

Universität
Basel

Fakultät für
Psychologie



Trust in Artificial Intelligence: Understanding and Calibrating Trust *and* Distrust in the Human–AI Interaction

Inauguraldissertation zur Erlangung der Würde eines Doktors der Philosophie vorgelegt der
Fakultät für Psychologie der Universität Basel von

Nicolas Dario Scharowski

aus Luzern (LU), Schweiz

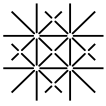
Basel, 2024

Originaldokument gespeichert auf dem Dokumentenserver der Universität Basel

edoc.unibas.ch



Dieses Werk ist lizenziert unter einer [Creative Commons Namensnennung-Nicht kommerziell 4.0](https://creativecommons.org/licenses/by-nc/4.0/)



Universität
Basel

Fakultät für
Psychologie



Genehmigt von der Fakultät für Psychologie auf Antrag von

Prof. Dr. Klaus Opwis (Erstgutachter)

Prof. Dr. Philipp Wintersberger (Zweitgutachter)

Datum des Doktoratsexamen: Freitag, 13. September 2024

Dekan:in der Fakultät für Psychologie



Erklärung zur wissenschaftlichen Lauterkeit

Ich erkläre hiermit, dass die vorliegende Arbeit ohne die Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel selbstständig verfasst ist. Zu Hilfe genommene Quellen sind als solche gekennzeichnet. Die veröffentlichten oder zur Veröffentlichung in Zeitschriften eingereichten Manuskripte wurden in Zusammenarbeit mit den Koautoren erstellt und von keinem der Beteiligten an anderer Stelle publiziert, zur Publikation eingereicht, oder einer anderen Prüfungsbehörde als Qualifikationsarbeit vorgelegt. Es handelt sich dabei um folgende Manuskripte:

1. **Scharowski, N.**, Perrig, S. A. C., Svab, M., Opwis, K., & Brühlmann, F. (2023). Exploring the effects of human-centered AI explanations on trust and reliance. *Frontiers in Computer Science*, 5. <https://doi.org/10.3389/fcomp.2023.1151150>.
2. **Scharowski, N.**, Perrig, S. A. C., Aeschbach, L. F., von Felten, N., Opwis, K., Wintersberger, P., & Brühlmann, F. (2023). To trust or distrust trust measures: Validating questionnaires for trust in AI. *Manuscript submitted for publication*.
3. **Scharowski, N.**, Benk, M., Kühne, S. J., Wettstein, L., & Brühlmann, F. (2023). 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). New York, NY, USA: ACM. <https://doi.org/10.1145/3593013.3593994>



Spezifizierung des eigenen Forschungsbeitrags zu den Manuskripten:

1. Eigener Beitrag nach [CRediT](#)¹:

- | | | |
|--|---|---|
| <input checked="" type="checkbox"/> Conceptualization | <input checked="" type="checkbox"/> Data curation | <input checked="" type="checkbox"/> Formal Analysis |
| <input type="checkbox"/> Funding acquisition | <input checked="" type="checkbox"/> Investigation | <input checked="" type="checkbox"/> Methodology |
| <input checked="" type="checkbox"/> Project administration | <input type="checkbox"/> Resources | <input type="checkbox"/> Software |
| <input type="checkbox"/> Supervision | <input type="checkbox"/> Validation | <input checked="" type="checkbox"/> Visualization |
| <input checked="" type="checkbox"/> Writing – original draft | | |
| <input checked="" type="checkbox"/> Writing – review & editing | | |

Das Manuskript wurde bisher für keine anderen Qualifikationsarbeiten eingereicht.

2. Eigener Beitrag nach [CRediT](#)¹:

- | | | |
|--|---|---|
| <input checked="" type="checkbox"/> Conceptualization | <input checked="" type="checkbox"/> Data curation | <input checked="" type="checkbox"/> Formal Analysis |
| <input type="checkbox"/> Funding acquisition | <input checked="" type="checkbox"/> Investigation | <input checked="" type="checkbox"/> Methodology |
| <input checked="" type="checkbox"/> Project administration | <input type="checkbox"/> Resources | <input type="checkbox"/> Software |
| <input type="checkbox"/> Supervision | <input type="checkbox"/> Validation | <input checked="" type="checkbox"/> Visualization |
| <input checked="" type="checkbox"/> Writing – original draft | | |
| <input checked="" type="checkbox"/> Writing – review & editing | | |

Das Manuskript wurde bisher für keine anderen Qualifikationsarbeiten eingereicht.

3. Eigener Beitrag nach [CRediT](#)¹:

- | | | |
|--|---|---|
| <input checked="" type="checkbox"/> Conceptualization | <input checked="" type="checkbox"/> Data curation | <input checked="" type="checkbox"/> Formal Analysis |
| <input checked="" type="checkbox"/> Funding acquisition | <input checked="" type="checkbox"/> Investigation | <input checked="" type="checkbox"/> Methodology |
| <input checked="" type="checkbox"/> Project administration | <input type="checkbox"/> Resources | <input type="checkbox"/> Software |
| <input checked="" type="checkbox"/> Supervision | <input type="checkbox"/> Validation | <input checked="" type="checkbox"/> Visualization |
| <input checked="" type="checkbox"/> Writing – original draft | | |
| <input checked="" type="checkbox"/> Writing – review & editing | | |

Das Manuskript wurde bisher für keine anderen Qualifikationsarbeiten eingereicht.

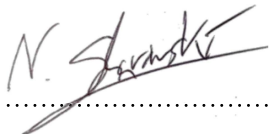
¹ <https://casrai.org/credit/>



Open-Science Aspekte der Manuskripte:

1. Preregistration: ja nein
 Open-Access-Publikation: ja nein
 Open-Access-Data/Analyse: ja nein
 Ort/URL der Daten und Analysen: <https://osf.io/bs6q3/>
2. Preregistration: ja nein
 Open-Access-Publikation: ja nein
 Open-Access-Data/Analyse: ja nein
 Ort/URL der Daten und Analysen: <https://osf.io/7cdne/>
3. Preregistration: ja nein
 Open-Access-Publikation: ja nein
 Open-Access-Data/Analyse: ja nein
 Ort/URL der Daten und Analysen: <https://osf.io/gzp5k/>

Ort, Datum Basel, 17. September 2024

 Signatur 

 Vorname Nachname Nicolas Dario Scharowski

Contents

Abstract	7
Introduction	8
Theoretical Background	12
Understanding Trust in AI	12
Defining Trust in AI	12
Modeling Trust in AI	14
Measuring Trust in AI	20
Calibrating and Increasing Warranted Trust and Distrust in AI	23
Summary of the Manuscripts	27
Manuscript 1: "Exploring the Effects of Human-Centered AI Explanations on Trust and Reliance"	30
Manuscript 2: "To Trust or Distrust Trust Measures: Validating Questionnaires for Trust in AI"	36
Manuscript 3: "Certification Labels for Trustworthy AI: Insights from an Empirical Mixed-Method Study"	41
General Discussion	46
Identifying and Addressing Challenges in Understanding Trust in AI	46
From Improved Measures to a Different Perspective on Trust <i>and</i> Distrust	49
Moving beyond XAI to Novel Approaches for Trustworthy AI	51
Limitations and Future Directions	56
Conclusion	58
Acknowledgments	81
Appendix	82

Abstract

Trust has emerged as a key measure in human–AI interaction in recent years. Trust in AI is recognized not only in academic research but also in industry and politics, where it is often considered a remedy for issues related to fairness, accountability, and transparency in AI systems. This importance of trust necessitates a thorough understanding of it. In particular, research into explainable AI (XAI) has suggested that explanations and other forms of transparency can increase trust in AI. However, the empirical evidence for this assumption is inconclusive. This dissertation explores potential reasons for this ambiguity and aims to contribute to a better understanding of end-users’ trust in AI. As such, manuscript 1 investigates post-hoc explanations and highlights the distinction between trust and behavioral measures of reliance, while also emphasizing the importance of human-related factors in AI-assisted decision-making. Manuscript 2 presents the first comprehensive validation study of trust questionnaires in the context of AI, advocating to consider both trust and distrust. Manuscript 3 explores certification labels as an alternative approach beyond traditional XAI methods to increase end-users’ warranted trust, demonstrating their potential. Overall, this dissertation seeks to provide a more holistic understanding of trust in AI by illustrating how to increase calibrated trust *and* distrust to a level warranted by the AI’s trustworthiness.

Introduction

Computers have long surpassed human capabilities in areas like math, logic, and information storage. However, with the advancement of artificial intelligence (AI), machines are now advancing into areas previously thought to be the exclusive domains of human competence. While AI-based systems have been used for autonomous driving (Wintersberger et al., 2018), product and price recommendations (Scharowski, Perrig, Svab, et al., 2023), and facial recognition in surveillance (Almeida et al., 2022) for some time, in today's technological landscape, AI also stands out as a generative force that is redefining the boundaries of machine capabilities. Modern generative AI can not only recognize patterns in data and draw conclusions from them but also use the patterns learned from the training inputs to generate new data, including texts (e.g., GPT-4), images (e.g., DALL-E), videos (e.g., SORA) and audio (e.g., WaveNet).

As the capabilities of AI continue to increase, there is also growing concern about maintaining human control over these intelligent technologies (Shneiderman, 2020).

This concern has been emphasized explicitly as a subgoal in the "seven grand challenges" for human-computer interaction (HCI) (Stephanidis et al., 2019). To ensure human control, ethical considerations for these systems — such as fairness, accountability, and transparency — have been proposed (Kaur et al., 2022; Lepri et al., 2018). Among these, the call for transparent AI has led to the emergence of dedicated multidisciplinary research areas, such as explainable artificial intelligence (XAI), which aims to provide understandable explanations of AI's decisions, actions, and processes (Liao et al., 2020; Lipton, 2018).

Developing transparent systems to ensure human control has long been a research focus in HCI (Shneiderman, 2020). It has drawn on research on expert systems, intelligent agents, recommender systems, and other adjacent fields such as automation (Abdul et al., 2018). However, as AI grows in complexity, ensuring transparency has become increasingly challenging (Burrell, 2016). Because of this complexity, AI is often characterized as a *opaque-box* (Suresh et al., 2021), meaning that AI can only be understood in terms of its inputs and outputs, while its internal mechanisms remain

opaque and inscrutable to direct observation (Burrell, 2016). The XAI community has thus proposed and investigated various methods for opening this opaque-box¹ or at least for explaining the outcomes generated by such systems in a post-hoc manner (see: Arrieta et al., 2020; Molnar, 2019; Speith, 2022, for a taxonomy and overview).

While addressing *how* AI can be made transparent is critical, it is equally important to consider *who* interacts with these systems (Ehsan, Passi, et al., 2021; Ehsan, Wintersberger, Liao, et al., 2021). The requirements, goals, and objectives of XAI can vary considerably depending on the specific task at hand and the stakeholders involved, who include developers, regulators, domain experts, and end-users (Arrieta et al., 2020; Langer et al., 2021; Suresh et al., 2021). This holistic perspective on the socio-technical environment in which people interact with AI systems has led to research efforts on human-centered explainable AI (HCXAI) (Ehsan, Wintersberger, Liao, et al., 2021; Ehsan et al., 2022). While it is essential to assess the specific needs and impact of XAI methods across these groups, this dissertation and its accompanying manuscripts primarily focus on end-users. This focus is important, as end-users have often been neglected by the developers of AI systems (Cheng et al., 2019; Du et al., 2019) and past research efforts in XAI have mainly used researchers' assumptions about what constitutes AI explanations without a thorough understanding of how end-users define, generate, select and evaluate explanations (Miller, 2019). This reveals a disconnect between experts' views and end-users' perspectives.

For end-users, not only are the requirements for XAI expected to be unique (Cheng et al., 2019; Langer et al., 2021), but so is the intended purpose behind XAI (Suresh et al., 2021). While the objectives of XAI are broad, ranging from supporting developers in debugging and debiasing AI models to enabling regulators to ensure that AI is compliant with laws or standards, the primary goal of XAI for end-users is to help building trust in AI (Jacovi et al., 2021; Suresh et al., 2021) which is also underlined by the vast number of studies on the topic. For this reason, this dissertation centers

¹ This dissertation deliberately avoids using the term *black-box* in contrast to *white-box* to describe AI that humans cannot easily decipher as suggested by the ACM's guidelines on diversity, equity, and inclusion. This choice avoids racially charged language and acknowledges that color-based descriptors carry culturally specific connotations that may limit their universal applicability.

around trust in AI as a central construct for end-users² in the human–AI interaction (Ueno et al., 2022). This emphasis on trust is not merely a response to the expanding capabilities of AI but also a recognition of the profound impact this technology can have on human lives (Jobin et al., 2019; Kaur et al., 2022). Trust plays a key role not only in people’s willingness to use and rely on automated systems (Hoff & Bashir, 2015) but also in the public acceptance of new technologies at the societal level (Knowles & Richards, 2021; Vorm & Combs, 2022).

Trust in AI, as explored in this dissertation, is not a monolithic construct but a multifaceted one that necessitates a thorough understanding of what trust entails and of what delineates it from other constructs. The endeavor to define trust in AI is marked by complexities and involves untangling a net of conceptual nuances, as trust is often conflated with related but distinct constructs such as reliance (Scharowski et al., 2022). This conflation not only complicates the establishment of a clear and agreed-upon definition of trust; rather, the choice of a particular definition also inevitably guides the theoretical and methodological approach to investigating the construct of interest (Aeschbach et al., 2021). Only after such conceptual challenges have been addressed and resolved can approaches and methods to build trust be meaningfully investigated in empirical human–AI research.

However, even if researchers have appropriately defined trust in AI and established a corresponding operationalization of it in their studies (i.e., the translation of a construct of interest that is not measurable or observable into something related to it that is measurable or observable; Slife et al., 2016), they still face challenges when measuring trust. There are a multitude of trust questionnaires (Kohn et al., 2021), but none of these have been specifically validated for the context of AI (Lai et al., 2023). This makes it difficult for researchers to make an informed decision in selecting appropriate trust questionnaires, which hinders the comparability of results across empirical studies and raises doubts about the validity of their findings.

² In line with prior work (Arrieta et al., 2020; Langer et al., 2021; Scharowski, Benk, et al., 2023; Seifert et al., 2019), end-users in this dissertation are defined as laypeople (i.e., non-experts in data science or machine learning) who may be affected directly or indirectly by the outcomes of AI systems; they are sometimes also referred to as data subjects (Knowles & Richards, 2021; Suresh et al., 2021).

