with perceived enjoyment, innovativeness and trust. *Technology in Society*, *55*, 100-110. https://doi.org/10.1016/j.techsoc.2018.06.007.

Andrews, J. C., Burton, S., & Kees, J. (2011). Is simpler always better? Consumer evaluations of front-of-package nutrition symbols. *Journal of Public Policy & Marketing*, 30(2), 175–190. https://doi.org/10.1509/jppm.30.2.175.

Arnold, M., Bellamy, R. K., Hind, M., Houde, S., Mehta, S., Mojsilović, A., ... & Varshney, K. R. (2019). Fact Sheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development*, *63*(4/5), 6-1. https://doi.org/10.1147/JRD.2019.2942288.

Avin, S., Belfield, H., Brundage, M., Krueger, G., Wang, J., Weller, A., ... & Zilberman, N. (2021). Filling gaps in trustworthy development of AI. Science, 374 (6573), 1327-1329. https://doi.org/10.1126/science.abi7176.

Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, *7*(1), 52-62. https://doi.org/10.61969/jai.1337500.

Binninger, A. S., & Robert, I. (2005). Les produits labellisés et le développement durable dans la perspective du consommateur: une étude exploratoire. In *21st French Marketing Association (AFM) Conference Proceedings*.

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. *Proceedings of the 2018 Chi conference on human factors in computing systems* (pp. 1-14). https://doi.org/10.1145/3173574.3173951.

Blair, M. M. (2008). The New Role for Assurance Services in Global Commerce. *Journal of Corporation Law 33*: 325-360.

Bostrom, N. (1998). How long before superintelligence?. Oxford Future of Humanity Institute. University of Oxford. *International Journal of Futures Studies*, 2(1), 1–9.

Bandy, J. (2021). Problematic machine behavior: A systematic literature review of algorithm audits. *Proceedings of the acm on human-computer interaction*, *5*(CSCW1), 1-34. https://doi.org/10.1145/3449148.

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., ... & Anderljung, M. (2020). Toward trustworthy AI development: mechanisms for supporting verifiable claims.

Bell, E., Bryman, A., & Harley, B. (2022). *Business research methods*. Oxford university press. https://doi.org/10.1093/hebz/9780198869443.001.0001

Burke, B., Cearley, D., Jones, N., Smith, D., Chandrasekaran, A., Lu, C. K., & Panetta, K. (2019, October 21). *Gartner top 10 strategic technology trends for 2020* - Smarter with Gartner. Retrieved November 16, 2023 from https://www.gartner.com/en/documents/3970506.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809-825. https://doi.org/10.1177/0022243719851788.

Choi, Y. (2023). A Study of Customer Acceptance of Artificial Intelligence Technology. *International Journal of E-Business Research (IJEBR)*, *19*(1), 1-14. https://doi.org/10.4018/IJEBR.323796.

Choung, H., David, P., & Ross, A. (2023). Trust in AI and Its Role in the Acceptance of AI Technologies. *International Journal of Human–Computer Interaction*, *39*(9), 1727-1739. https://doi.org/10.1080/10447318.2022.2050543.

Charness, G., Gneezy, U., & Kuhn, M. (2012). Experimental Methods: Between-Subject and Within-Subject Design. *Journal Of Economic Behavior & Organization,* 81(1), 1-8. https://doi.org/10.1016/j.jebo.2011.08.009

Cihon, P., Kleinaltenkamp, M. J., Schuett, J., & Baum, S. D. (2021). AI certification: Advancing ethical practice by reducing information asymmetries. *IEEE Transactions on Technology and Society*, *2*(4), 200-209. https://doi.org/10.1109/TTS.2021.3077595.

Colquitt, J. A., & Rodell, J. B. (2011). Justice, trust, and trustworthiness: A longitudinal analysis integrating three theoretical perspectives. *Academy of management journal*, *54*(6), 1183-1206. https://doi.org/10.5465/amj.2007.0572.

Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022, June). Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1571-1583). https://doi.org/10.1145/3531146.3533213.

Crockett, K. A., Gerber, L., Latham, A., & Colyer, E. (2021). Building trustworthy AI solutions: A case for practical solutions for small businesses. *IEEE Transactions on Artificial Intelligence*.

Davis, F. D. (1985). *A technology acceptance model for empirically testing new end-user information systems: Theory and results* (Doctoral dissertation, Massachusetts Institute of Technology).

Edelman (2019, January 20). *Edelman Trust Barometer Global Report*. Retrieved November 2, 2023 from https://www.edelman.com/trust/2019-trust-barometer.

Egea, J. M. O., & González, M. V. R. (2011). Explaining physicians' acceptance of EHCR systems: An extension of TAM with trust and risk factors. *Computers in Human Behavior*, *27*(1), 319-332. https://doi.org/10.1016/j.chb.2010.08.010.

European Commission. (2020, February 19). *White Paper on Artificial Intelligence: A European approach to excellence and trust*. Retrieved November 22, 2023 from https://commission.europa.eu/system/files/2020-02/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

European Union (2023): *Artificial intelligence act*. Retrieved November 22, 2023 from https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf.

Evans, J. R., & Mathur, A. (2005). The value of online surveys. *Internet research*, *15*(2), 195-219. https://doi.org/10.1108/10662240510590360.

Falco, G., Shneiderman, B., Badger, J., Carrier, R., Dahbura, A., Danks, D., ... & Yeong, Z. K. (2021). Governing AI safety through independent audits. *Nature Machine Intelligence*, *3*(7), 566-571. https://doi.org/10.1038/s42256-021-00370-7.

Ferrario, A. (2023). Justifying our Credences in the Trustworthiness of AI Systems: A Reliabilistic Approach. *Available at SSRN 4524678*. https://doi.org/10.2139/ssrn.4524678.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2021). An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Ethics, governance, and policies in artificial intelligence*, 19-39. https://doi.org/10. 1007/s11023-018-9482-5.

Gefen, D. (2000). E-commerce: the role of familiarity and trust. *Omega*, *28*(6), 725-737. https://doi.org/10.1016/s0305-0483(00)00021-9.

Gefen, D., & Straub, D. W. (2004). Consumer trust in B2C e-Commerce and the importance of social presence: experiments in e-Products and e-Services. *Omega*, *32*(6), 407-424. https://doi.org/10.1016/j.omega.2004.01.006.

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

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

Goodman, S., Vanderlee, L., Acton, R., Mahamad, S., & Hammond, D. (2018). The impact of front-of-package label design on consumer understanding of nutrient amounts. *Nutrients*, *10*(11), 1624. https://doi.org/10.3390/nu10111624.

Gorton, M., Tocco, B., Yeh, C. H., & Hartmann, M. (2021). What determines consumers' use of eco-labels? Taking a close look at label trust. *Ecological Economics*, *189*, 107173. https://doi.org/10.1016/j.ecolecon.2021.107173.

Grunert, K. G., Hieke, S., & Wills, J. (2014). Sustainability labels on food products: Consumer motivation, understanding and use. *Food policy*, *44*, 177-189. https://doi.org/10.1016/j.foodpol.2013.12.001.

Gursoy, D., Chi, O. H., Lu, L., & Nunkoo, R. (2019). Consumers acceptance of artificially intelligent (AI) device use in service delivery. *International Journal of Information Management*, *49*, 157-169. https://doi.org/10.1016/j.ijinfomgt.2019.03.008.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and machines*, *30*(1), 99-120. https://doi.org/10.1007/ s11023-020-09517-8.