Additionally, empirical research should critically examine under what circumstances increasing trust in AI is desirable in the first place. A uniform increase in trust is inappropriate; instead trust should be reflective of the actual capabilities and limitations of AI. For trust to be warranted, any increase in it should be contingent and calibrated based on the inherent trustworthiness of the AI (Jacovi et al., 2021). For this reason, we will argue that in circumstances where AI is untrustworthy, it might even be advisable to maintain a certain level of warranted distrust toward AI.

This dissertation focuses on two central questions: (I) how to understand and measure trust in AI, and (II) how to calibrate and increase trust and distrust to a level that is warranted. The manuscripts and contributions that formed this dissertation addressed these questions either explicitly or implicitly and contributed to resolving some of the challenges and issues raised above. Manuscript 1 (i.e., Scharowski, Perrig, Svab, et al., 2023) investigated the impact of human-centered post-hoc explanations on trust and trust-related behavioral measures, inspiring several follow-up projects that attempted to gain a more nuanced understanding of trust in AI. One of these follow-up projects resulted in manuscript 2 (i.e., Scharowski, Perrig, Aeschbach, et al., 2023), which investigated the psychometric quality of trust questionnaires and reflected on the implications arising from an identified two-factor structure of trust *and* distrust within the context of AI. Manuscript 3 of this dissertation (i.e., Scharowski, Benk, et al., 2023) explored approaches to increase warranted trust beyond XAI methods by considering how certification labels could communicate to end-users that an AI has been audited and deemed trustworthy by qualified auditors.

Collectively, these manuscripts set out to explore the complex terrain of trust in AI, with a particular focus on end-users. This dissertation aims to deepen the understanding of trust in the context of AI by critically examining the role of XAI as a means for increasing calibrated trust and distrust. It seeks to contribute to the broader discourse on human-centered AI, which advocates for systems that are not only advanced in their capabilities but also aligned with human values and needs, ensuring that AI is worthy of the trust placed in it by the people whom it should serve.

Theoretical Background

The following sections examine the questions of (I) how to understand and measure trust in AI and (II) how to calibrate and increase trust and distrust to a level that is warranted. The challenges associated with answering these questions will be identified and discussed.

Understanding Trust in AI

Understanding trust in AI requires exploring its definitions, models, and measurements and appreciating the challenges involved. Although these challenges are discussed in distinct sections, it is important to recognize that they are interconnected and cannot be understood in isolation. Issues encountered in one stage of empirical research invariably influence the next stage, or as Durlak and DuPre (2008) put it: "science cannot study what it cannot measure accurately and cannot measure what it does not define" (p. 342). For example, the way trust is conceptualized (definition) informs the theoretical framework it is embedded in (modeling), which in turn affects how trust is quantified in empirical settings (measurement). Although I acknowledge these dependencies, each of these aspects will be presented and discussed in an individual section for clarity and ease of comprehension.

Defining Trust in AI

The concept of trust has been dissected and discussed across multiple disciplines for decades, with contributions from fields such as philosophy (Baier, 1986), sociology (Gambetta, 2000), and economics (Berg et al., 1995). This cross-disciplinary focus has led to diverse interpretations and conceptualizations of trust, ranging from the expectation of benevolence in interpersonal contexts (Mayer et al., 1995) to strategic reciprocity in economic models (Berg et al., 1995) and moral considerations in philosophical approaches to trust (Baier, 1986). However, the importance of trust is not limited to interactions between people; it has also been highlighted as a critical factor in human interactions with computers (Riegelsberger et al., 2005), computer-mediated

services (Beldad et al., 2010), automation (Lee & See, 2004), and robots (Hancock et al., 2011). Because trust in AI is considered an elaborate form of automation, current research on human–AI interaction has adopted and extended definitions from trust in automation to trust in AI (Jacovi et al., 2021).

For this reason, the most commonly used definition of trust in the literature on human–AI interaction (Sassmannshausen et al., 2023; Ueno et al., 2022; Vereschak et al., 2021) is attributable to Lee and See (2004)'s definition of trust in automation as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (p. 54). This definition is based on the influential work by Mayer et al. (1995), who emphasize the critical role of uncertainty, vulnerability, and risk in the formation of trust. Uncertainty, vulnerability, and risk have been widely regarded as necessary conditions for the existence of trust (Buçinca et al., 2020; Castelfranchi & Falcone, 2010; Hoff & Bashir, 2015; Rousseau et al., 1998; Vereschak et al., 2021). The definition of trust in automation by Lee and See (2004) was adopted for trust in AI throughout this dissertation for three reasons: (I) its widespread use as a definition for trust in AI; (II) its emphasis on risk and vulnerability for trust to be a meaningful concept; and (III) its broad applicability, as the definition does not specifically require the trusted party to be a human, a robot, or AI.

However, defining trust within the literature on human–AI interaction faces two challenges that both complicate integrating empirical findings into a cohesive theoretical framework. First, only a minority of studies explicitly provide a definition of trust (Bach et al., 2024; Sassmannshausen et al., 2023; Vereschak et al., 2021). Second, the literature is marked by a multitude of definitions and conceptualizations of trust (Benk et al., 2022; Ueno et al., 2022; Vereschak et al., 2021), which sometimes converge with related but distinct constructs such as *perceived trustworthiness* (Schlicker et al., 2022), *calibrated trust* (Wischnewski et al., 2023), *reliance* (Scharowski et al., 2022), and *warranted trust* (Jacovi et al., 2021). If researchers do not clearly delineate between these constructs, theoretical conflation can occur in which different facets of trust and related constructs are lumped together under one single term, neglecting the nuances

that distinguish these constructs from one another (Kohn et al., 2021). Such conflation hinders the development of theory, as well-defined constructs form the building blocks of theory and models (Shoemaker et al., 2004).

The theoretical conflation between trust and related constructs can result in operationalizations that intend to measure trust but instead measure trust-related constructs (e.g., reliance), leading to misunderstandings regarding what is actually investigated in a study (Scharowski et al., 2022). For example, Lai and Tan (2019) defined trust as "the percentage of instances for which humans follow the machine prediction" (p. 5) and thus operationalized it as a behavioral measure. This approach contrasts with the commonly held perspective that considers trust, either explicitly or implicitly, as an attitude (Castelfranchi & Falcone, 2010; Vereschak et al., 2021), that is, as a subjective psychological construct. Such constructs are unobservable features (e.g., psychological traits or abilities; Hopkins, 1998) that are typically measured via questionnaires (Scharowski et al., 2022). In contrast to trust as an attitude, in this dissertation reliance is defined as a behavior "that follows from the advice of the system" (Scharowski et al., 2022, p. 3), which means that is open to different approaches of measurement than trust.

Modeling Trust in AI

A definition necessitates the explicit or implicit adoption of the theoretical model in which it is embedded. This adoption further determines how the construct of interest is operationalized and measured (Aeschbach et al., 2021). Beyond the aforementioned multitude of definitions, there is also a wide variety of models of trust in circulation (e.g., Davis, 1989; Hoff & Bashir, 2015; Lee & See, 2004; Liao & Sundar, 2022; Madsen & Gregor, 2000; Mayer et al., 1995; McKnight & Chervany, 2001a; Toreini et al., 2020). However, only around 23% of empirical studies on human–AI interaction explicitly refer to a model of trust, and of those, the majority use different rather than the same model (Ueno et al., 2022). This echoes concerns about a "theory crisis" in psychology in which research findings are not sufficiently embedded in a robust theoretical framework and an overabundance of theories predominate. This can lead to a fragmentation of

disconnected findings, making it challenging to build cumulative knowledge in the field (Eronen & Bringmann, 2021; Oberauer & Lewandowsky, 2019).

While an in-depth exploration of the overlaps and dissimilarities between the various models of trust is beyond the scope of this dissertation, most of the models involve a task to be executed under some risk or with uncertainty of the outcome, necessitating a trustor (i.e., the entity that trusts) to depend on a trustee (i.e., the entity that is trusted in). To ensure a coherent discussion and given the adoption of Lee and See (2004)'s definition of trust in AI throughout this dissertation, I will mainly focus on the model of trust and its components proposed by Mayer et al. (1995), whom Lee and See (2004) used as a foundation for their model of trust in automation. Within Mayer et al. (1995)'s model, trust is built on the trustee's attributes, which are called *factors of trustworthiness*. These factors include ability (i.e., the set of skills, competencies, and attributes relevant within a specific domain), benevolence (i.e., the degree to which the trustee is believed to act in the trustor's interest beyond just self-centered motivations), and integrity (i.e., the trustor's belief that the trustee adheres to a set of principles deemed acceptable by the trustor). Building on Mayer et al. (1995), Lee and See (2004) expanded these factors of trustworthiness to automation, introducing performance (i.e., *what* the automation does), process (i.e., *how* the automation works), and purpose (i.e., *why* the automation was developed) as the foundation of trust. For example, based on the repeated demonstration of a self-driving car in effortlessly navigating through a city, a trustor might perceive it as having high ability. This assessment contributes to the trustor's perception of the self-driving car as trustworthy.

More recent studies have introduced AI-specific trustworthiness factors (e.g., Kaplan et al., 2023; Liao & Sundar, 2022; Thornton et al., 2021; Toreini et al., 2020). For example, drawing on Mayer et al. (1995) and Lee and See (2004), Liao and Sundar (2022) defined ability (i.e., an AI's capabilities, such as making predictions and generating answers), intention benevolence (i.e., an AI's underlying purpose and the ethical considerations guiding its development), and process integrity (i.e., an AI's appropriateness and the reliability of its decision-making processes) as key attributes of

AI trustworthiness. However, due to the similarity of these terms to Lee and See (2004)'s basis of trust, they are summarized in Figure 1 under the general concept of *actual trustworthiness*. In this regard, it is important to differentiate between the trustor's *perceived trustworthiness* and the trustee's actual trustworthiness (Schlicker et al., 2022). The actual trustworthiness corresponds to the inherent attributes of the trustee, while the perceived trustworthiness is merely the trustor's assessment of these attributes, which to some extent remain unknown (Mayer et al., 1995; Schlicker et al., 2022). For this reason, the trustor can never have complete knowledge of the actual trustworthiness of the trustee and their perceived trustworthiness will always involve a degree of uncertainty and risk. At this point, vulnerability comes into play as key factor in the existence of trust, echoing Simmel (1908)'s assertion that "the omniscient needs no trust, the completely ignorant cannot trust" (p. 263). Consequently, the trustor remains vulnerable to the actions and the attributes of the AI system that were not perceived or for which there is uncertainty (Schlicker et al., 2022). In this sense, trust can be regarded as the willingness to accept vulnerability and the corresponding uncertainty (i.e., the potential mismatch between perceived and actual trustworthiness) (Mayer et al., 1995; Schlicker et al., 2022).

In addition to recognizing uncertainty, risk, and vulnerability as necessary conditions for trust, most models acknowledge that trust is both dynamic (i.e., subject to change over time; Hoff and Bashir, 2015) and multifaceted (Lee & See, 2004). Different factors residing in human predisposition (i.e., human-related factors), the AI (i.e., AI-related factors), and the human-AI interaction within the environment (i.e., environmental factors) contribute to the formation of trust and the resulting trust-related behavior (Hoff & Bashir, 2015). However, this is where the general consensus between these models seems to end, and complexity and ambiguity begin to emerge. Indeed, a systematic literature review by Sassmannshausen et al. (2023) identified as many as 479 factors that are believed to influence trust formation in the context of AI. These findings indicate that researchers keep uncovering and introducing new factors of trust (Sassmannshausen et al., 2023), such as *interaction frequency*, *level of automation*,

demographics, familiarity with AI, and many others, leading to a certain inflation of models of trust in AI. Due to the variety of these factors, they are only indicated and not explicitly described in Figure 1.

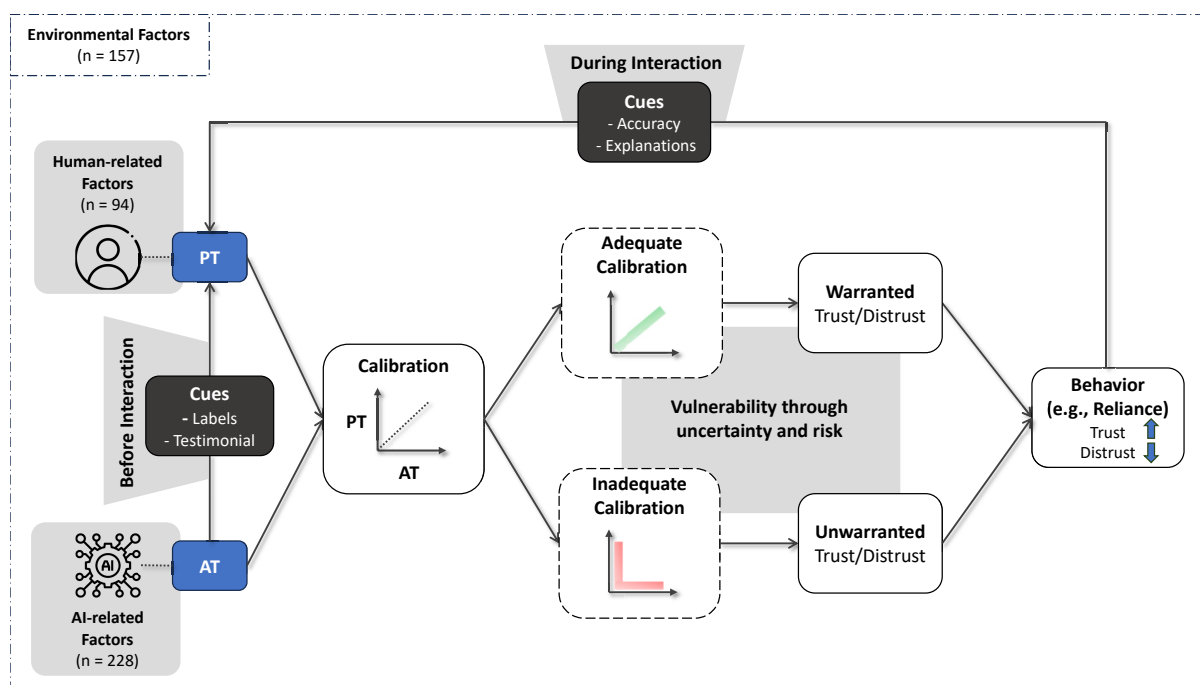


Figure 1

Model adapted from Mayer et al. (1995) and Lee and See (2004) illustrating the dynamics of the formation of trust and distrust based on the calibration of actual trustworthiness (AT) and perceived trustworthiness (PT). The model emphasizes the adequate calibration through trustworthiness cues before and during the interaction to form warranted or unwarranted trust and distrust that influences corresponding behavior (e.g., reliance). The factors influencing trust ($N = 479$) include AI-related, human-related, and environmental considerations (Sassmannshausen et al., 2023).

Figure 1 illustrates a concise adaptation of the trust model by Mayer et al. (1995) with Lee and See (2004)'s considerations regarding trust calibration for the context of AI. In this model, trust as a dynamic process based on the interaction with AI can be increased, decreased, maintained, broken, or repaired after it has been broken (Glikson & Woolley, 2020). Trust is thus process-oriented and recursive, and it contains a temporal dimension in which certain components of trust appear earlier in the model than others. For example, most models treat trust as an antecedent guiding reliance. However, the two share a probabilistic relationship and not a deterministic one (Körber,

2019), as attitudes do not always translate into behavior (Ajzen & Fishbein, 1980). This implies that even if a system is trusted, reliance must not necessarily follow from it and vice versa (Kirlik, 1993; Körber, 2019). For this reason, trust and reliance remain conceptually distinct in this model. Consequently, research should not treat the corresponding attitudinal (e.g., trust) and behavioral measures (e.g., reliance) interchangeably, which echoes the appeal of the last section.

Considering the potential mismatch between how trustworthy a system actually is and how it is perceived by the trustor, there is a risk that the resulting trust might be unjustified. Recognizing this issue, Lee and See (2004) coined the concept of *trust calibration*, which refers to aligning an individual's perceived trustworthiness with the system's actual trustworthiness (Lee & See, 2004). Two kinds of mismatches in calibration can occur: the individual's resulting trust may exceed the system's trustworthiness, leading to misuse (i.e., over-reliance; Parasuraman and Riley, 1997), or an individual's trust may fall short of the system's trustworthiness, resulting in disuse (i.e., under-reliance; Parasuraman and Riley, 1997) of the system. These two mismatches are referred to in Figure 1 as *inadequate calibration*. For example, a trustor might place too much trust in a self-driving car and show over-reliance on it in navigating through severe weather conditions beyond its ability, risking their own safety. Conversely, a trustor may place too little trust in a self-driving car and show under-reliance on it to manage routine navigation under perfect conditions, forfeiting the chance to enhance safety. Ideally, individuals should display calibrated trust, where the level of trust accurately matches the trustworthiness of the system. Building on this idea, a related concept for reliance calibration was introduced by Schemmer et al. (2023), where the appropriateness of reliance can be evaluated for classification problems. According to this concept, reliance is deemed appropriate when the trustor relies on themselves in cases where they are correct and the AI is incorrect, and relies on the AI when they are incorrect and the AI is correct (Schemmer et al., 2023).

Acknowledging that trust and reliance should match the actual trustworthiness of a system, Wischniewski et al. (2023) urged the XAI community not to aim for increasing

users' trust uniformly but to focus explicitly on increasing calibrated trust. In this regard, Jacovi et al. (2021) introduced the notion of *warranted* and *unwarranted* trust for AI (see Figure 1). They defined warranted trust as trust that is calibrated with trustworthiness; otherwise, trust is unwarranted if it is not calibrated with trustworthiness. This notion of warranted and unwarranted trust brings about an interesting distinction; if an AI system is untrustworthy (e.g., has low ability), then not only is a person's trust unwarranted but also their distrust is warranted (Jacovi et al., 2021). This means that if a system is untrustworthy, it may not be sufficient for people simply not to trust the system; rather, it may be advisable for them to actively distrust it. Jacovi et al. (2021) argued that while the key motivation of XAI is commonly framed as simply increasing trust in AI systems, a more precise motivation should be either (I) to increase calibrated trust in trustworthy AI or (II) to increase calibrated distrust in untrustworthy AI. For this reason, we extended Figure 1 to include the construct of distrust.

This distinction emphasizes the theoretical importance of distrust and its consideration in XAI research. However, the research community has predominantly focused on trust (Peters & Visser, 2023; Scharowski & Perrig, 2023), and while this focus has yielded important insights into how trust in AI can be developed and maintained, distrust as a separate construct has remained relatively understudied (Ueno et al., 2022). This unilateral perspective on trust ignores decades of research (e.g., Lewicki et al., 2006; Luhmann, 1979; McKnight & Chervany, 2001a; Ou & Sia, 2009; Saunders et al., 2014; Sitkin & Roth, 1993) that has extensively explored both the coexistence and independence of trust *and* distrust. There are theoretical reasons to treat trust and distrust as independent constructs — rather than as mere opposites — with distinct antecedents and consequences (Cacioppo & Berntson, 1994; Chang & Fang, 2013; Lewicki et al., 1998; McKnight & Chervany, 2001b).

For example, trust and distrust can both help to manage uncertainty and complexity, which were shown in the previous section to be inherent in trust-based relations, but in different ways. While trust reduces complexity by leading individuals to engage in

actions that expose them to vulnerability (i.e., undesirable outcomes are removed from consideration to form positive expectations; Kroeger, 2019), distrust reduces complexity by leading individuals not to engage in actions that expose them to vulnerability (i.e., undesirable outcomes are accentuated in their consideration to form negative expectations; Kroeger, 2019). Given these theoretical reflections, distinguishing between trust and distrust seems justified and could inform future XAI research. This distinction would allow not only to consider warranted trust in trustworthy AI but also warranted distrust in untrustworthy AI, aligning more closely with the refined motivations of XAI outlined by Jacovi et al. (2021). However, this distinction can only be realized if the two constructs can be measured separately and appropriately.

Measuring Trust in AI

After introducing models how trust is formed and recognizing that the perceived trustworthiness should ideally be calibrated with the actual trustworthiness of the AI, the question arises of how the resulting level of trust (and distrust) can be measured. Although trust evolves over time, most empirical studies do not account for this dynamic nature. Trust is mainly measured either before or after an interaction, with only a minority of studies measuring it multiple times during an interaction with AI (Ueno et al., 2022; Vereschak et al., 2021). In human–AI interaction, trust is assessed in various ways ranging from attitudinal to behavioral measurements (Kohn et al., 2021; Mohseni et al., 2020; Vereschak et al., 2021).

As mentioned earlier, behavior as a more directly observable phenomenon that can be measured through, for example, *agreement rate*, *decision time*, *trust games*, *compliance*, and *reliance* as well as through *physiological measures* (Mohseni et al., 2020; Vereschak et al., 2021). Although these are often referred to as "measures of trust," it was argued earlier, that trust defined as an attitude, should be assessed using attitudinal measures. While behavioral measures are also important in human–AI research, studies that rely solely on behavioral measures, such as reliance, do not genuinely measure trust but trust-related behavior (e.g., Ahn et al., 2024). This conflation can lead to misleading interpretations of research findings, as the results of attitudinal and behavioral trust

may appear inconsistent or contradictory (Scharowski et al., 2022).

Subjective measures of trust include *interviews*, *open-ended question*, and *think-aloud protocols* (Mohseni et al., 2020; Vereschak et al., 2021). That being said, *questionnaires* are the primary source of subjective measurement (Vereschak et al., 2021); Ueno et al. (2022) estimated that 89% of publications measure trust via questionnaires. However, similar to the plethora of definitions and models, there are also a variety of questionnaires for measuring trust in AI (e.g., Hoffman et al., 2023; Jian et al., 2000; Körber, 2019; Madsen & Gregor, 2000; Mayer & Davis, 1999; Merritt, 2011; Schaefer, 2016). These originated from different disciplines such as human–human trust, human–agent trust, human–automation trust, and human–robot trust.

Even with this variety, new measures and scales for trust continue to be developed and introduced to the field (e.g., Hoffman et al., 2023). This makes it difficult for researchers to arrive at an informed decision about which scale to use and also raises the question of whether the models underlying these trust questionnaires are sufficiently similar for comparisons of empirical results. For example, some questionnaires measure constructs associated with trustworthiness, such as a system’s capability and benevolence (e.g., Cai et al., 2019; Mayer & Davis, 1999), rather than trust itself. In addition, only 6% of papers on human–AI interaction measure distrust (Ueno et al., 2022), a fact that underlines the aforementioned disregard for this construct. Indeed, many trust questionnaires seem to lack a clear connection to any underlying model, and conversely, many trust models do not sufficiently specify how different measures of trust are incorporated in their theoretical models (Kohn et al., 2021). This disconnect between models and measures further complicates systematic scientific investigations on trust (e.g., meta-analyses), as it remains unclear which specific aspects of trust were measured in a study and how they fit within a theoretical model (Kohn et al., 2021).

Despite (or maybe because of) this variety of questionnaires available, researchers often develop their own scales (e.g., Merritt, 2011; Yin et al., 2019) or use single-item questions (e.g., Yu et al., 2017) to assess trust. While using self-developed scales and single-item measures can offer certain benefits, such as being better tailored to the

study context or causing less task disruption for participants, these measures usually lack a rigorous design and validation process (Furr, 2011) and are often only used in a single study, rendering it difficult to compare their results with other studies (Flake & Fried, 2020). For this reason, Wischnewski et al. (2023) recommended using validated and standardized trust questionnaires that have undergone scrutiny to ensure their psychometric quality, including objectivity, reliability, and validity. However, this recommendation poses challenges for researchers seeking to measure trust in AI, as there exist no validated trust questionnaires in the context of AI.

The absence of validated questionnaires for trust in AI compels researchers to adopt questionnaires from other domains, and they often necessitate modifying the scale's items to fit the new study context. Vereschak et al. (2021) estimated that more than half of all empirical studies in research on human–AI interaction introduce such modifications to the original, validated questionnaires (e.g., changing "the system is dependable" to "the artificial intelligence is dependable"). However, terminological differences affect people's perception and assessment of technology (Langer et al., 2022), and any modification of a questionnaire can alter its reliability and validity, prompting questions about whether the modified scale still measures the intended construct. Consequently, any modification to a questionnaire demands a reevaluation of its psychometric quality (Furr, 2011; Juniper, 2009), a practice often overlooked in current research on human–AI interaction (Vereschak et al., 2021). At best, the use of non-validated trust questionnaires in the context of AI makes it challenging for other researchers to replicate or build on existing work. At worst, this practice leads to ambiguous or contradictory results that hinder progress in XAI and human–AI interaction altogether. In response to these challenges, manuscript 2 of this dissertation (i.e., Scharowski, Perrig, Aeschbach, et al., 2023) conducted an extensive validation of trust questionnaires that are used to measure trust in AI.

These last sections examined question (I), how to understand and measure trust in AI. For this, the challenges involved in defining, modeling, and measuring trust and how these are interconnected were outlined. Our research efforts in pursuing rigorous definitions (i.e., Scharowski et al., 2022), valid and reliable measures (i.e., Perrig, Scharowski, & Brühlmann, 2023), and a more nuanced understanding of the dimensionality (i.e., Scharowski & Perrig, 2023) of trust have attempted to address some of these challenges. The next section will approach question (II), how to calibrate and increase trust and distrust to a level that is warranted. For this, XAI methods and other approaches to calibrate trust and distrust will be introduced.

Calibrating and Increasing Warranted Trust and Distrust in AI

Because trust in AI is fundamentally grounded in the perception of trustworthiness, calibrating trust and distrust begins with effectively communicating an AI system's attributes. In this context, researchers refer to the concept of *trustworthiness cues*. (de Visser et al., 2014; Liao & Sundar, 2022; Schlicker et al., 2022). Trustworthiness cues are any information about an AI's attributes (e.g., ability, benevolence, integrity) that can contribute to a user's trust assessment (Liao & Sundar, 2022). These cues essentially act as evidence of the AI's trustworthiness. For example, if an AI explains its output or decision (e.g., through post-hoc explanations), these explanations might act as a cue for the AI's ability, whereas compliance with regulations and ethical standards (e.g., through an AI certification label) could signify the AI's integrity.

End-users then use these cues as heuristics (i.e., mental rules of thumb) to make judgments about the perceived trustworthiness of the AI (Schlicker et al., 2022).

However, as with any heuristic, these judgments can be flawed, which is why Liao et al. (2020) stress the importance of ensuring that trustworthiness cues are both truthful and relevant. Ideally, end-users hold *justified true beliefs* about the trustworthiness of an AI system that are attributable to its actual trustworthiness and not accidental or random (Ferrario & Loi, 2022), pointing again to the calibration of the perceived trustworthiness of the end-user with the AI's actual trustworthiness.

Take the following example in relation to Figure 1: before interacting with a self-driving car, an end-user can only infer its actual trustworthiness from cues such as testimonials from friends or certification labels from an audit. These cues influence the perceived trustworthiness of the end-user, but they must be truthful and recognizable for them to be correctly incorporated into their calibration. Because the end-user can never have complete knowledge of the self-driving car's actual trustworthiness, there is risk and uncertainty of a potential mismatch between the perceived and actual trustworthiness. Only if the end-user willingly accepts the resulting vulnerability does trust form as a meaningful construct. The resulting levels of trust and distrust can then either increase (in the case of trust) or decrease (in the case of distrust) the likelihood that the end-user will display behavioral manifestations of trust, such as relying on the self-driving car. The experiences during the interaction, such as the car's accuracy in identifying pedestrians or its ability to provide explanations for unexpected driving behavior, also serve the end-user as trustworthiness cues. This in turn leads to a feedback effect on the perceived trustworthiness and the subsequent levels of trust and distrust.