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2013). Partial least squares structural equation modeling: Rigorous applications, better results and higher acceptance. *Long range planning*, *46*(1-2), 1-12. https://doi.org/10.1016/j.lrp.2013.01.001.

Harari, Y. N. (2016). *Homo Deus: A brief history of tomorrow*. random house.

Hayes, A. F. (2018). PROCESS macro for SPSS and SAS. The PROCESS macro for SPSS and SAS. *Introduction to mediation, moderation, and conditional PROCESS analysis, second edition: A regression-based approach*.

Hayes, A. F., & Preacher, K. J. (2014). Statistical mediation analysis with a multicategorical independent variable. *British journal of mathematical and statistical psychology*, *67*(3), 451-470.

High-Level Expert Group on Artificial Intelligence (AI HLEG) (2019, April 8). *Ethics Guidelines for Trustworthy Artificial Intelligence*. https:// ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1.

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

Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2020). The dataset nutrition label. *Data Protection and Privacy*, *12*(12), 1. https://doi.org/10.5040/9781509932771.ch-001.

IEEE (2008, August). IEEE Standard for Software Reviews and Audits. *IEEE Std 1028-2008* (Aug 2008), pp. 1–53. https://doi.org/10.1109/IEEESTD.2008.4601584.

ISO, I. (2004). IEC 17000: 2004 Conformity assessment—Vocabulary and general principles. *International Organization for Standardization*.

ISO 24028:2020. (2020). *Information Technology–Artificial Intelligence–Overview of Trustworthiness in Artificial Intelligence.* Standard. International Organization for Standardization.

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021, March). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 624-635). https://doi.org/10.1145/3442188.3445923.

Jahn, G., Schramm, M., & Spiller, A. (2005). The reliability of certification: Quality labels as a consumer policy tool. *Journal of Consumer Policy*, *28*, 53-73. https://doi.org/10.1007/s10603-004-7298-6.

Jakesch, M., Buçinca, Z., Amershi, S., & Olteanu, A. (2022, June). How different groups prioritize ethical values for responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 310-323). https://doi.org/10.1145/3531146.3533097.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature machine intelligence*, *1*(9), 389-399. https://doi.org/10.1038/s42256-019-0088-2.

Johnson, D., & Grayson, K. (2005). Cognitive and affective trust in service relationships. *Journal of Business research*, *58*(4), 500-507. https://doi.org/10.1016/S0148-2963(03)00140-1.

Jones, A., Neal, B., Reeve, B., Mhurchu, C. N., & Thow, A. M. (2019). Front-of-pack nutrition labelling to promote healthier diets: current practice and opportunities to strengthen regulation worldwide. *BMJ global health*, *4*(6), e001882. https://doi.org/10.1136/bmjgh-2019-001882.

Kapania, S., Siy, O., Clapper, G., SP, A. M., & Sambasivan, N. (2022, April). " Because AI is 100% right and safe": User attitudes and sources of AI authority in India. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1-18). https://doi.org/10.1145/3491102.3517533.

Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2020). Requirements for trustworthy artificial intelligence—A review. In *Proceedings of the International Conference on Network-Based Information Systems*. 105–115. https://doi.org/10.1007/978-3-030-57811-4_11.

Kaur, D., Uslu, S., & Durresi, A. (2021). Requirements for trustworthy artificial intelligence–a review. In *Advances in Networked-Based Information Systems: The 23rd International Conference on Network-Based Information Systems (NBiS-2020) 23* (pp. 105-115). Springer International Publishing. https://doi.org/10.1007/978-3-030-57811-4_11.

Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: a review. *ACM Computing Surveys (CSUR)*, *55*(2), 1-38. https://doi.org/10.1145/3491209.

King, A. A., Lenox, M. J., & Terlaak, A. (2005). The strategic use of decentralized institutions: Exploring certification with the ISO 14001 management standard. *Academy of management journal*, *48*(6), 1091-1106. https://doi.org/10.5465/amj.2005.19573111.

Knowles, B., & Richards, J. T. (2021, March). The sanction of authority: Promoting public trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 262-271). https://doi.org/10.1145/3442188.3445890.

Krafft, P. M., Young, M., Katell, M., Huang, K., & Bugingo, G. (2020). Defining AI in policy versus practice. In Proceedings of the AAAI/ ACM *Conference on AI, Ethics, and Society* (pp. 72–78). https://doi. org/10.1145/3375627.3375835

Kumar, A., Braud, T., Tarkoma, S., & Hui, P. (2020, March). Trustworthy AI in the age of pervasive computing and big data. In *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)* (pp. 1-6). IEEE. https://doi.org/10.1109/PerComWorkshops48775.2020.9156127.

Lakkaraju, H., Adebayo, J., & Singh, S. (2020, December 7). Explaining machine learning predictions: State-of-the-art, challenges, and opportunities. *NeurIPS Tutorial.* Retrieved November 26, 2023, from https://neurips.cc/virtual/2020/public/tutorial_59e711d152de7bec7304a8c2ecaf9f0f.html.

Larceneux, F. (2004). 20ième Congrès de l'AFM. *Impacts des stratégies de labellisation sur le processus de décision des consommateurs: le cas du label biologique*.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50-80. https://doi.org/10.1518/hfes.46.1.50.30392.

Lewis, D., Hogan, L., Filip, D., & Wall, P. J. (2020). Global challenges in the standardization of ethics for trustworthy AI. *Journal of ICT Standardization*, 123-150. https://doi.org/10.13052/jicts2245-800X.823.

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., ... & Zhou, B. (2023). Trustworthy AI: From principles to practices. *ACM Computing Surveys*, *55*(9), 1-46. https://doi.org/10.1145/3555803.

Liao, Q. V., & Sundar, S. S. (2022, June). Designing for responsible trust in AI systems: A communication perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1257-1268). https://doi.org/10.1145/3531146.3533182.

Lukyanenko, R., Maass, W., & Storey, V. C. (2022). Trust in artificial intelligence: From a Foundational Trust Framework to emerging research opportunities. *Electronic Markets*, *32*(4), 1993-2020. https://doi.org/10.1007/s12525-022-00605-4.

Malhotra, N. K., Nunan, D., & Birks, D. F. (2017). *Marketing research: An applied approach*. Pearson.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review*, *20*(3), 709-734. https://doi.org/10.2307/258792.

Mazey, N. (2018). Initial trust in emerging technologies and the effect of threats to privacy.

Mazzù, M. F., Baccelloni, A., Romani, S., & Andria, A. (2022). The role of trust and algorithms in consumers' front-of-pack labels acceptance: a cross-country investigation. *European Journal of Marketing*, *56*(11), 3107-3137. https://doi.org/10.1108/ejm-10-2021-0764.

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information systems research*, *13*(3), 334-359. https://doi.org/10.1287/isre.13.3.334.81.

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

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature machine intelligence*, *1*(11), 501-507. https://doi.org/10.1038/s42256-019-0114-4.

Morik, K., Kotthaus, H., Heppe, L., Heinrich, D., Fischer, R., Pauly, A., & Piatkowski, N. (2021). The care label concept: a certification suite for trustworthy and resource-aware machine learning. *arXiv preprint arXiv:2106.00512*.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and engineering ethics*, *26*(4), 2141-2168. https://doi.org/10.1007/s11948-019-00165-5.

Nordheim, C. B., Følstad, A., & Bjørkli, C. A. (2019). An initial model of trust in chatbots for customer service—findings from a questionnaire study. *Interacting with Computers*, *31*(3), 317-335. https://doi.org/10.1093/iwc/iwz022.

OECD (2019). Scoping the OECD AI Principles: Deliberations of the Expert Group on Artificial Intelligence at the OECD (AIGO). *OECD Digital Economy Papers*. https://doi.org/10.1787/d62f618a-en.

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

Rossi, F. (2018). Building trust in artificial intelligence. *Journal of international affairs*, *72*(1), 127-134. https://doi.org/10.3733/ca.2018a0015.

Russell, S. J., & Norvig, P. (2010). *Artificial intelligence a modern approach*. London.

Salloum, S. A., & Al-Emran, M. (2018). Factors affecting the adoption of E-payment systems by university students: Extending the TAM with trust. International Journal of Electronic Business, 14(4), 371-390. https://doi.org/10.1504/ijeb.2018.10019536.

Scharowski, N., Benk, M., Kühne, S. J., Wettstein, L., & Brühlmann, F. (2023, June). Certification Labels for Trustworthy AI: Insights From an Empirical Mixed-Method Study. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 248-260). https://doi.org/10.1145/3593013.3593994

Schebesta, H. (2019). *Control in the label: Self-declared, certified, accredited? On-pack consumer communication about compliance control in voluntary food schemes from a legal perspective* (pp. 143-161). Springer International Publishing. https://doi.org/10.1007/978-3-030-02499-4_7.

Schuitema, G., Aravena, C., & Denny, E. (2020). The psychology of energy efficiency labels: Trust, involvement, and attitudes towards energy performance certificates in Ireland. *Energy Research & Social Science*, *59*, 101301. https://doi.org/10.1016/j.erss.2019.101301.

Seifert, C., Scherzinger, S., & Wiese, L. (2019). Towards generating consumer labels for machine learning models. In *2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI)* (pp. 173-179). IEEE. https://doi.org/10.1109/CogMI48466.2019.00033.

Sim, M., Kim, S. Y., & Suh, Y. (2022). Sample size requirements for simple and complex mediation models. *Educational and Psychological Measurement*, *82*(1), 76-106. https://doi.org/10.1177/00131644211003261.

Stadelmann, M., & Schubert, R. (2018). How do different designs of energy labels influence purchases of household appliances? A field study in Switzerland. *Ecological economics*, *144*, 112-123. https://doi.org/10.1016/j.ecolecon.2017.07.031.

Storey, V. C., Lukyanenko, R., Parsons, J., & Maass, W. (2022). Explainable AI: Opening the black box or Pandora's box? *Communications of the ACM*, 65(4), 27–29. https://doi.org/10.1145/ 3490699

Stuurman, K., & Lachaud, E. (2022). Regulating AI. A label to complete the proposed Act on Artificial Intelligence. *Computer Law & Security Review*, *44*, 105657. https://doi.org/10.1016/j.clsr.2022.105657.

Taufique, K. M., Nielsen, K. S., Dietz, T., Shwom, R., Stern, P. C., & Vandenbergh, M. P. (2022). Revisiting the promise of carbon labelling. *Nature Climate Change*, *12*(2), 132-140. https://doi.org/10.1038/s41558-021-01271-8.

The Economist. (2019, December 18). *Don't trust AI until we build systems that earn trust.* Retrieved November 17, 2023 from https://www.economist.com/open-future/2019/12/18/dont-trust-ai-until-we-build-systems-that-earn-trust.

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets*, *31*, 447-464. https://doi.org/10.1007/s12525-020-00441-4.

Tonkin, E., Wilson, A. M., Coveney, J., Webb, T., & Meyer, S. B. (2015). Trust in and through labelling–a systematic review and critique. *British Food Journal*, *117*(1), 318-338. https://doi.org/10.1108/BFJ-07-2014-0244.

Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C. G., & Van Moorsel, A. (2020, January). The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 272-283). https://doi.org/10.1145/3351095.3372834.

Tsikriktsis, N. (2005). A review of techniques for treating missing data in OM survey research. *Journal of operations management*, *24*(1), 53-62. https://doi.org/10.1016/j.jom.2005.03.001.

Varshney, K. R. (2019). Trustworthy machine learning and artificial intelligence. *XRDS: Crossroads, The ACM Magazine for Students*, *25*(3), 26-29. https://doi.org/10.1145/3313109.

Vehovar, V., Toepoel, V., & Steinmetz, S. (2016). Non-probability sampling. In C. Wolf, D. Joye, T. W. Smith, & Y. Fu (Eds.), *The SAGE Handbook of Survey Methodology* (Vol. 1, pp. 329-345). SAGE Publications Ltd. https://doi.org/10.4135/9781473957893.n22.

Velčovská, Š., & Del Chiappa, G. (2015). The food quality labels: Awareness and willingness to pay in the context of the Czech Republic. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, *63*(2), 647-658. https://doi.org/10.11118/actaun201563020647.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS quarterly*, 425-478. https://doi.org/10.2307/30036540.

Wilson, C., Ghosh, A., Jiang, S., Mislove, A., Baker, L., Szary, J., ... & Polli, F. (2021, March). Building and auditing fair algorithms: A case study in candidate screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 666-677). https://doi.org/10.1145/3442188.3445928.

Woods, S., Walters, M., Koay, K. L., & Dautenhahn, K. (2006, March). Comparing human robot interaction scenarios using live and video based methods: towards a novel methodological approach. In *9th IEEE International Workshop on Advanced Motion Control, 2006.* (pp. 750-755). IEEE.

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019). Explainable AI: A brief survey on history, research areas, approaches and challenges. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8* (pp. 563-574). Springer International Publishing. https://doi.org/10.1007/978-3-030-32236-6_51.

Yampolskiy, R. V. (2015). *Artificial superintelligence: a futuristic approach*. cRc Press. https://doi.org/10.1201/b18612.

Yang, R., & Wibowo, S. (2020). Risks and uncertainties in citizens' trust and adoption of e-government: A proposed framework. ACIS 2020 Proceedings. 80, https://aisel.aisnet.org/ acis2020/ 80

Yurrita, M., Murray-Rust, D., Balayn, A., & Bozzon, A. (2022, June). Towards a multi-stakeholder value-based assessment framework for algorithmic systems. In *Proceedings of the 2022*

ACM Conference on Fairness, Accountability, and Transparency (pp. 535-563). https://doi.org/10.1145/3531146.3533118.

Zicari, R. V., Brodersen, J., Brusseau, J., Düdder, B., Eichhorn, T., Ivanov, T., ... & Westerlund, M. (2021). Z-Inspection®: a process to assess trustworthy AI. *IEEE Transactions on Technology and Society*, *2*(2), 83-97. https://doi.org/10.1109/TTS.2021.3066209.