Two of the manuscripts comprising this dissertation explored potential ways to increase end-users' warranted trust in AI. Specifically, we investigated the impact of human-centered post-hoc explanations (i.e., Scharowski, Perrig, Svab, et al., 2023) and certification labels for trustworthy AI (i.e., Scharowski, Benk, et al., 2023). Numerous cues can be used to assess the trustworthiness of an AI system, but in recent years, XAI research in particular has explored a variety of methods for providing some level of transparency in several hundred publications annually (Speith, 2022). Researchers have begun to categorize different XAI methods; interested readers seeking further information can refer to the work of Arrieta et al. (2020), Molnar (2019), and Speith (2022), who have provided comprehensive overviews and taxonomies of these methods. However, despite the popularity of XAI as a area of research, empirical evidence on whether XAI methods effectively increase warranted trust has remained inconclusive (e.g. Cheng et al., 2019; Ehsan et al., 2019; Kizilcec, 2016; Nothdurft et al., 2013; Papenmeier et al., 2022; Poursabzi-Sangdeh et al., 2021; Zhang et al., 2020).

I argue that potential reasons for these ambiguous findings, as outlined in the previous sections, may include:

- I inconsistent definitions of trust in AI and its conflation with related but distinct constructs;
- II inadequate integration of relevant constructs into coherent models to enable proper investigation of their relationships;
- III inappropriate use or adoption of measures that are not validated in the context of trust in AI.

These ambiguous findings (to which, as we shall see, manuscript 1 also contributed) have sparked criticism of XAI and prompted explorations of alternative approaches to increasing warranted trust and distrust in AI. For example, Knowles and Richards (2021) argued that public-facing AI explanations are ineffective because end-users typically lack the expertise to assess the trustworthiness of AI on the basis of such XAI methods. Instead, they advocated for a regulatory ecosystem that can guarantee AI trustworthiness through institutional authority and the power to sanction untrustworthy AI. Within this framework, XAI methods would only be utilized by qualified auditors who could use AI explanations and other forms of AI documentation to verify that an AI could be trusted (Knowles & Richards, 2021).

Such considerations inspired manuscript 3 on certification labels (i.e., Scharowski, Benk, et al., 2023) as a trustworthiness cue for end-users. Certification labels signal that an AI has been deemed trustworthy by an audit. We think that communicating the outcomes of such audits also represents a form of transparency. If mandatory, certification labels have the potential both to increase warranted trust, when a label signals trustworthiness, and to increase warranted distrust, if the absence of a label signals untrustworthiness.

To this point, this dissertation has attempted to (I), provide a better understanding of trust in AI in terms of definitions, models, and measures of trust while also outlining certain challenges and (II), introducing approaches and methods to calibrate and increase trust and distrust to a level that is warranted. Next, a comprehensive overview of each of the individual manuscripts that have shaped this dissertation and contributed to addressing some of the challenges raised will be provided. The remainder of the dissertation is structured as follows: First, a summary of each of the three main manuscripts will be presented; the summaries will detail their motivations, objectives, methodologies, results, and discussions. This will be followed by a collective examination and final synthesis of the key findings and implications drawn from both the manuscripts and some additional related contributions.

Summary of the Manuscripts

The following manuscripts constitute this thesis. Manuscript 1 and 3 have already been published. Manuscript 2 is under review.

1. **Scharowski, N.**, Perrig, S. A. C., Svab, M., Opwis, K., & Brühlmann, F. (2023). Exploring the effects of human-centered AI explanations on trust and reliance. *Frontiers in Computer Science*, 5. <https://doi.org/10.3389/fcomp.2023.1151150>
2. **Scharowski, N.**, Perrig, S. A. C., Aeschbach, L. F., von Felten, N., Opwis, K., Wintersberger, P., & Brühlmann, F. (2023). To trust or distrust trust measures: Validating questionnaires for trust in AI. *Manuscript submitted for publication*.
3. **Scharowski, N.**, Benk, M., Kühne, S. J., Wettstein, L., & Brühlmann, F. (2023). 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). New York, NY, USA: ACM. <https://doi.org/10.1145/3593013.3593994>

The following publications and contributions are related to this thesis and referenced in it but were omitted for the sake of brevity and focus.

- **Scharowski, N.**, Opwis, K., & Brühlmann, F. (2021). Initial evidence for biased decision-making despite human-centered AI explanations. *CHI 2021 Workshop: Operationalizing Human-Centered Perspectives in Explainable AI*.
<https://osf.io/preprints/osf/5jzmb>
- **Scharowski, N.**, Perrig, S. A. C., von Felten, N., & Brühlmann, F. (2022). Trust and reliance in XAI – Distinguishing between attitudinal and behavioral measures. *CHI 2022 TRAIT Workshop on Trust and Reliance in AI-Human Teams*. <https://doi.org/10.48550/arXiv.2203.12318>
- **Scharowski, N.**, & Perrig, S. A. C. (2023). Distrust in (X)AI – Measurement artifact or distinct construct? *CHI 2023 TRAIT Workshop on Trust and Reliance in AI-Human Teams*. <https://doi.org/10.48550/arXiv.2303.16495>
- Benk, M., Wettstein, L., Schlicker, N., Wangenheim, F., & **Scharowski, N.** (2024). Bridging the Knowledge Gap: Understanding User Expectations for Trustworthy LLM Standards. *Manuscript submitted for publication*.
- Perrig, S. A. C., **Scharowski, N.**, & Brühlmann, F. (2023). Trust issues with trust scales: Examining the psychometric quality of trust measures in the context of AI. *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. <https://doi.org/10.1145/3544549.3585808>
- Perrig, S. A. C., **Scharowski, N.**, Brühlmann, F., von Felten, N., Opwis, K., & Aeschbach, L. F. (2024). Independent validation of the Player Experience Inventory: Findings from a large set of video game players. *Proceedings of the CHI Conference on Human Factors in Computing Systems*.
<https://doi.org/10.1145/3613904.3642270>

- Spiess, F., **Scharowski, N.**, Haller, A., Memeti, Z., Schuldt, H., & Brühlmann, F. (2024). Bringing Video Browsing to Virtual Reality: Empirical Evaluation of a Novel Multimedia Drawer. *Proceedings of the 2024 International Conference on Multimedia Retrieval*. <https://doi.org/10.1145/3652583.3658077>
- Perrig, S. A. C., Aeschbach, L. F., **Scharowski, N.**, von Felten, N., Opwis, K., & Brühlmann, F. (2022). Measurement practices in UX research: A systematic quantitative literature review. *Frontiers in Computer Science*, 6. <https://doi.org/10.3389/fcomp.2024.1368860>

Manuscript 1: "Exploring the Effects of Human-Centered AI Explanations on Trust and Reliance"

Motivation and Aim of the Study. Within the discourse on transparent AI, it is crucial to differentiate between the terms *interpretability* and *explainability*. While both terms refer to transparency methods and are often used interchangeably (Arrieta et al., 2020), they differ in their approach to achieving transparency. Interpretability renders AI transparent by a directly observable or visually interpretable decision-making processes and is often linked to the concept of clear-box AI. Conversely, explainability accepts the opaque-box paradigm for AI systems and focuses on providing explanations of how a particular output or decision was reached, essentially serving as a form of post-hoc interpretability (Ehsan et al., 2019; Lipton, 2018; Mohseni et al., 2020).

In this initial study, we explored post-hoc explanations to enhance end-users' trust in AI (Stephanidis et al., 2019). In this context, *post-hoc* refers to the fact that the explanations for a specific recommendation (or prediction, outcome, etc.) are only provided after a computation has been carried out by the AI (Lipton, 2018). This approach avoids rendering the AI system as a whole transparent, as it is often unfeasible to disclose the inner workings of complex AI systems (e.g., deep neural networks). Instead, post-hoc explanations are provided that are similar to how humans rationalize their decisions. The human brain has also been characterized as opaque, as people usually do not have direct access to its fundamental decision-making processes (e.g., neuronal activation). Nevertheless, humans can explain their decisions after the fact in a way that other individuals can understand.

For this reason, post-hoc explanations seemed to be a particularly promising XAI method for rendering AI systems more transparent and trustworthy to end-users. We reviewed the XAI literature to identify additional factors that contribute to compelling explanations from a human perspective; we thus aimed to identify factors related to human-centered explainable AI (Ehsan, Wintersberger, Liao, et al., 2021; Ehsan et al., 2022). By doing so, we not only narrowed down the number of potential post-hoc explanations to be investigated but also hoped to optimize the effect that these

explanations potentially have on end-users. This seemed particularly important as empirical studies have shown mixed results regarding the impact of AI explanations on warranted trust in AI systems (Kästner et al., 2021; Langer et al., 2021).

Drawing on past work (Adadi & Berrada, 2018; Ehsan et al., 2019; Mittelstadt et al., 2019) and specific criteria recognized to contribute to meaningful explanations for people as defined by Miller (2019), we focused on explanations that are *selective*, *contrastive*, and akin to human *social* interactions. Considering these criteria, we identified two promising human-centered post-hoc explanations: feature-importance and counterfactual explanations. Given the lack of empirical research on human-centered explanations, an empirical investigation into the efficacy of feature-importance and counterfactual explanations seemed warranted and promising for fostering end-users' trust in AI.

Method. Employing a mixed design, this study compared feature-importance and counterfactual explanations with a control condition in the context of real-estate valuation. The experimental design consisted of a 3×2 structure, with the explanation conditions as the between-subject factor and the type of AI recommendation (increasing or decreasing the price) as the within-subject factor.

Given that the added explanatory information of counterfactual explanations are both selective and contrastive, we hypothesized they would have a greater impact on end-users' trust than feature-importance explanations. We distinguished trust as a psychological construct from reliance as a trust-related behavior, formulating hypotheses for each. The independent variable was the explanation condition, and the dependent variables included *weight of advice* (WOA) from the literature on advice taking (Harvey & Fischer, 1997) to measure reliance and the Trust between People and Automation Scale (TPA) by Jian et al. (2000) for trust.

The experiment was conducted online via Amazon Mechanical Turk ($N = 380$). The sample predominantly comprised male participants (61%), and the average age was 37 years ($M = 37.03$, $SD = 10.15$, $min = 18$, $max = 69$). The task involved estimating subleasing prices for different apartments. The task was designed to mimic actual

decision-making scenarios with apartment listings from an existing real-estate marketplace. Based on an apartment's features and amenities (e.g., number of bedrooms, distance to public transit), participants had to estimate an initial subleasing price ($T1$). After estimating $T1$, an algorithm (introduced to participants as an AI system) provided a price recommendation (R). After being presented with the price recommendation, participants could decide if they desired to approach the price recommendation or not and settled on a final subleasing price ($T2$). $T1$, R , and $T2$ were crucial components of WOA, which measured the degree to which people changed their behavior and moved their initial estimate toward the advice. To account for the within-subject factor (type of AI recommendation), half of the apartments were randomly selected for the AI to recommend an increase to the initial price, and for the other half it recommended a decrease to the initial price. This allowed for an independent assessment of the two types of AI recommendations on participants' reliance. After completing the task, participants filled out the TPA and provided demographic information.

Results. Regarding reliance, participants in both conditions approached the AI recommendations, resulting in a positive WOA ($M = 0.69$, $SD = 0.36$). Their adjustment more than averaged their initial price estimation with the AI recommendation. The study utilized linear mixed-effect models (LMEMs) to examine the impact of feature-importance and counterfactual explanations. Overall, feature-importance and counterfactual explanations across all apartments did not significantly affect the WOA compared to the control group.

However, the predictor that accounted for the type of AI recommendation (increase or decrease in price) was highly significant ($p < .001$), with β -estimates ranging between 95% CI [0.03, 0.07]. A closer investigation revealed that the effect of explanations could only be revealed by considering the type of AI recommendation (see Figure 2). When the AI recommended to decrease the price of an apartment, counterfactual explanations led to a significant increase in WOA by approximately 4% compared to the control ($\beta = 0.04$, 95% CI β [0.01, 0.08], $t(378) = 2.31$, $p = .02$), while feature-importance

explanations showed a non-significant increase of 2% ($\beta = 0.02$, 95% $CI \beta[-0.01, 0.05]$, $t(378) = 1.10$, $p = .27$).

Regarding trust, a Kruskal–Wallis test indicated that the mean ratings for the TPA were not significantly different between the conditions ($H(2) = 1.54$, $p = .46$). Crucially, an exploratory factor analysis (EFA) revealed a two-factor structure for the TPA. The first five items of the TPA loaded on one factor (.79 – .89), while the remaining seven items loaded on a second factor (.56 – .85), corresponding to the trust receptively distrust items of the TPA questionnaire.

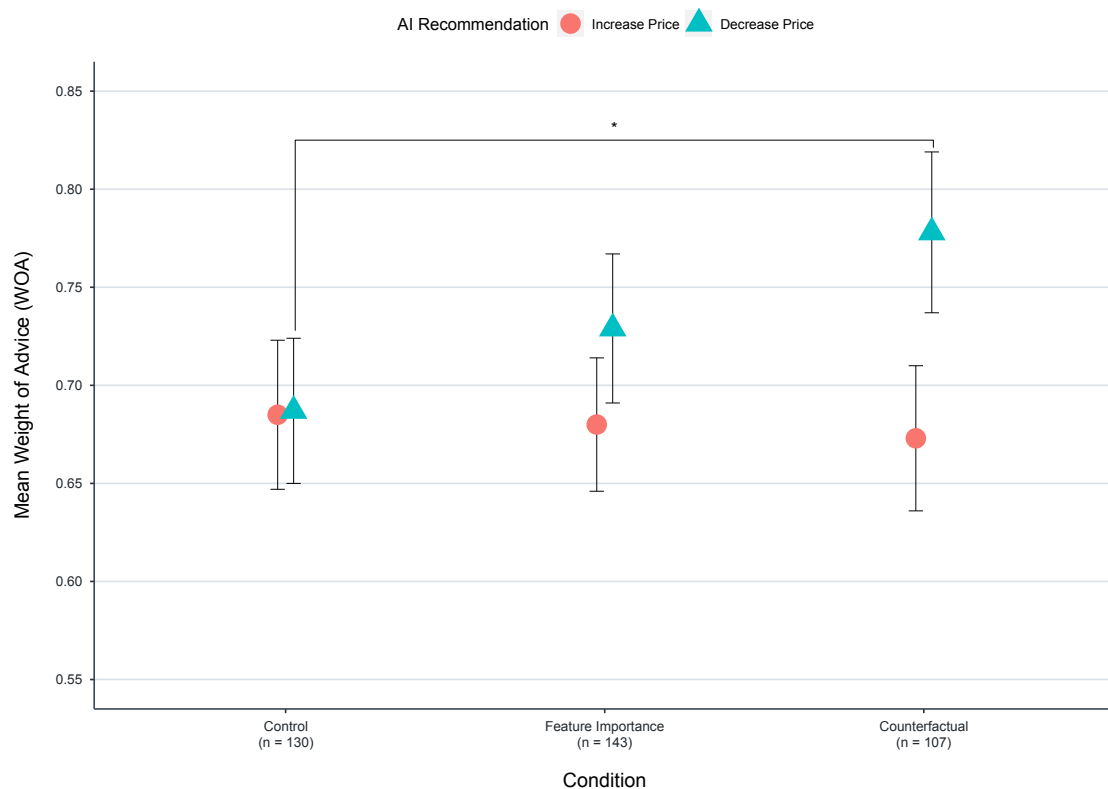


Figure 2

*Mean WOA for each condition and type of recommendation. Note that the y-axis is scaled to better visualize the effect. The error bars depict 95% confidence intervals. * Statistically significant difference with $p < .05$.*

Discussion and Conclusion. The findings of this initial study suggested that the participants generally relied on the AI recommendations since they adopted nearly 70% of the AI recommendations when updating their prior beliefs to form their final price estimate. This is in alignment with existing literature that suggests that people

generally rely on AI in decision-making processes (Logg et al., 2019). Surprisingly, the type of AI recommendation (increasing vs. decreasing the price) played a crucial role in influencing users' reliance, and counterfactual explanations were only significantly effective in increasing reliance when the AI recommended a price decrease. Therefore, while we could show that human-centered post-hoc explanations significantly increased AI reliance, in our study, this effect was contingent on the type of recommendation (increasing vs. decreasing the price), underscoring the importance of contextual and human-related factors such as the type of the decision-making task at hand.