Appendices

Appendix 1: Survey 1

Informed consent

Welcome and thank you for participating in this experiment on **trustworthy artificial intelligence (AI)**. I, Joana Glaum, am conducting this experiment as part of my Master Thesis at Católica Lisbon School of Business and Economics, under the supervision of Dr. Filipa de Almeida. This study consists of a scenario and multiple questions related to it. It will take around **4 minutes** to complete. The purpose is to gain insights into users' trust in and acceptance of AI systems. Your participation will contribute to research on trustworthy AI. Please answer as honestly as possible. All answers will be kept strictly confidentially and are anonymous. This means that it will not be possible to link your responses to your identity. The data collected will be used for research purposes only and may be presented in my thesis or disseminated in academic journals, always in an aggregated form, never about any individual response. I ask you to take the study in one go, without interruptions. There are no expected side effects of participating in this study beyond those associated with looking at a computer screen for circa 4 minutes. You may change your mind and drop out at any point of the study during its completion. If you have any questions regarding this study, please do not hesitate to contact me: Joana Glaum (s-jmglaum@ucp.pt). By continuing you agree to participate. Thank you!

Q1 Do you consent to participate in this study?

- ○ I consent (1)
- ○ I do not consent (2)

Understanding of AI

Thank you for consenting to participate. To start, please answer the following question:

Q2 How would you rate your understanding of Artificial Intelligence?

- ○ Extremely bad (1)
- ○ Moderately bad (2)
- ○ Slightly bad (3)
- ○ Neither good or bad (4)
- ○ Slightly good (5)
- ○ Moderately good (6)
- ○ Extremely good (7)

Scenario

If Condition = No Label

Now, I would like to ask you to imagine the scenario described next. Please make an effort to imagine yourself in the described situation and answer as realistically as possible. Thank you!

Consider the situation where you are applying for a new job at a company. The company is using an AI system* called *MyJob* for evaluating job applications. You will be required to fill out a form, uploading your CV, and submit them along with personal information like address, marital status, employment status and references to MyJob. Once assessed, MyJob will determine based on the provided information whether or not you will be invited for an interview.

*AI system stands for 'Artificial intelligence system' which is "*a software that is developed with specific techniques and approaches and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with.*" (OECD, 2019)

Scenario_label

If Condition = Label

Now, I would like to ask you to imagine the scenario described next. Please make an effort to imagine yourself in the described situations and answer as realistically as possible.
Thank you!

Consider the situation where you are applying for a new job at a company. The company is using an AI system* called *MyJob* for evaluating job applications. You will be required to fill out a form, uploading your CV, and submit them along with personal information like address, marital status, employment status

and references to MyJob. Once assessed, MyJob will determine based on the provided information whether or not you will be invited for an interview. MyJob has received a certification label for trustworthy AI, as communicated by the logo below. The label was awarded by a foundation specialized in assessing AI trustworthiness. In this process, a panel of experts verified whether a digital service or product offered by a company (e.g., an AI as in the scenarios presented) meets certain criteria regarding trustworthy AI. These criteria were proposed by the European Commission's high-level expert group on AI. Take your time to get a close look at the label.

*AI system stands for 'Artificial intelligence system' which is "*a software that is developed with specific techniques and approaches and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with*." (OECD, 2019)

Trust in AI

Thank you. Now, please answer the following questions to the best of your knowledge.

Q3 To what extent do you agree to the following statements regarding the presented AI system?

	Strongly disagree (1)	Disagree (2)	Somewhat disagree (3)	Neither agree nor disagree (4)	Somewhat agree (5)	Agree (6)	Strongly agree (7)
MyJob cares about my well-being. (1)	○	○	○	○	○	○	○
MyJob is sincerely concerned about addressing the problems of human users. (2)	○	○	○	○	○	○	○
MyJob tries to be helpful and does not operate out of selfish interest. (3)	○	○	○	○	○	○	○

	Strongly disagree (1)	Disagree (2)	Somewhat disagree (3)	Neither agree nor disagree (4)	Somewhat agree (5)	Agree (6)	Strongly agree (7)
MyJob is truthful in its dealings with me. (4)	○	○	○	○	○	○	○
MyJob keeps its commitments and delivers on its promises. (5)	○	○	○	○	○	○	○
I have never used a smartphone. (6)	○	○	○	○	○	○	○
MyJob is honest and does not abuse the information and advantage it has over its users. (7)	○	○	○	○	○	○	○
MyJob would perform its role in evaluating job applications well. (8)	○	○	○	○	○	○	○
MyJob is competent in evaluating job applications. (9)	○	○	○	○	○	○	○
MyJob has the skills and competence to accurately evaluate job applications. (10)	○	○	○	○	○	○	○
MyJob is reliable in making decisions about job applications. (11)	○	○	○	○	○	○	
MyJob is dependable. (12)	○	○	○	○	○	○	○

Assuming a reality where you have the possibility to choose whether to use MyJob and other AI systems similar to it, please answer the following questions:

Q4 To what extent do you agree to the following statements regarding the **presented AI system**?

	Strongly disagree (1)	Disagree (2)	Somewhat disagree (3)	Neither agree nor disagree (4)	Somewhat agree (5)	Agree (6)	Strongly agree (7)
I intend to use AI systems like MyJob for assessing my job applications in the future. (1)	○	○	○	○	○	○	○
I will use AI systems like MyJob in the future for assessing my job applications. (2)	○	○	○	○	○	○	○
I have concerns about AI systems like MyJob being used for assessing my job applications. (3)	○	○	○	○	○	○	○

Final questions

Thank you for your answers. To end, I would like you to please answer a couple of demographic questions, as well as questions related to the experiment. I emphasize that all answers are anonymous and confidential, which implies that I am unable to link your responses to your person.

Q5 Answering the questions of this study was...

- ○ Extremely difficult (1)
- ○ Moderately difficult (2)
- ○ Slightly difficult (3)
- ○ Neither easy nor difficult (4)
- ○ Slightly easy (5)
- ○ Moderately easy (6)
- ○ Extremely easy (7)

Q6 Imagining the previously described scenario was...

- ○ Extremely difficult (1)
- ○ Moderately difficult (2)
- ○ Slightly difficult (3)
- ○ Neither easy nor difficult (4)
- ○ Slightly easy (5)
- ○ Moderately easy (6)
- ○ Extremely easy (7)

Q7 What is your gender?

- ○ Male (1)
- ○ Female (2)
- ○ Other (3) _____
- ○ Prefer not to say (4)

Q8 How old are you?

_____

Q9 Where are you from?

▼ Afghanistan (1) ... Zimbabwe (1357)

Q10 What is you highest level of education?

- ○ Less than Secondary education (1)
- ○ Secondary education (2)
- ○ Bachelor’s degree (3)
- ○ Master's degree (4)
- ○ Doctoral degree (5)
- ○ Other (6) _____

Q11 What is your current employment status?

- ○ Employed (1)
- ○ Freelancer (2)
- ○ Unemployed (3)
- ○ Student (4)
- ○ Worker and Student (5)
- ○ Retired (6)
- ○ Other (7) _____

Q12 How comfortable are you with the English language?

- ○ Extremely uncomfortable (1)
- ○ Somewhat uncomfortable (2)
- ○ Neither comfortable nor uncomfortable (3)
- ○ Somewhat comfortable (4)
- ○ Extremely comfortable (5)

Q13 How much attention did you pay during this survey?

- ○ None at all (1)
- ○ A little (2)
- ○ A moderate amount (3)
- ○ A lot (4)
- ○ A great deal (5)

Q14 Through which channel or platform did you access this survey?

- ○ E-Mail (1)
- ○ LinkedIn (2)
- ○ Social Media (Instagram, Facebook) (3)
- ○ WhatsApp (4)
- ○ Other (5) _____

Q15 Do you have any comments you would like to share with the researcher? If so, please write them in the box below. Otherwise, just leave it blank

_____

Debriefing

Thank you for your participation in this study. In this study I actually want to study if a label for trustworthy AI affects end-users' trust in and the acceptance of AI systems. For that, I manipulated the scenario by assigning one half of participants to a condition in which the AI system had received a label for trustworthiness, and one half to a control condition (AI system without any label for trustworthiness). I did not disclose the full goal of the tasks you were exposed to as doing so would render the results of the current study not informative.

Appendix 2: Survey 2

Informed consent

Welcome and thank you for participating in this experiment on **trustworthy artificial intelligence (AI)**. I, Joana Glaum, am conducting this experiment as part of my Master Thesis at Católica Lisbon School of Business and Economics, under the supervision of Dr. Filipa de Almeida. This study consists of a scenario

and multiple questions related to them. It will take around **5 minutes** to complete. The purpose is to examine the importance of different requirements for trustworthy AI. Your participation will contribute to research on trustworthy AI. Please answer as honestly as possible. All answers will be kept strictly confidentially and are anonymous. This means that it will not be possible to link your responses to your identity. The data collected will be used for research purposes only and may be presented in my thesis or disseminated in academic journals, always in an aggregated form, never about any individual response. I ask you to take the study in one go, without interruptions. There are no expected side effects of participating in this study beyond those associated with looking at a computer screen for circa 5 minutes. You may change your mind and drop out at any point of the study during its completion. If you have any questions regarding this study, please do not hesitate to contact me: Joana Glaum (s-jmglaum@ucp.pt). By continuing you agree to participate. Thank you!

Q1 Do you consent to participate in this study?

- ○ I consent (1)
- ○ I do not consent (2)

Understanding of AI

Thank you for consenting to participate. To start, please answer the following question:

Q2 How would you rate your understanding of Artificial Intelligence?

- ○ Extremely bad (1)
- ○ Moderately bad (2)
- ○ Slightly bad (3)
- ○ Neither good or bad (4)
- ○ Slightly good (5)
- ○ Moderately good (6)
- ○ Extremely good (7)

Scenario

If Condition = No Label

Now, I would like to ask you to imagine the scenario described next. Please make an effort to imagine yourself in the described situation and answer as realistically as possible. Thank you!

Consider the situation where you are applying for a new job at a company. The company is using an AI system* called *MyJob* for evaluating job applications. You will be required to fill out a form, uploading your CV, and submit them along with personal information like address, marital status, employment status and references to MyJob. Once assessed, MyJob will determine based on the provided information whether or not you will be invited for an interview.

*AI system stands for 'Artificial intelligence system' which is "*a software that is developed with specific techniques and approaches and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with.*" (OECD, 2019)

Scenario_label 1

If Condition = Certification Label 1

Now, I would like to ask you to imagine the scenario described next. Please make an effort to imagine yourself in the described situations and answer as realistically as possible. Thank you!

Consider the situation where you are applying for a new job at a company. The company is using an AI system* called *MyJob* for evaluating job applications. You will be required to fill out a form, uploading your CV, and submit them along with personal information like address, marital status, employment status and references to MyJob. Once assessed, MyJob will determine based on the provided information whether or not you will be invited for an interview. MyJob has received a certification label for trustworthy AI, as communicated by the logo below. The label was awarded by a foundation specialized in assessing AI trustworthiness. In this process, a panel of experts verified whether a digital service or product offered by a company (e.g., an AI as in the scenarios presented) meets certain criteria regarding trustworthy AI. These criteria were proposed by the European Commission's high-level expert group on AI. Take your time to get a close look at the label.

*AI system stands for 'Artificial intelligence system' which is "*a software that is developed with specific techniques and approaches and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with*." (OECD, 2019)

Scenario_label 2

If Condition = Certification Label 2 (with TAI requirements)

Now, I would like to ask you to imagine the scenario described next. Please make an effort to imagine yourself in the described situations and answer as realistically as possible. Thank you!

Consider the situation where you are applying for a new job at a company. The company is using an AI system* called *MyJob* for evaluating job applications. You will be required to fill out a form, uploading your CV, and submit them along with personal information like address, marital status, employment status and references to MyJob. Once assessed, MyJob will determine based on the provided information whether or not you will be invited for an interview. MyJob has received a certification label for trustworthy AI, as communicated by the logo on the left side. The label was awarded by a foundation specialized in assessing AI trustworthiness. In this process, a panel of experts verified whether a digital service or product offered by a company (e.g., an AI as in the scenarios presented) meets certain criteria regarding trustworthy AI. These criteria were proposed by the European Commission's High-level Expert on AI. The criteria are listed on the right side next to the label. Take your time to get a close look at the label.

*AI system stands for 'Artificial intelligence system' which is "*a software that is developed with specific techniques and approaches and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with*." (OECD, 2019)

Trust in AI

Thank you. Now, please answer the following questions to the best of your knowledge.

Q3 To what extent do you agree to the following statements regarding the presented AI system

	Strongly disagree (1)	Disagree (2)	Somewhat disagree (3)	Neither agree nor disagree (4)	Somewhat agree (5)	Agree (6)	Strongly agree (7)
MyJob cares about my well-being. (1)	○	○	○	○	○	○	○
MyJob is sincerely concerned about addressing the problems of human users. (2)	○	○	○	○	○	○	○
MyJob tries to be helpful and does not operate out of selfish interest. (3)	○	○	○	○	○	○	○
MyJob is truthful in its dealings with me. (4)	○	○	○	○	○	○	○
MyJob keeps its commitments and delivers on its promises. (5)	○	○	○	○	○	○	○
I have never used a smartphone. (6)	○	○	○	○	○	○	○
MyJob is honest and does not abuse the information and advantage it has over its users. (7)	○	○	○	○	○	○	○
MyJob would perform its role in evaluating job applications well. (8)	○	○	○	○	○	○	○
MyJob is competent in evaluating job applications. (9)	○	○	○	○	○	○	○
MyJob has the skills and competence to accurately evaluate job applications. (10)	○	○	○	○	○	○	○

	Strongly disagree (1)	Disagree (2)	Somewhat disagree (3)	Neither agree nor disagree (4)	Somewhat agree (5)	Agree (6)	Strongly agree (7)
MyJob is reliable in making decisions about job applications. (11)	○	○	○	○	○	○	○
MyJob is dependable. (12)	○	○	○	○	○	○	○

Assuming a reality where you have the possibility to choose whether to use MyJob and other AI systems similar to it, please answer the following questions:

Q4 To what extent do you agree to the following statements regarding the **presented AI system**?

	Strongly disagree (1)	Disagree (2)	Somewhat disagree (3)	Neither agree nor disagree (4)	Somewhat agree (5)	Agree (6)	Strongly agree (7)
I intend to use AI systems like MyJob for assessing my job applications in the future. (1)	○	○	○	○	○	○	○
I will use AI systems like MyJob in the future for assessing my job applications. (2)	○	○	○	○	○	○	○
I have concerns about AI systems like MyJob being used for assessing my job applications. (3)	○	○	○	○	○	○	○

Criteria for TAI

Thank you! In the following, I would like to ask you to rate the importance of the seven requirements for trustworthy AI.

Q5 How important are these values for you in the design of trustworthy AI systems that interact with us?