Participants consistently relied more heavily on AI recommendations to decrease the price than on recommendations to increase the price. We argued that this preference could be attributed to certain cognitive biases, particularly loss aversion (Tversky & Kahneman, 1991). Loss aversion suggests that people prefer avoiding losses over acquiring equivalent gains. In the context of our study, participants seemed more concerned about the potential loss (unsuccessful sublease) than the potential gain (higher sublease price), potentially influencing their reliance on AI recommendations to decrease prices. In line with past work (Kliegr et al., 2021; Wang et al., 2019), we argued that loss aversion and similar biases (e.g., base-rate neglect, confirmation bias) could considerably influence human-AI interaction, potentially leading to irrational and suboptimal choices in cases where cognitive biases emerge in the decision-making process.

However, the study design did not allow us to equally investigate the moderating effect of the type of recommendation on trust, as we assessed users' trust only once after the entire task. At the same time, reliance could be measured on the level of each apartment. In summary, our findings did not indicate a consistent effect of human-centered post-hoc explanations on trust, and the potential effect of cognitive biases on trust needs to be further explored. Our findings are thus in line with other research that has provided mixed evidence regarding the effect of XAI methods on trust. Crucially, an investigation of the TPA's factor structure using an EFA implied a two-factor solution, differentiating between trust and distrust, as previously observed

outside the AI context (Spain et al., 2008). This insight has inspired our subsequent research on the psychometric quality of trust questionnaires and the dimensionality of trust.

Manuscript 2: "To Trust or Distrust Trust Measures: Validating Questionnaires for Trust in AI"

Motivation and Aim of the Study. Based on the psychometric findings obtained in the first manuscript, we further investigated trust as a psychological construct and how it can be measured using questionnaires. As mentioned earlier in this dissertation, measuring trust is not without challenges. The motivation for this second study was thus to provide researchers with more reliable and valid questionnaires for measuring trust in AI. Only with validated and standardized measurements that withstand psychometric scrutiny can researchers rely on the results of their studies. Validated measurements also form the basis for replicating or building on existing work. Without validation, findings remain ambiguous or inconsistent, impeding progress in human–AI interaction and XAI research (Lai et al., 2023; Wischnewski et al., 2023).

Within this area of research, the TPA by Jian et al. (2000) is by far the most commonly used questionnaire for measuring trust (Hoff & Bashir, 2015; Kohn et al., 2021; Ueno et al., 2022; Vereschak et al., 2021; Wischnewski et al., 2023). However, the TPA was developed in the context of automation and never validated in the context of AI. In fact, there is only one designated questionnaire for trust in AI, the Trust Scale for the AI Context (TAI) by Hoffman et al. (2023), and it too has not been validated. As a result, researchers often adopt questionnaires from other fields, use self-designed questionnaires, or employ single-items to measure trust, all of which in their own way raise the issue of adherence to psychometric standards.

Another pending research question concerns the dimensionality of trust (Kopp, 2024; Scharowski & Perrig, 2023). As outlined earlier, there are different perspectives on whether trust should be considered one-dimensional (with high trust and low trust at the extremes of one dimension) or two-dimensional (with trust and distrust on two separate dimensions). This uncertainty is reflected by researchers working with the TPA who take either one of two approaches (Ueno et al., 2022): (I) recoding the negatively formulated items of the scale and forming a single trust score or (II) refraining from recoding and creating one score for trust and a second score using the

remaining items for distrust. Our validation study aimed to address the question of the dimensionality of trust and to demonstrate the practical and theoretical implications of considering trust and distrust as two separate constructs.

Method. We conducted a pre-registered 2×2 mixed-design online experiment, recruiting 1,500 participants from the US via Prolific. The study involved two independent variables: the type of AI application (automated vehicle or chatbot) and the trustworthiness condition (trustworthy or untrustworthy). This formed a crossover design with four scenarios (condition \times application). This design ensured that each participant encountered both a trustworthy and an untrustworthy scenario for either AI application. Inspired by prior work (Holthausen et al., 2020; Schaefer, 2016), each participant was presented two out of four pre-recorded videos depicting interactions with an automated vehicle (AV) or a chatbot showcasing trustworthy or untrustworthy AI behavior.

After each video, participants completed the TPA, TAI, and other related scales, such as the Situational Trust Scale (STS; Dolinek & Wintersberger, 2022), the STS for Automated Driving (STS-AD; Holthausen et al., 2020), and the Positive and Negative Affect Schedule (PANAS; Watson et al., 1988), which were used for psychometric validation. The trustworthy AV scenario depicted safe urban driving without automation failure, while the trustworthy chatbot provided correct answers to basic knowledge questions (e.g., "a mouse is smaller than an elephant"). In contrast, the untrustworthy scenarios included an AV that seemed not to slow down for a pedestrian on a crosswalk and a chatbot that provided incorrect answers for the same basic knowledge questions (e.g., "a mouse is bigger than an elephant").

The survey tool verified that participants watched the video, and data quality was further controlled through instructed response items and a self-reported data-quality item as suggested by Brühlmann et al. (2020), leaving us with 1,485 participants and 2,970 valid responses for the data analysis. Participants were spread evenly across the four scenarios and showed an even gender distribution with an average age of 42.98 years ($SD = 13.95$, $min = 18$, $max = 82$).

Results. A variety of psychometric evaluation methods were employed. Initially, we assessed the quality of individual items through metrics such as descriptive statistics, item difficulty and variance, discriminatory power, and inter-item correlations. Most items were found to be inconspicuous, leading to the decision to use overall data for subsequent analyses.

Table 1

Descriptive statistics for all collected measures, separate per condition.

Construct	Chatbot trustworthy		Chatbot untrustworthy		AV trustworthy		AV untrustworthy	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
TPA trust	4.53	1.33	1.78	1.27	4.13	1.43	1.86	1.23
TPA distrust	2.13	1.23	4.60	1.64	2.63	1.24	4.69	1.37
TAI trust	3.55	0.78	1.73	0.81	3.14	0.89	1.65	0.80
SDS-AD situational trust	-	-	-	-	4.58	1.20	1.52	0.96
SDS situational trust	5.49	0.90	2.27	1.18	-	-	-	-
PANAS positive affect	2.65	0.97	2.36	0.87	2.73	0.95	2.41	0.78
PANAS negative affect	1.17	0.42	1.58	0.79	1.44	0.63	2.25	0.98

Note: Responses could range from 1 to 5 for the TAI, and from 1 to 7 for all other measures.

We then used confirmatory factor analysis (CFA) and, where necessary, exploratory factor analysis (EFA) to validate the the construct validity and theoretical models of the scales. The CFA revealed that the TPA’s originally proposed single-factor model did not fit well, while the TAI’s single-factor model was mainly supported. We therefore decided to perform further analyses, exploring alternative models for the TPA while concluding that no such efforts were necessary for the TAI. Due to the poor fit of the TPA’s single-factor model, an EFA was conducted that suggested a two-factor solution. The factor analysis also identified two problematic items (4 and 12), leading to their removal for an improved version of the scale. This version of the TPA was then examined in an alternative CFA and exhibited improved model fit, leading to the conclusion that the TPA should be used as a two-factor model without items 4 and 12. The validity of the two scales was assessed using criterion validity and convergent and divergent validity. For criterion validity, two-way ANOVAs were employed to test the scales across the different conditions and application areas. They revealed, as hypothesized, that the condition (trustworthy vs. untrustworthy) had significant effects on TPA trust and distrust scores and on the TAI score, with a negligible effect for application area (AV vs. chatbot). For convergent and divergent validity, correlations

among the scales and related measures confirmed the expected patterns, supporting the convergent and divergent validity of the scales.

Crucially, while the two trust scores of the TAI and TPA correlated positively with positive affect and negatively with negative affect, the pattern was reversed for the TPA distrust score, demonstrating that distrust and trust were associated with different affects. Reliability was examined through internal consistency indicators, and both scales demonstrated good to excellent alpha and omega coefficients. Finally, we investigated the stability of the model fit for the two scales across the four scenarios (condition \times application), using this as an indicator of the scale's measurement invariance. The TAI model fit mostly well in all scenarios, whereas the TPA's two-factor model only fit in the "untrustworthy chatbot" scenario.

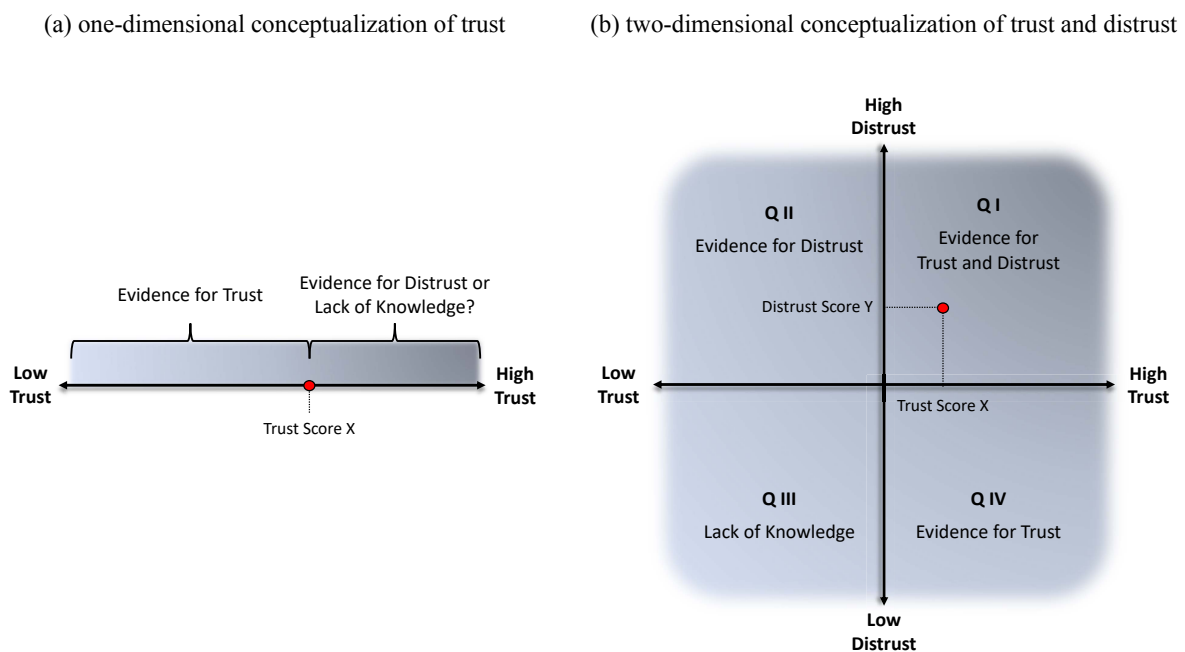


Figure 3

Conceptual frameworks of trust and distrust. (a) the one-dimensional conceptualization places trust on a single continuum ranging from low to high trust (adapted from Castelfranchi and Falcone (2010)). (b) the two-dimensional conceptualization of trust and distrust separates trust and distrust scores into two distinct dimensions. Quadrant I: high trust, high distrust. Quadrant II: low trust, high distrust. Quadrant III: low trust, low distrust. Quadrant IV: high trust, low distrust (adapted from Lewicki et al. (1998)).

Discussion and Conclusion. Based on these results, we formulated recommendations and guidance for researchers aiming to measure trust in AI with the TPA or TAI. The TAI's performance as a one-dimensional measure of trust that captures the spectrum from low trust to high trust was robust across different AI scenarios and conditions. In contrast, the originally proposed single-factor model for the TPA could not be supported, which is in line with previous research (Spain et al., 2008). Instead, our findings suggested a two-factor solution distinguishing between trust and distrust. To account for this two-factor structure, we recommended averaging the distrust items of the TPA to obtain a composite distrust score without any reversal while averaging the remaining items to calculate a trust score. Additionally, we suggested removing the trust items 4 and 12 when applying the scale in the context of AI. Although these measures resulted in some improvement, the performance of the TPA indicated that further work on the scale is required.

These findings led us to argue for a more holistic understanding of trust *and* distrust as two distinct, independent constructs. Trust and distrust were associated with different affects, which provides additional empirical support for a two-dimensional conceptualization that aligns with previous theoretical research in the field of automation and interpersonal trust (e.g., Lewicki et al., 2006; Luhmann, 1979; McKnight & Chervany, 2001a; Ou & Sia, 2009; Saunders et al., 2014; Sitkin & Roth, 1993). While our study could not conclusively resolve whether trust and distrust constitute the same construct at opposite ends of a continuum or should be treated as separate constructs, we emphasized the added value and opportunities of a two-dimensional understanding for XAI and human–AI interaction research. To do so, we adopted the 2×2 framework proposed by Lewicki et al. (1998), which highlights the simultaneous coexistence and development of trust or distrust over time (see Figure 3). Such a distinction could account for both *warranted trust* in trustworthy AI and *warranted distrust* in untrustworthy AI, aligning more closely with the introduced objectives of XAI by Jacovi et al. (2021). Finally, we encouraged future research to improve the understanding and measurement of both trust and distrust.