	Not important at all (1)	Not important (2)	Somewhat unimportant (3)	Neither unimportant nor important (4)	Somewhat important (5)	Important (6)	Extremely important (7)
Human agency and oversight: There is human oversight and control throughout the lifecycle of AI-systems. (1)	○	○	○	○	○	○	○
Privacy and data governance: Competent authorities are involved in the implementation of legal frameworks and guidelines for testing and certification of AI-enabled products and services. (2)	○	○	○	○	○	○	○
Technical robustness and safety: AI-systems are developed in a responsible manner with proper consideration of risks. (3)	○	○	○	○	○	○	○
Transparency: There are transparency requirements that reduce the opacity of AI-systems. (4)	○	○	○	○	○	○	○

Q5 How important are these values for you in the design of trustworthy AI systems that interact with us?

	Not important at all (1)	Not important (2)	Somewhat unimportant (3)	Neither unimportant nor important (4)	Somewhat important (5)	Important (6)	Extremely important (7)
Diversity, non-discrimination and fairness: Rules designed to protect fundamental human rights, such as equality, are applied in the conception of AI-systems. (5)	○	○	○	○	○	○	○
Societal and environmental well-being: AI-systems are programmed to conform to the best standards of sustainability and address like issues climate change and environmental justice. (6)	○	○	○	○	○	○	○
Accountability: AI-systems are programmed to, at any step, be accountable for considering their impact in the world. (7)	○	○	○	○	○	○	○

Q6 Are there any other factors or criteria not mentioned in this survey that you believe are crucial for determining the trustworthiness of AI systems?

_____

Final questions

Thank you for your answers. To end, I would like you to please answer a couple of demographic questions, as well as questions related to the experiment. I emphasize that all answers are anonymous and confidential, which implies that I am unable to link your responses to your person.

Q7 Answering the questions of this study was...

- ○ Extremely difficult (1)
- ○ Moderately difficult (2)
- ○ Slightly difficult (3)
- ○ Neither easy nor difficult (4)
- ○ Slightly easy (5)
- ○ Moderately easy (6)
- ○ Extremely easy (7)

Q8 Imagining the previously described scenario was...

- ○ Extremely difficult (1)
- ○ Moderately difficult (2)
- ○ Slightly difficult (3)
- ○ Neither easy nor difficult (4)
- ○ Slightly easy (5)
- ○ Moderately easy (6)
- ○ Extremely easy (7)

Q9 What is your gender?

- ○ Male (1)
- ○ Female (2)
- ○ Other (3) _____
- ○ Prefer not to say (4)

Q10 How old are you?

_____

Q11 Where are you from?

▼ Afghanistan (1) ... Zimbabwe (1357)

Q12 What is you highest level of education?

- ○ Less than Secondary education (1)
- ○ Secondary education (2)
- ○ Bachelor’s degree (3)
- ○ Master's degree (4)
- ○ Doctoral degree (5)
- ○ Other (6) _____

Q13 What is your current employment status?

- ○ Employed (1)
- ○ Freelancer (2)
- ○ Unemployed (3)
- ○ Student (4)
- ○ Worker and Student (5)
- ○ Retired (6)
- ○ Other (7) _____

Q14 How comfortable are you with the English language?

- ○ Extremely uncomfortable (1)
- ○ Somewhat uncomfortable (2)
- ○ Neither comfortable nor uncomfortable (3)
- ○ Somewhat comfortable (4)
- ○ Extremely comfortable (5)

Q15 How much attention did you pay during this survey?

- ○ None at all (1)
- ○ A little (2)
- ○ A moderate amount (3)
- ○ A lot (4)
- ○ A great deal (5)

Q16 Through which channel or platform did you access this survey?

- ○ E-Mail (1)
- ○ LinkedIn (2)
- ○ Social Media (Instagram, Facebook) (3)
- ○ WhatsApp (4)
- ○ Other (5) _____

Q17 Do you have any comments you would like to share with the researcher? If so, please write them in the box below. Otherwise, just leave it blank

_____

Debriefing

Thank you for your participation in this study. In this study I actually want to study if a label for trustworthy AI affects end-users' trust in and the acceptance of AI systems. Moreover, I want to measure the importance of the seven requirements for trustworthy AI that were proposed by the European Commission's high-level

expert group for AI.

For that, I manipulated the scenario by assigning one third of participants to a condition in which the AI system had received a label for trustworthiness, but without communicating the trustworthy requirements; one third to a condition where the label displayed the requirements for trustworthy AI, and one third to a control condition (AI system without any label for trustworthiness and communication of requirements).

On the basis of the results of the survey, I aim to examine which of the seven requirements are most powerful in communicating trustworthiness to end-users of AI systems.

I did not disclose the full goal of the tasks you were exposed to as doing so would render the results of the current study not informative.

Data Analysis Study 1

Appendix 3: Demographics

Descriptive Statistics for Age

	N	Minimum	Maximum	Mean	Std. Deviation
How old are you?	121	21.00	60.00	27.6446	5.86779
Valid N (listwise)	121				

Frequency table for Gender

		N	%
Male		51	39.5%
Female		75	58.1%
Missing	*System*	3	2.3%

Frequency table for Education

		N	%
Secondary education		4	3.1%
Bachelor's degree		64	49.6%
Master's degree		51	39.5%
Doctoral degree		1	0.8%
Other		1	0.8%
Missing	System	8	6.2%

Frequency table for Employment Status

	N	%
Employed	59	45.7%
Freelancer	4	3.1%
Unemployed	1	0.8%
Student	41	31.8%
Worker and Student	21	16.3%
Missing System	3	2.3%

Frequency table for Country of Residence

	N	%
Austria	2	1.6%
Belgium	5	3.9%
Denmark	4	3.1%
France	1	0.8%
Germany	87	67.4%
Netherlands	2	1.6%
New Zealand	1	0.8%
Poland	2	1.6%
Portugal	13	10.1%
Switzerland	1	0.8%
Missing System	11	8.5%

Descriptive Statistics for Understanding of AI

	N	Minimum	Maximum	Mean	Std. Deviation
How would you rate your understanding of Artificial Intelligence?	128	2	7	5.30	1.07
Valid N (listwise)	128				

Appendix 4: Scale Reliability

A reliability analysis was conducted on the full sample of Study 1 to test for the Cronbach's alpha (α) of the three scales used. According to Hair et al. (2013), a construct is considered reliable if the α value is greater than .70. The results of the reliability analysis revealed that both the affective trust scale (α = .86) and the cognitive trust scale (α = .93) showed high internal consistency between the items and are reliable enough to predict the respective variables. Similarly, the DV scale was also found reliable (α = .87). The reliability analysis was performed on the full sample, before the data cleaning.

Reliability Statistics

	Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
Affective Trust	.86	.86	6
Cognitive Trust	.93	.93	5
Behavioral intention to use	.87	.87	3

Appendix 5: Mixed ANOVA

A mixed ANOVA was performed on the full sample to evaluate the effect of the label scenario (absence vs. presence) on end-users' trust (cognitive and affective) and their behavioral intention to use the AI. The effect of the different label conditions on end-users' trust was significant at the 5%-level, $F(1, 134) = 27.14$, $p < .001$, such that the certification label had a greater influence on both cognitive ($M = 5.11$, $SD = 1.28$, $t(134) = 4.91$, $p < .001$) and affective trust ($M = 4.64$, $SD = 1.01$, $t(134) = 4.79$, $p < .001$) in the AI system, compared to the scenario where the AI did not receive any label for TAI (Affective: $M = 3.81$, $SD = 1.14$; Cognitive: $M = 4.20$, $SD = 1.15$). This supports **H1a** and since the label leads to higher levels of cognitive than effective trust, **H1b** is also supported. Moreover, the effect of the AI label on behavioral intention to use ($M = 4.55$, $SD = 1.37$) was above and beyond the effect of the control scenario ($M = 3.42$, $SD = 1.31$, $t(134) = 5.03$, $p < .001$), which supports **H2**.

Descriptive Statistics

	Scenario	Mean	Std. Deviation	N
Affective Trust	.00	3.81	1.14	68
	1.00	4.64	1.01	68
	Total	4.28	1.15	136
Cognitive Trust	.00	4.2	1.15	68
	1.00	5.11	1.28	68
	Total	4.66	1.29	136
Behavioral_Trust	.00	3.43	1.32	68
	1.00	4.56	1.37	68
	Total	3.99	1.46	136

Tests of Between-Subjects Effects

Measure: MEASURE_1
Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[a]
Intercept	7519,86	1	7519.86	2191.51	<.001	.94	2191.51	1.000
Scenario	93,13	1	93.13	27.14	<.001	.17	27.14	.999
Error	459,80	134	3.43					

a. Computed using alpha = .05

Appendix 6: Hayes PROCESS Model 4

A multiple mediation analysis was performed to test **H3,** whether cognitive and affective trust significantly mediate the relationship between the label scenario and the behavioral intention to use the AI. The Hayes PROCESS Model 4 was performed on the full sample and the covariate "understanding of AI" was not included. Table 1 shows the results for the a-path, and Table 2 for the b-path. The bootstrap confidence interval for the indirect effect of affective trust was entirely above zero (CI [.15, .63]), suggesting that affective trust mediates the association between the label scenario and the acceptance of the AI system. The same mediating effect was found for cognitive trust on the relationship between the label scenario and behavioral intention to use as the bootstrap

intervall did not include zero (CI [.20, .75]). Therefore, **H3** was supported. As the contrast indicates that the difference between the mediation effect is significant (C1 CI [.13, .41]), and according to the b-paths stronger for cognitive (b1 = .48) than affective trust (b2 = .44), **H3b** can also be supported.

Variable		**Affective trust (M1)**					**Cognitive trust (M2)**			
		B	***SE***	***p***	***β***		***B***	***SE***	***p***	***β***
X (label scenario)	*a1*	.83	.18	.000	.72	*a2*	.91	.21	.000	.69
		$R^2 = .1303$					$R^2 = .1227$			
		$F(1,134) = 20.07, p < .001$					$F(1,134) = 18.74, p < 001$			

Table 1: Hayes PROCESS Model 4 a-path results

Variable		**Behavioral intention to use (Y)**			
		B	***SE***	***p***	***β***
X (label scenario)	*c*	.33	.17	.049	.22
M1 (Affective Trust)	*b1*	.44	.10	.000	.35
M2 (Cognitive Trust)	*b2*	.48	.09	.000	.43
		$R^2 = .5924$			
		$F(3,132) = 63.95, p < .001$			

Table 2: Hayes PROCESS Model 4 b-path results

Data Analysis Study 2

Appendix 7: Demographics

Descriptive Statistics for Age

	N	Minimum	Maximum	Mean	Std. Deviation
How old are you?	218	19.00	60.00	28.03	6.71
Valid N (listwise)	218				

Frequency table for Gender

		N	%
Male		74	33.5%
Female		144	65.2%
Missing	System	3	1.4%

Frequency table for Education

		N	%
Less than Secondary education		2	0.9%
Secondary education		10	4.5%
Bachelor’s degree		107	48.4%
Master's degree		89	40.3%
Doctoral degree		6	2.7%
Other		1	0.5%
Missing	System	6	2.7%

Frequency table for Employment status

		N	%
Employed		108	48.9%
Freelancer		8	3.6%
Unemployed		3	1.4%
Student		78	35.3%
Worker and Student		20	9.0%
Retired		1	0.5%
Missing	System	3	1.4%

Frequency table for Country of Residence

		N	%
Afghanistan		2	0.9%
Albania		1	0.5%
Australia		1	0.5%
Austria		19	8.6%
Belgium		7	3.2%
Croatia		1	0.5%
Czech Republic		2	0.9%
Denmark		8	3.6%
France		9	4.1%
Germany		139	62.9%
Hungary		1	0.5%
Nepal		1	0.5%
Netherlands		1	0.5%
Norway		1	0.5%
Portugal		9	4.1%
Zimbabwe		1	0.5%
Missing	System	18	8.1%

Descriptive Statistics for Understanding of AI

	N	Minimum	Maximum	Mean	Std. Deviation
How would you rate your understanding of Artificial Intelligence?	221	1	7	5.48	.91
Valid N (listwise)	221				

Appendix 8 : Scale Reliability

A reliability analysis was conducted on the full sample of Study 2 to test for the Cronbach's α of the three scales used. The results of the reliability analysis revealed that the affective trust scale (α = .94) and the cognitive trust scale (α = .97) were found reliable since the α value is greater than .70. Thus, both constructs showed high internal consistency between the items and are reliable enough to predict the respective variables. Similarly, the DV scale was also found reliable (α = .92).

Reliability Statistics

	Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
Affective Trust	.94	.94	6
Cognitive Trust	.97	.97	5
Behavioral intention to use	.92	.93	3

Appendix 9 : Mixed ANOVA

The mixed ANOVA revealed a significant difference between the effect of the independent variable (no label vs. certification label vs. label with requirements) on user trust in AI (cognitive and affective trust), $F(2, 219)= 172.94$, $p < .001$. A planned contrast indicated that the effect of the certification label 2 on affective as well as cognitive trust was above and beyond the effect of the AI system without any label (Affective trust: $t(139) = 4.12$, $p < .001$; Cognitive trust: $t(145) = 4.20$, $p < .001$). The same effect was found between the effect of the AI with the certification label 1 and the control condition (Affective trust: $t(120) = 10.89$, $p < .001$; Cognitive trust: $t(141) = 12.38$, $p < .001$). Moreover, the effect of the certification label 2 on cognitive as well as affective trust was above and beyond the effect of the AI system with the certification label 1 (Affective trust: $t(138) = 13.27$, $p < .001$; Cognitive trust: $t(139) = 16.70$, $p < .001$).

The results demonstrate that the certification label 2 had the greatest influence on both cognitive ($M = 6.29$; $SD = .82$) and affective trust ($M = 5.28$, $SD = .75$) in the AI system, followed by the certification label from Study 1 (Cognitive: $M = 5.70$, $SD = .87$; Affective: $M = 4.83$; $SD = .57$) and the AI system without any label had the lowest impact on cognitive ($M = 3.78$; $SD = .99$) and affective trust ($M = 3.44$, $SD = .92$). Thus, the results of the mixed ANOVA support the findings from Study 1 as well as **H1a**. Moreover, the findings suggest that displaying trustworthiness requirements on the label translates into higher trust in the AI system. Therefore, **H4** is also supported. As in Study 1, it can also be derived from the results that in both cases the label leads to higher levels of cognitive than effective trust. **H1b** is therefore supported.