Manuscript 3: "Certification Labels for Trustworthy AI: Insights from an Empirical Mixed-Method Study"

Motivation and Aim of the Study. In this third study, we explored approaches beyond XAI methods for ascertaining and communicating the trustworthiness of AI to end-users. A growing body of work has started to recognize the critical role of AI auditing for ensuring fairness, accountability, and governance (Avin et al., 2021; Knowles & Richards, 2021; Toreini et al., 2020). While audits have been a long-standing practice in industries where safety is critical, such as aerospace and medicine (Costanza-Chock et al., 2022), the development of standardized AI auditing practices, regulatory guidelines, and auditing frameworks is still evolving (Bandy, 2021; Costanza-Chock et al., 2022; Raji et al., 2020). *How* the outcomes of AI audits can be effectively communicated to end-users presented a promising research opportunity as past research had mainly focused on AI documentation that summarizes and synthesizes information about the models and training datasets used by AI, such as model cards (Crisan et al., 2022; Mitchell et al., 2019), datasheets, (Gebru et al., 2021), and external scorecards (Floridi et al., 2022).

These types of documentation play a crucial role in AI governance by allowing auditors and regulators to determine whether certain principles of trustworthy AI (e.g., fairness, robustness, privacy) have been met (Knowles & Richards, 2021). However, they are tailored to AI practitioners and regulators (Seifert et al., 2019; Yurrita et al., 2022) rather than end-users, who might not have access to the technical information that AI documentation provides or the expertise to understand it (Arnold et al., 2019; Knowles & Richards, 2021). We proposed AI certification labels as a means for communicating to end-users that an audit has considered an AI system trustworthy. Such labels would act as trustworthiness cues (Schlicker et al., 2022) that signal compliance with ethical principles. Despite the growing theoretical and practical interest in AI certification labels (e.g., Fraunhofer Institute for Telecommunications & Heinrich Hertz Institute, HHI, n.d.; Hallensleben et al., 2020), no empirical research has been conducted on end-users' attitudes toward these labels, including their impact on trust in and willingness to use

AI. Our study aimed to fill this research gap by contributing to understanding how certification labels can effectively communicate AI's trustworthiness.

Method. We used a mixed-method research approach that combined semi-structured interviews ($N = 12$) and a subsequent online survey ($N = 302$). Our approach was grounded in a scenario-based method (Binns et al., 2018; Jakesch et al., 2022; Kapania et al., 2022) that used six real-world AI application areas (i.e., medical diagnosis, loan approval, hiring processes, music preferences, route planning, and price comparisons) that covered both low- and high-stakes scenarios. This categorization was crucial as it mirrors the risk-based distinction in AI regulation, such as those by the EU AI Act (European Commission, 2021). As a certification label, we adopted the "Digital Trust Label" developed by the non-profit foundation Swiss Digital Initiative. The label is based on a catalogue of 35 auditable criteria across four categories (security, data protection, reliability, and fair user interaction).

For the qualitative part of the study, we conducted 12 interviews with end-users from diverse backgrounds that lasted 60–90 minutes and explored their attitudes toward AI certification labels. Interviewees ranked the AI scenarios to validate our low- and high-stakes categorizations. They were then introduced to the certification label and were solicited for feedback on the comprehensibility, perceived benefits, and potential drawbacks of such a label. Building on this qualitative foundation, we subsequently designed an online survey to quantitatively assess the effect of certification labels.

For the survey, we used a sample that was representative of the Swiss population regarding age and gender according to the census. The survey employed a within-subjects design and was divided into three parts. First, participants chose one low-stakes and one high-stakes scenario and rated their trust and willingness to use AI. Second, they were introduced to the certification label and asked to evaluate its importance and impact on AI acceptance. In the third part of the survey, we reassessed trust in and willingness to use AI for the same scenarios, but this time with an AI that had received the introduced certification label. This allowed us to measure the effect that a label has on trust in and willingness to use an AI system. Both quantitative and

qualitative results provided a comprehensive understanding of end-users' attitudes toward AI certification labels and their impact on trust in and willingness to use AI in different scenarios.

Results. The findings of a qualitative content analysis (Mayring & Fenzl, 2019) revealed that end-users have nuanced attitudes toward AI certification labels. These attitudes were categorized into *opportunities*, *facilitators*, *limitations*, and *inhibitors*. For example, certification labels were perceived as an opportunity to increase trust, transparency, and fairness, and participants believed that the label's criteria covered relevant concerns, particularly regarding data security and protection. Facilitators for effective certification labels included the need for additional label information, such as details about the auditing process and the independence of the party awarding the label. However, participants also noted limitations and inhibitors to the effectiveness of certification labels. Concerns were raised that such labels do not address all AI-related issues equally, explicitly pointing out the lack of performance measures like accuracy. The potential that an overabundance of different labels could lead to confusion among end-users and the subjective nature of specific label criteria (e.g., fairness) were identified as inhibitors for the label's effectiveness.

In the survey, introducing a certification label significantly increased the trust in and willingness to use the AI in both low- and high-stakes scenarios. The presence of a certification label resulted in the most profound increase in the trust ($M_{\Delta} = 9.12$, $SD = 17.92$, $t(301) = 8.84$, $p < .001$) and willingness to use ($M_{\Delta} = 8.41$, $SD = 17.69$, $t(301) = 8.26$, $p < .001$) ratings in high-stakes scenarios, followed by trust ($M_{\Delta} = 6.57$, $SD = 13.26$, $t(301) = 8.61$, $p < .001$) and willingness to use ($M_{\Delta} = 4.60$, $SD = 17.03$, $t(301) = 4.70$, $p < .001$) ratings in low-stakes scenarios. A majority of the survey participants (63%) preferred the use of certification labels in high-stakes scenarios, stating the complexity and lack of personal expertise in these scenarios as reasons for their preference. Moreover, most participants (71%) expressed a greater likelihood of accepting AI decisions if the AI had received a certification label.

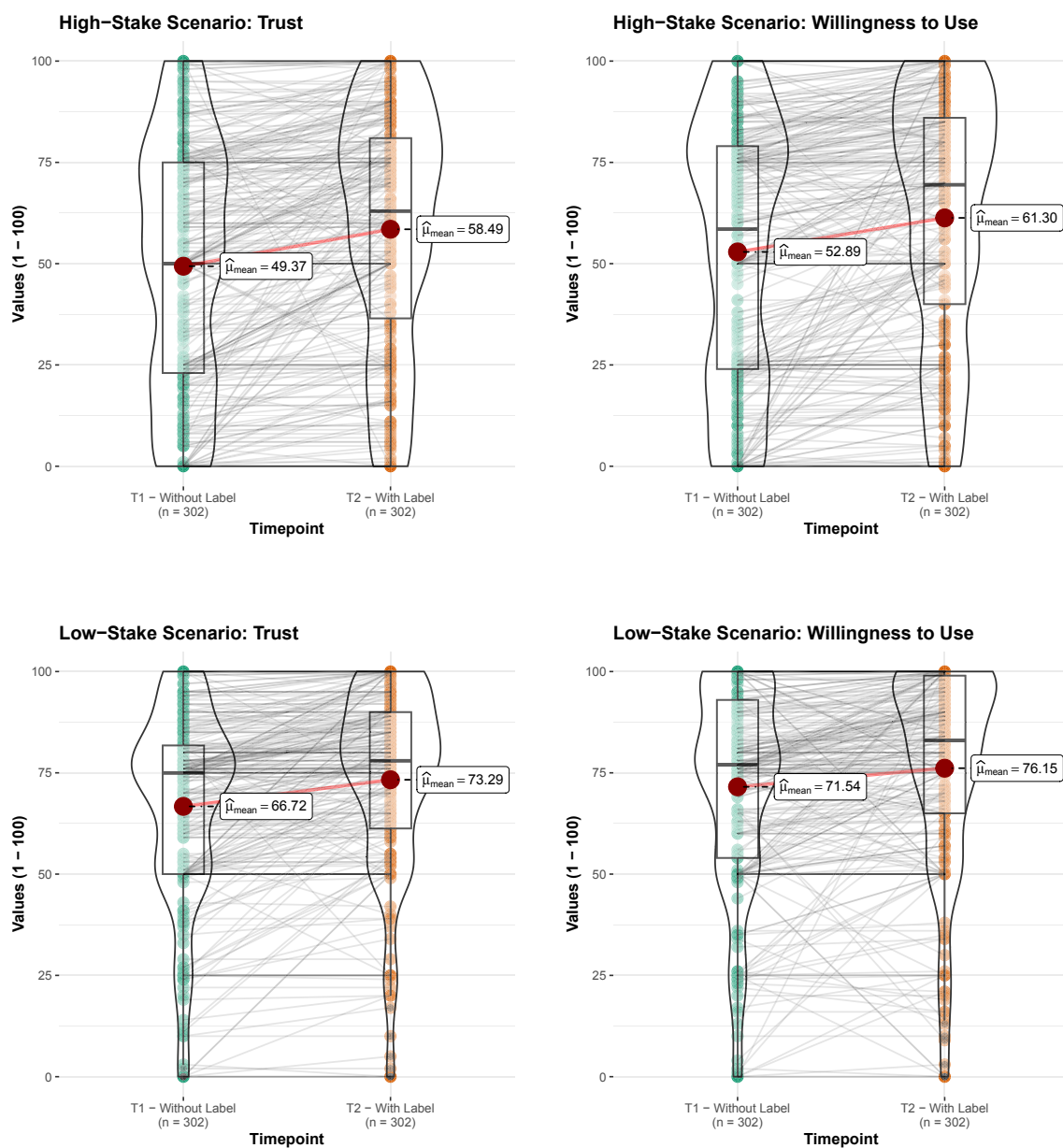


Figure 4

Plots showing individual scores for trust and willingness to use and their respective changes from T1 (without label) to T2 (with label). The plots also depict the medians, means, and distribution of the aggregated low-stakes and high-stakes scenarios. All comparisons revealed statistically significant differences.

Discussion and Conclusion. The quantitative findings demonstrated that certification labels can increase warranted trust in and willingness to use AI in both low- and high-stakes scenarios. These findings resonate with current debates in AI regulation and policy (European Commission, 2020; Stuurman & Lachaud, 2022), which have suggested voluntary labeling for low-stakes AI. However, we found that certification labels could also effectively convey regulatory compliance in high-stakes scenarios, demonstrating the potential for using certification labels in areas that involve more risk. The qualitative part of our study revealed that the effectiveness of AI certification labels face similar challenges to those already identified for other labeling or auditing procedures. For example, the issue of an overabundance of labels, each with different standards, is a challenge also recognized for eco-friendly labels, where numerous labels can create confusion among consumers (Harbaugh et al., 2011). This suggests a need for the harmonization and regulation of certification labels across industries. Furthermore, there is a risk that labels are adopted without meeting the promised criteria, as has been the case with the CE (conformité européenne) logo, which some products have used without actually being manufactured according to EU quality standards (Consumer Research Associates Ltd, 2007). Ensuring that organizations genuinely comply with label criteria is crucial. Otherwise, certification labels might be perceived as mere formalities or "empty promises" that do not reflect truthful trustworthiness cues. However, if such considerations are considered and incorporated into the design process, our study demonstrated that certification labels could be a promising component of an AI trustworthiness ecosystem (Avin et al., 2021) by fostering calibrated trust and willingness to use AI technologies among end-users.

General Discussion

The goal of this dissertation was (I) to contribute to a deeper understanding of trust in AI by highlighting challenges related to its definitions, models, and measures, and (II) to explore approaches and methods to calibrate and increase trust and distrust to a level that is warranted. The manuscripts that form the cornerstones of this dissertation addressed different facets of the introduced challenges in one way or another. In this last section, a final synthesis of the three manuscripts with other relevant studies will present an overarching discussion that could inform future work on trust in AI in XAI and human–AI interaction research.

Identifying and Addressing Challenges in Understanding Trust in AI

Earlier in this dissertation, it was argued that the reasons why the evidence for the effectiveness of XAI methods in increasing warranted trust has remained inconclusive may include (I) inconsistent definitions of trust in AI and its conflation with related but distinct constructs. Manuscript 1 contributed to this conviction and highlighted the complexities involved when investigating the effects of XAI methods on trust in and reliance on AI. It revealed that the effect of post-hoc explanations tailored to human understanding varied between measures of trust and reliance. This finding emphasized the importance of selecting appropriate measures for clearly defined concepts and further motivated a workshop paper (i.e., Scharowski et al., 2022) that outlined theoretical reasons for distinguishing trust from reliance. A more detailed elaboration and the justification for this distinction was presented earlier in this dissertation. Empirical XAI research seems to have sometimes conflated behavioral measures (i.e., reliance) with attitudinal measures (i.e., trust), which could explain the ambiguous research findings, as these measures relate to distinct constructs.

The results of manuscript 1 further indicated that the effects of post-hoc explanations could not be readily understood without considering additional human-related factors in AI-assisted decision-making. In the context of our study, cognitive biases and heuristics that individuals exhibited when receiving AI recommendations had a substantial impact

on the effectiveness of AI explanations. This finding led us to articulate challenge (II), the inadequate integration of relevant constructs into coherent models to enable proper investigation of their relationships. It also motivated a workshop paper (i.e., Scharowski & Opwis, 2021) in which we argued that such cognitive biases have been relatively neglected in XAI research. This contrasts with the extensive focus on biases within training data and the consequent inaccuracies and prejudices in AI predictions (Fazelpour & Danks, 2021). Only more recently has the XAI community begun to incorporate insights into how people reason or make decisions and how these factors can compromise human–AI interaction (e.g., Kliegr et al., 2021; Wang et al., 2019).