Descriptive Statistics

	Scenario	Mean	Std. Deviation	N
Affective_Trust	.00	3.45	.92	73
	1.00	4.83	.57	73
	2.00	5.29	.75	76
	Total	4.53	1.09	222
Cognitive_Trust	.00	3.79	.99	73
	1.00	5.71	.87	73
	2.00	6.29	.82	76
	Total	5.28	1.39	222
Behavioral_Trust	.00	3.02	.88	73
	1.00	5.09	.82	73
	2.00	5.37	.94	76
	Total	4.50	1.37	222

Tests of Between-Subjects Effects

Measure: MEASURE_1
Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[a]
Intercept	15087.37	1	15087.37	8428.64	<.001	.98	8428.64	1.00
Scenario	619.14	2	309.57	172.94	<.001	.61	345.88	1.00
Error	392.01	219	1.79					

a. Computed using alpha = .05

Appendix 10 : Hayes PROCESS Model 4

To test the robustness of the findings from Study 1, a simple mediation analysis was performed by the PROCESS SPSS macro to test whether **H3** also holds true for the data of Study 2. The Hayes PROCESS Model 4 was performed on the full sample and the covariate "understanding of AI" was not included. Table 1 shows the results for the a-path, and Table 2 for the b-path. Using indicator coding, the control group was set as reference group and the treatments group were compared to it. The bootstrap confidence interval for the indirect effect of affective trust based on 5,000 bootstrap resamples was entirely above zero (X1: CI [.101, .71]; X2: CI [1.02, 1.87]), suggesting that

affective trust mediates the association between the label scenarios and the acceptance of the AI system. The same mediating effect was found for cognitive trust on the relationship between the label scenarios and the DV as the bootstrap intervall did not include zero (X1: CI [1.27, 1.48]; X2: CI [1.02, 1.87]). Thus, support was found for **H3** that cognitive and affective trust significantly mediated the relationship between the label scenarios and the behavioral intention to use the AI. As the contrast indicates that the difference between the mediation effect is significant (C1 CI [.22, .58]), and according to the b-paths stronger for cognitive (b1 = .57) than affective trust (b2 = .28), **H3b** can also be supported. The results also contributes to the robustness of the findings of Study 1.

Variable		**Affective trust (M1)**					**Cognitive trust (M2)**			
		B	***SE***	***p***	***β***		***B***	***SE***	***p***	***β***
X1 (label)	*a1*	1.38	.12	.000	1.26	*a2*	1.92	.14	.000	1.37
X2 (label with requirements)	*a1a*	1.83	.12	.000	1.68	*a2a*	2.50	.14	.000	1.79
		$R^2 = .5151$					$R^2 = .5891$			
		$F(2, 219) = 116.29, p < .001$					$F(2, 219) = 157.00, p < .001$			

Table 1: Hayes PROCESS Model 4 a-path results

Variable		**Behavioral intention to use (Y)**			
		B	***SE***	***p***	***β***
X1 (label)	*c1*	.59	.12	.000	.43
X2 (label with requirements)	*c2*	.41	.13	.0039	.30
M1 (Affective Trust)	*b1*	.28	.07	.0003	.23
M2 (Cognitive Trust)	*b2*	.57	.06	.000	.58
		$R^2 = .8379$			
		$F(4, 217) = 280.33, p < .001$			

Table 2: Hayes PROCESS Model 4 b-path results

Appendix 11: Repeated measures ANOVA

To test whether one of the seven requirements for TAI was perceived as especially important by end-users regarding their interaction with AI systems, a repeated measures ANOVA was performed on the data of the participants in the control group before the elimination of participants. This approach was chosen because in the control group, an influence of the AI labels on the importance perception of the requirements could be prevented. The Mauchly's Test of Sphericity indicated that the assumption of sphericity is violated, $p < .001$, however the Greenhouse Geisser test indicates a significant difference at a 5% level between the importance of the requirements, $p < .001$. From the results of the repeated measures ANOVA, it becomes evident that *Transparency* is of greatest importance for end-users regarding their interaction with AI systems ($M = 6.32$, $SD = .671$). Moreover, the importance of this requirement was significantly different to those of the other requirements for TAI (*MD* to second highest requirement = .76, $p < .001$). *Privacy and Data Governance* was rated as the second most important requirement for TAI ($M = 5.55$, $SD = .713$), however, it was not significantly more important than the third requirement ($MD = .30$, $p < .114$). Thus, **H5** can only partially be supported.

Descriptive Statistics

	Mean	Std. Deviation	N
Human agency and oversight	4.70	.90	71
Privacy and data governance	5.55	.71	71
Technical robustness and safety	5.24	.73	71
Transparency	6.32	.67	71
Diversity, non-discrimination and fairness	5.18	.78	71
Societal and environmental well-being	4.49	1.22	71
Accountability	4.87	.92	71

Pairwise Comparisons Table

Measure: MEASURE 1

(I) Requirements	(J) Requirements	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
Human agency and oversight	2	-.859*	.119	<.001	-1.235	-.483
	3	-.563*	.111	<.001	-.914	-.213

	4	-1.634*	.133	<.001	-2.054	-1.213
	5	-.493*	.118	.002	-.866	-.119
	6	.197	.148	1.000	-.270	.665
	7	-.183	.137	1.000	-.614	.247
Privacy and Data Governance	1	.859*	.119	<.001	.483	1.235
	3	.296	.103	.114	-.029	.621
	4	-.775*	.090	<.001	-1.059	-.490
	5	.366*	.110	.030	.018	.714
	6	1.056*	.150	<.001	.584	1.529
	7	.676*	.123	<.001	.287	1.065
Technical Robustness and Safety	1	.563*	.111	<.001	.213	.914
	2	-.296	.103	.114	-.621	.029
	4	-1.070*	.109	<.001	-1.413	-.728
	5	.070	.093	1.000	-.222	.362
	6	.761*	.121	<.001	.379	1.142
	7	.380*	.101	.007	.062	.699
Transparency	1	1.634*	.133	<.001	1.213	2.054
	2	.775*	.090	<.001	.490	1.059
	3	1.070*	.109	<.001	.728	1.413
	5	1.141*	.116	<.001	.776	1.506
	6	1.831*	.168	<.001	1.302	2.360
	7	1.451*	.120	<.001	1.072	1.829
Diversity, non-discrimination and fairness	1	.493*	.118	.002	.119	.866
	2	-.366*	.110	.030	-.714	-.018
	3	-.070	.093	1.000	-.362	.222
	4	-1.141*	.116	<.001	-1.506	-.776
	6	.690*	.111	<.001	.340	1.040
	7	.310*	.084	.009	.045	.575
Societal and environmental well-being	1	-.197	.148	1.000	-.665	.270
	2	-1.056*	.150	<.001	-1.529	-.584
	3	-.761*	.121	<.001	-1.142	-.379
	4	-1.831*	.168	<.001	-2.360	-1.302
	5	-.690*	.111	<.001	-1.040	-.340
	7	-.380*	.097	.004	-.686	-.075
Accountability	1	.183	.137	1.000	-.247	.614
	2	-.676*	.123	<.001	-1.065	-.287
	3	-.380*	.101	.007	-.699	-.062
	4	-1.451*	.120	<.001	-1.829	-1.072

	5	-.310*	.084	.009	-.575	-.045
	6	.380*	.097	.004	.075	.686

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.