By providing empirical evidence of the critical role of such human-related factors, we advocated for the XAI community to consider potential biases and heuristics — e.g., loss aversion (Tversky & Kahneman, 1991), framing effects (Tversky & Kahneman, 1981), or confirmation bias (Wason, 1960) — to fully realize the potential of AI-assisted decision-making. Manuscript 1 revealed that cognitive biases can obscure the effectiveness of XAI methods, suggesting that human-related factors may contribute more to the formation of trust than AI explanations. Indeed, a recent meta-analysis by Kaplan et al. (2023) indicated that the effect sizes associated with transparency as an AI-related factor of trust were relatively small ($d = 0.24$), especially compared to human-related factors of trust in AI, such as propensity to trust ($d = 0.70$), attitudes toward AI ($d = 1.05$), and expertise ($d = 0.47$). Reflecting on the 479 factors identified by Sassmannshausen et al. (2023) as being influential in the formation of trust in AI, empirical research should not simply add more and more constructs to existing trust models. Instead, researchers should focus theory-driven on the most relevant factors and investigate their relationships to increase warranted trust and distrust in AI.

With respect to such contextual factors, Buginca et al. (2020) pointed out that empirical studies in XAI often do not employ actual decision-making tasks that evoke vulnerability through risk. In such experiments, trust does not form because its definition inherently requires that individuals willingly expose themselves to risk and accept vulnerability (Lee & See, 2004; Mayer et al., 1995). This critical role of risk

underscores the importance of contextual factors in the formation of trust in AI. For example, while the risk involved in interacting with a chatbot is comparatively low, trusting a self-driving car can mean the difference between life and death. In addition, the embodiment of AI necessarily limits the types of tasks end-users can perform with it. Chatbots can only chat, while self-driving cars, being physically embodied, present a broader range of behavior and potential interactions (Kaplan et al., 2023). This contrast raises the question of how comparable these interactions with different forms of AI genuinely are and how meaningful it is to generalize this diversity into an overall construct of *trust in AI* in the first place. Depending on the context, different factors in the human–AI interaction may be more or less important in the formation of trust, and it seems crucial that such considerations are taken into account in models and frameworks of trust in AI.

In order to address challenges (I) and (II), we not only advocated for rigorous definitions of trust in AI and a clear distinction between attitudinal and behavioral measures but also for the necessity of integrating trust, trust-related constructs, and other influential factors within a coherent theoretical model or framework. Without such consistencies in definitions and measures and their embedding in a solid theoretical foundation, the empirical claims of XAI research will likely remain ambiguous. Finally, manuscript 1 uncovered psychometric limitations in the TPA by Jian et al. (2000), as the most widely used questionnaire for measuring trust in AI. These limitations were particularly evident since the initially suggested one-factor model for the TPA could not withstand psychometric scrutiny. Instead, a factor analysis favored a two-factor model differentiating trust and distrust. These initial findings, along with their theoretical implications for XAI, were published both as an extended abstract (i.e., Perrig, Scharowski, & Brühlmann, 2023) and a subsequent workshop paper (i.e., Scharowski & Perrig, 2023). They also led us to formulate a third challenge hindering progress in empirical research on XAI and human–AI interaction in general, namely, (III) the inappropriate use or adoption of measures that are not validated in the context of trust in AI.

From Improved Measures to a Different Perspective on Trust *and* Distrust

To address this third challenge, manuscript 2 presented a comprehensive validation study of the TPA as the most commonly used questionnaire for measuring trust in AI and the recently introduced TAI questionnaire by Hoffman et al. (2023). Investigating psychometric quality is particularly important because a variety of trust questionnaires exist, none of which had been previously explicitly validated for AI contexts. While the TAI displayed satisfactory psychometric qualities, the findings of manuscript 2 once again suggested a two-factor solution for the TPA that emphasizes the distinction between trust and distrust. This finding provided a two-dimensional perspective on the measurement of trust in AI, resonating with previous work that separates trust and distrust as two distinct psychological constructs (e.g., Lewicki et al., 2006; Luhmann, 1979; McKnight & Chervany, 2001a; Ou & Sia, 2009; Saunders et al., 2014).

As discussed earlier in this dissertation, the development of trust should be calibrated with the AI's actual trustworthiness for trust to be warranted (Jacovi et al., 2021). This leads to the conclusion that trust in untrustworthy AI is undesirable, while appropriately calibrated distrust in such AI can likewise be considered warranted. Consequently, Jacovi et al. (2021) reasoned that the goals of XAI should be extended from increasing calibrated trust in trustworthy AI to increasing calibrated distrust in untrustworthy AI. Being able to measure *both* constructs appropriately would thus contribute to these objectives. This two-dimensional perspective on trust *and* distrust challenges the normative notion prevalent in societal and scholarly discourses that trust is inherently "good" or "positive," while distrust is "bad" or "negative" (Ou & Sia, 2010). It also critiques the idea that distrust is "an obstacle, a speedbump on the highway of inevitable progress, and skeptical users are pitted in opposition to the evocation of a greater public good" (Krüger & Wilson, 2023, p. 1757).

We argued that distrust is more than merely the absence of trust and advocated for a two-dimensional conceptualization that recognizes trust and distrust, each capable of varying levels of intensity ranging from low to high (see Figure 3). This conceptualization offers certain advantages over a one-dimensional understanding of

trust. For example, within the TPA, items of trust and distrust are aggregated and effectively merged into a single trust score. Such procedures can obscure the underlying reasons for a specific trust score, as it is conceivable that this score is caused either by genuine distrust or merely by a lack of knowledge regarding the AI's trustworthiness (Castelfranchi & Falcone, 2010). Within a one-dimensional conceptualization of trust, it is not possible to meaningfully distinguish between these two cases. However, psychologically, these two scenarios seem worth differentiating, as this same level of trust can be associated with entirely different emotions that result in different behavior.

For instance, a lack of knowledge may spark curiosity in a person and suspend their decision to rely on an AI system while they gather additional cues to form their judgment. In the case of genuine distrust, this judgment has already been made; the person may feel skepticism and decide not to rely on the AI. Therefore, a two-dimensional understanding of trust and distrust provides additional information to differentiate between such cases. This information would make it further possible to categorize individuals based on their respective trust and distrust levels and identify different user groups. For example, some users might display high trust and low distrust, others low trust and high distrust, and yet others might exhibit low levels of both trust and distrust, indicating a lack of knowledge (see Figure 3). Such categorizations could deepen our understanding of the specific concerns and needs of these groups, providing practitioners and researchers with a decision basis to either (I) increase the trustworthiness of their AI systems in the case of low levels of trust or (II) decrease the AI's untrustworthiness in the case of high levels of distrust, aligning more closely with the extended goals of XAI as envisioned by Jacovi et al. (2021).

Crucially, the factors contributing to trust differ from those contributing to distrust (Lewicki et al., 1998). In other areas of HCI, it has been empirically demonstrated that specific design attributes distinctly contribute to the formation of trust while others contribute to distrust (Seckler et al., 2015). For example, privacy issues increase distrust, while visible security signs increase trust. Similarly, within the realm of AI, particularly XAI, specific cues may signal trustworthiness (e.g., certification labels),

while other cues could indicate potential untrustworthiness (e.g., accuracy measures). Treating trust and distrust as mere opposites on a single dimension oversimplifies this dynamic, as this framework presupposes that any cue that increases trust inevitably decreases distrust. However, treating the two constructs as lying on separate dimensions allows for different trustworthiness cues to either increase trust or decrease distrust. Within this framework, changes in one dimension would not necessarily lead to a corresponding change in the other dimension (Ou & Sia, 2010).

While the idea of a simultaneous manifestation of both high levels of trust and distrust may seem counterintuitive or even paradoxical, manuscript 2 detailed how such cases can be rationalized and outlined scenarios where it may be desirable for both trust and distrust to be warranted and calibrated. Since AI is a powerful technology that increasingly has the potential for both positive and negative impacts on society, it may be appropriate to reevaluate the value and adaptive function of distrust and start recognizing it as a protective mechanism that encourages individuals to take necessary precautions against potential risks. For this reason, distrust should be considered a distinct construct deserving of the same level of consideration and attention as trust. The second manuscript of this dissertation thus not only addressed challenge (III), the inappropriate use or adoption of measures that are not validated in the context of trust in AI, but it also provided a richer framework for examining a broader spectrum of attitudes in human–AI interactions. This two-dimensional perspective allows for a more precise and targeted calibration of trust and distrust in response to trustworthy or untrustworthy AI.

Moving beyond XAI to Novel Approaches for Trustworthy AI

This dissertation identified potential reasons for the ambiguous findings regarding the impact of XAI methods on warranted trust and distrust. It contains manuscripts that helped formulate the challenges preventing more conclusive evidence and outlined how these challenges could be addressed. However, the assumption of a straightforward link between XAI and trust has become so irrefutable that some authors have termed it the

"explainability-trust hypothesis" (Kästner et al., 2021; Peters & Visser, 2023). Indeed, the explainability-trust hypothesis appears to have been uncritically adopted in politics and industry, as reflected in legal documents (e.g., European Commission, 2019; The White House, 2022) and strategic frameworks (e.g., IBM Research, 2023; Microsoft Research, 2022), despite a lack of conclusive evidence supporting it. This is particularly disconcerting because it has been pointed out that explanations can be generated in a way that misleads end-users into trusting AI that is not genuinely trustworthy (Lakkaraju & Bastani, 2020) and there are concerns that explanations could be used by AI developers as an "ethical fig leaf" to cover up any shortcomings of the AI (Starke et al., 2023). Similarly, Krüger and Wilson (2023) warned against the resulting "blind trust" in such cases and the dangers associated with treating trust merely as a resource to be exploited for wealth creation and as a veil for private and state actors to push the advancement of AI systems in the name of trust. This concluding section will discuss additional reasons why XAI has not lived up to its expectations and, lastly, introduces novel approaches that extend beyond traditional XAI methods as a way forward.

The seminal work by Ribeiro et al. (2016) offers insights into why XAI may not be as effective for end-users as expected. The authors introduced LIME, a widely adopted XAI method that explains AI predictions, for example, by visualizing how specific symptoms (e.g., headache, sneezing) contribute to a predicted disease (e.g., the flu). They emphasized that the utility of AI explanations is less in their role in identifying *accurate* predictions and more in signaling when predictions might be *inaccurate* and should, therefore, not be trusted. In other words, the utility of AI explanations is twofold: they can signal trustworthiness to form warranted trust or signal untrustworthiness to form warranted distrust. This second utility seems to have been often overlooked. It highlights the role of XAI in assisting stakeholders such as developers with debugging AI during its development or supporting auditors in evaluating AI systems once developed. However, Ferrario and Loi (2022) argued that after an AI system has been deployed, its actual trustworthiness — for example, its ability to make accurate predictions — remains unchanged by XAI methods available to

end-users. Any improvements in the AI's ability are contingent on further debugging or other means of increasing its trustworthiness (Ferrario & Loi, 2022). Consequently, XAI can only indirectly contribute to more trustworthy AI in its development and evaluation stage, but not after deployment, where AI explanations alone do not warrant trust.

To ensure that trust does not become a mere means to an end, efforts to increase end-users' trust should directly focus on increasing the actual trustworthiness of AI. This view also implies that not only are XAI methods unable to contribute to warranted trust among end-users, but AI explanations that reveal potential flaws in predictions could instead contribute to warranted distrust (Kästner et al., 2021). Particularly within the confines of a one-dimensional conceptualization of trust, where a distinction between trust and distrust cannot be readily made, any increase in warranted distrust will necessarily be reflected in low levels of trust (see Figure 3). The ambiguous findings regarding the relationship between XAI and trust may be due to the underappreciated utility of XAI for increasing warranted distrust and the inability to even account for this utility within the limitations of a one-dimensional understanding of trust.

Concerns of this nature prompted manuscript 3 of this dissertation, which explored novel approaches beyond XAI for ensuring trustworthy AI. In this final study, we investigated certification labels as trustworthiness cues. Certification labels can effectively communicate to end-users that an AI has been audited according to specific criteria and deemed trustworthy by qualified experts. While certification labels are not an XAI method per se, we regard communicating the outcomes of such AI audits through labels as a form of transparency. The findings of this manuscript demonstrated that if certain prerequisites are met, certification labels can increase end-users' warranted trust in both low- and high-stakes AI scenarios. Moreover, our qualitative data showed both the opportunities and limitations of such certification labels and highlighted essential factors to consider for their successful implementation.

Crucially, certification labels as trustworthiness cues remove the need for end-users to comprehend explanations or other forms of XAI methods because they signal compliance with governance structures and institutional authorities, which Knowles and

Richards (2021) deemed crucial for ensuring trustworthy AI. Through the associated audit process, which evaluates the AI based on certain criteria, certification labels can provide end-users with a *justified true belief* in the trustworthiness of AI (Ferrario & Loi, 2022), assuming that the audit is genuinely credible. In some sense, trust is effectively replaced by supervision, monitoring, and oversight. This echoes the proverb "trust is good, control is better," which implies that if the control mechanisms of an audit remove uncertainty and vulnerability for end-users, the necessity and functionality of trust disappear with it.

Moreover, the use of certification labels to establish warranted trust in AI resonates with concerns that have been largely and deliberately excluded in this dissertation, particularly philosophical arguments (e.g., Baier, 1986) about whether trust is even a meaningful concept in relation to inanimate objects. As mentioned at the beginning of this dissertation, trust can be regarded as the acceptance of vulnerability (Mayer et al., 1995). However, since objects do not have any understanding of vulnerability and lack intentionality, they cannot, according to the argument, purposefully betray the trust placed in them. In this view, only moral agents meet the necessary conditions for maintaining or exploiting trust relationships (Hawley, 2014).

Although I diverge from these views — as anthropomorphism causes people to attribute intention and agency to AI (Li & Suh, 2022), thereby resulting in trust relations *as if* AI was a moral actor (Coeckelbergh, 2014) — this perspective is worth considering. Certification labels align more closely with the moral prerequisite of trust envisioned by these critics because they shift end-users' trust relation from the AI to trusting the auditors who have awarded the certification label. As a consequence, not only is the role of the trustee in the trust relation shifted from the AI to the auditors, but such labels also have, in our view, the potential to increase both warranted trust (if the presence of a certification label signals trustworthiness) and warranted distrust (if the absence of a certification label signals untrustworthiness).

One way forward that reflects the above discussions about different approaches and methods beyond XAI to increase end-users' warranted trust and distrust may be

adopting a more holistic view for communicating the trustworthiness of AI. It might be useful to consider XAI methods as *one* component in a broader AI-trustworthiness ecosystem, where developers or auditors can use XAI methods to ascertain whether an AI is trustworthy. With regard to end-users, the explainability-trust hypothesis may oversimplify the sociotechnical environment in which people interact with AI systems. Increasing warranted trust for end-users may require more than just designing systems whose results can be explained. Moving forward, it seems crucial to appreciate the interplay between different XAI methods in the AI life cycle, in auditing practices and legal guidelines, and within the general public discourse about the types of AI we as a society want to engage with in the first place.

This leads me to conclude that XAI should not be seen as a universal remedy but as one among many approaches to ensuring trustworthy AI, a conclusion similar to those made regarding XAI and fairness (Deck et al., 2023). While the XAI community is still engaged in ongoing discussions on the challenges addressed in this dissertation, the AI industry is surging forward, leading to an ever-deeper integration of AI in our daily lives. Generative AI introduces new questions and presents unique challenges in thinking about trust and using XAI in, for example, LLM-based applications. This warrants research focused on ensuring trustworthiness for these applications and efforts on identifying which trustworthiness cues end-users deem relevant for these forms of AI (e.g., Benk et al., 2024). While the future will reveal what further advancements in AI will hold, I remain convinced that it is essential to develop human-centered systems, ensuring that human needs and values guide the development of AI.

Limitations and Future Directions

While manuscript 1 of this dissertation highlighted that a distinction between trust as a psychological construct and reliance as behavior is theoretically justified and useful, implementing this distinction in actual experiments can be challenging. Behavior can be measured at several points in an interaction, however, measuring trust via lengthy questionnaires is not as feasible. This leads to a different granularity of these measures, complicating a direct comparison of trust and reliance.

One possible approach to obtaining the same granularity could be to measure trust with single-items, which have been shown to have a similar sensitivity to trust violations as questionnaires with multiple-items in the context of human–robot interaction (Nesbet et al., 2022). However, there are several reasons against such an approach, and researchers should critically question whether single-items are suitable for measuring complex constructs such as trust (Loo, 2002). Despite acknowledging the issues that come with single-items, we also made use of them in manuscript 2 because of certain benefits in their application, such as less task disruption for participants. Researchers should thus always weigh the advantages and disadvantages of using single-items instead of questionnaires, which are longer but potentially less error-prone and thus more valid. In a similar vein, we encouraged future research to explore the dynamic nature of trust, which has received less attention because trust is rarely measured multiple times during an interaction (Vereschak et al., 2021). This can lead to the false impression that trust reaches a stable state over time, when in fact it changes dynamically throughout the human–AI interaction. Ideally, this dynamic could be assessed by combining trust and reliance measures, which could provide interesting insights into the dynamic changes in trust and reliance over time.

Second, while we provided theoretical reasons for and emphasized certain advantages of distinguishing between trust and distrust, methodological limitations of statistical factor analysis could be an alternative explanation for the revealed two-factor structure of the TPA, and we recognize these limitations. Negatively formulated items, such as the distrust items of the TPA, can lead to response patterns that may load on two

distinct factors in a factor analysis and distort factor structures. This has been shown for scales of usability (Lewis & Sauro, 2017; Lewis et al., 2013) and website aesthetics (Perrig, von Felten, et al., 2023). However, whereas separate constructs lacked theoretical justifications in these cases, we have extensively presented theoretical arguments to justify a distinction between trust and distrust.

Ultimately, the underlying structure of psychological constructs like trust is rooted not in statistical considerations but in theoretical ones (Fried, 2020). Future research could explore how trust and distrust as distinct constructs contribute differently to behavioral outcomes in the human–AI interaction. For example, trust and distrust might impact reliance differently, with trust potentially increasing the likelihood of reliance more than distrust decreasing it. Such considerations of the constructs’ predictive power for behavior could be an interesting opportunity for future research. Understanding these dynamics could enable a more nuanced understanding of human–AI interaction, with differing strategies for effectively increasing warranted trust to mitigate under-reliance or to increase warranted distrust to prevent over-reliance.

Finally, the usefulness of talking about *trust in AI* as an overall construct for all forms of AI can be questioned. Not only do end-users arguably have little knowledge of the areas of life in which AI is already being used (Selwyn & Gallo Cordoba, 2022), but terminological differences in descriptions of AI also influence the capabilities attributed to such systems (Langer et al., 2022). Together with the differences in risk perception that we could show in Manuscript 2 for automated vehicles compared to chatbots, this raises the question of how meaningful a general construct of trust is across different applications, contexts, and interaction possibilities. This seems like asking someone “how much they trust electricity,” although the exact types of electricity-powered applications seem to matter when assessing their trust. As the adoption of AI progresses, future research should investigate how different manifestations and embodiments of AI affect end-users’ trust and which factors genuinely are important in these interactions, taking into account the 479 factors identified by Sassmannshausen et al. (2023) as influential in building trust in AI.

Conclusion

This dissertation has laid out the complexities and challenges associated with understanding trust in AI and has explored approaches and methods to calibrate and increase both trust and distrust to a level that is warranted. These interconnected challenges emerge from attempts to define trust and persist through the operationalization processes of modeling and measuring it. The manuscripts that form this dissertation have identified and addressed some of these challenges.

Manuscript 1 illustrated that trust, as an attitude, and reliance, as a behavior, yield different results when applying XAI methods, underscoring the importance of consistent definitions of trust and its distinction from related but distinct constructs. The findings of manuscript 1 also suggested that human-related factors, such as the influence of cognitive biases, may obscure the effects of XAI methods and that adequately integrating such factors into models of trust is crucial.

Manuscript 2 assessed the psychometric quality of questionnaires as the most relevant measures of trust in AI. This work contributed to more reliable and valid measures and provided guidance for researchers and practitioners using trust questionnaires.

Furthermore, it expanded the one-dimensional understanding of trust to a more holistic and nuanced view of trust *and* distrust, emphasizing the advantages of this broader perspective for XAI and human–AI research.

Manuscript 3 explored approaches to trustworthy AI beyond XAI, namely certification labels as cues that signal credible audit processes. We elaborated on the benefits of this approach for end-users, such as avoiding the need for them to comprehend AI explanations. Moreover, the quantitative and qualitative findings demonstrated the potential of such approaches to increase both warranted trust and warranted distrust. Overall, this dissertation aimed to contribute to a deeper understanding of trust in AI and, more specifically, how trust is defined, modeled, and measured and how to meaningfully calibrate trust in AI. It offered perspectives on increasing both trust and distrust to ensure that both are accurately calibrated with AI's trustworthiness to be warranted.

Despite the identified challenges, I remain convinced that trust matters, be it in interactions between humans or between humans and AI. However, for trust and distrust to be effectively investigated, it is essential to acknowledge and address the challenges discussed in this dissertation and similar ones. Only through dedicated scrutiny can we advance our understanding of trust in AI. Such an understanding is fundamental for developing human-centered AI that safeguards human control over this powerful technology, ensuring that people interact with AI that has genuinely earned their trust.

References

- Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–18. <https://doi.org/10.1145/3173574.3174156>
- 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>
- Aeschbach, L. F., Perrig, S. A. C., Weder, L., Opwis, K., & Brühlmann, F. (2021). Transparency in measurement reporting: A systematic literature review of CHI PLAY. *Proc. ACM Hum.-Comput. Interact.*, 5(CHI PLAY). <https://doi.org/10.1145/3474660>
- Ahn, D., Almaatouq, A., Gulabani, M., & Hosanagar, K. (2024). Impact of model interpretability and outcome feedback on trust in AI. *Proceedings of the CHI Conference on Human Factors in Computing Systems*. <https://doi.org/10.1145/3613904.3642780>
- Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Prentice-Hall.
- Almeida, D., Shmarko, K., & Lomas, E. (2022). The ethics of facial recognition technologies, surveillance, and accountability in an age of artificial intelligence: A comparative analysis of US, EU, and UK regulatory frameworks. *AI and Ethics*, 2(3), 377–387. <https://doi.org/10.1007/s43681-021-00077-w>
- Arnold, M., Bellamy, R. K. E., Hind, M., Houde, S., Mehta, S., Mojsilović, A., Nair, R., Ramamurthy, K. N., Olteanu, A., Piorkowski, D., Reimer, D., Richards, J., Tsay, J., & Varshney, K. R. (2019). Factsheets: Increasing trust in AI services through supplier’s declarations of conformity. *IBM Journal of Research and Development*, 63(4/5), 6:1–6:13. <https://doi.org/10.1147/JRD.2019.2942288>
- Arrieta, A. B., Díaz-Rodríguez, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F.

- (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Avin, S., Belfield, H., Brundage, M., Krueger, G., Wang, J., Weller, A., Anderljung, M., Krawczuk, I., Krueger, D., Lebensold, J., et al. (2021). Filling gaps in trustworthy development of AI. *Science*, 374(6573), 1327–1329. <https://doi.org/10.1126/science.abi7176>
- Bach, T. A., Khan, A., Hallock, H., Beltrão, G., & Sousa, S. (2024). A systematic literature review of user trust in AI-enabled systems: An HCI perspective. *International Journal of Human–Computer Interaction*, 40(5), 1251–1266. <https://doi.org/10.1080/10447318.2022.2138826>
- Baier, A. (1986). Trust and antitrust. *Ethics*, 96(2), 231–260. <https://doi.org/10.1086/292745>
- Bandy, J. (2021). Problematic machine behavior: A systematic literature review of algorithm audits. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1). <https://doi.org/10.1145/3449148>
- Beldad, A., de Jong, M., & Steehouder, M. (2010). How shall I trust the faceless and the intangible? A literature review on the antecedents of online trust. *Computers in Human Behavior*, 26(5), 857–869. <https://doi.org/10.1016/j.chb.2010.03.013>
- Benk, M., Wettstein, L., Schlicker, N., von Wangenheim, F., & Scharowski, N. (2024). *Bridging the knowledge gap: Understanding user expectations for trustworthy LLM standards* [Manuscript submitted for publication].
- Benk, M., Tolmeijer, S., von Wangenheim, F., & Ferrario, A. (2022). The value of measuring trust in AI - A socio-technical system perspective. *CHI TRAIT Workshop*. <https://doi.org/10.48550/arXiv.2204.13480>
- Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and economic behavior*, 10(1), 122–142. <https://doi.org/10.1006/game.1995.1027>

- 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*, 1–14. <https://doi.org/10.1145/3173574.3173951>
- Brühlmann, F., Petralito, S., Aeschbach, L., & Opwis, K. (2020). The quality of data collected online: An investigation of careless responding in a crowdsourced sample. *Methods in Psychology*, 2, 100022. <https://doi.org/10.1016/j.metip.2020.100022>
- Buçinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L. (2020). Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 454–464. <https://doi.org/10.1145/3377325.3377498>
- Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big data & society*, 3(1), 2053951715622512. <https://doi.org/10.1177/205395171562251>
- Cacioppo, J. T., & Berntson, G. G. (1994). Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates. *Psychological Bulletin*, 115(3), 401–423. <https://doi.org/10.1037/0033-2909.115.3.401>
- Cai, C. J., Jongejan, J., & Holbrook, J. (2019). The effects of example-based explanations in a machine learning interface. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 258–262. <https://doi.org/10.1145/3301275.3302289>
- Castelfranchi, C., & Falcone, R. (2010). *Trust theory: A socio-cognitive and computational model*. John Wiley & Sons. <https://doi.org/10.1002/9780470519851>
- Chang, Y.-S., & Fang, S.-R. (2013). Antecedents and distinctions between online trust and distrust: Predicting high-and low-risk internet behaviors. *Journal of Electronic Commerce Research*, 14(2), 149.

- Cheng, H.-F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M., & Zhu, H. (2019). Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. <https://doi.org/10.1145/3290605.3300789>
- Coeckelbergh, M. (2014). The moral standing of machines: Towards a relational and non-cartesian moral hermeneutics. *Philosophy & technology*, 27, 61–77. <https://doi.org/10.1007/s13347-013-0133-8>
- Consumer Research Associates Ltd. (2007). *EFTA study on certification and marks in Europe*. <https://www.efta.int/sites/default/files/publications/study-certification-marks/executive-summary.pdf>
- Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022). Who audits the auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1571–1583. <https://doi.org/10.1145/3531146.3533213>
- Crisan, A., Drouhard, M., Vig, J., & Rajani, N. (2022). Interactive model cards: A human-centered approach to model documentation. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 427–439. <https://doi.org/10.1145/3531146.3533108>
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319–340. <https://doi.org/10.2307/249008>
- Deck, L., Schoeffer, J., De-Arteaga, M., & Kühl, N. (2023). A critical survey on fairness benefits of XAI. *arXiv preprint arXiv:2310.13007*.
- de Visser, E. J., Cohen, M., Freedy, A., & Parasuraman, R. (2014). A design methodology for trust cue calibration in cognitive agents. *Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments: 6th International Conference, VAMR 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014, Proceedings, Part I* 6, 251–262. https://doi.org/10.1007/978-3-319-07458-0_24

- Dolinek, L., & Wintersberger, P. (2022). Towards a generalized scale to measure situational trust in AI systems. *CHI 2022 TRAIT Workshop on Trust and Reliance in AI-Human Teams*.
- Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. *Commun. ACM*, 63(1), 68–77. <https://doi.org/10.1145/3359786>
- Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American journal of community psychology*, 41, 327–350. <https://doi.org/10.1007/s10464-008-9165-0>
- Ehsan, U., Passi, S., Liao, Q. V., Chan, L., Lee, I., Muller, M., Riedl, M. O., et al. (2021). The who in explainable AI: How AI background shapes perceptions of AI explanations. *arXiv preprint arXiv:2107.13509*.
- Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., & Riedl, M. O. (2019). Automated rationale generation: A technique for explainable AI and its effects on human perceptions. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 263–274. <https://doi.org/10.1145/3301275.3302316>
- Ehsan, U., Wintersberger, P., Liao, Q. V., Mara, M., Streit, M., Wachter, S., Riener, A., & Riedl, M. O. (2021). Operationalizing human-centered perspectives in explainable AI. *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. <https://doi.org/10.1145/3411763.3441342>
- Ehsan, U., Wintersberger, P., Liao, Q. V., Watkins, E. A., Manger, C., Daumé III, H., Riener, A., & Riedl, M. O. (2022). Human-centered explainable AI (HCXAI): Beyond opening the black-box of AI. *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*. <https://doi.org/10.1145/3491101.3503727>
- Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, 16(4), 779–788. <https://doi.org/10.1177/1745691620970586>

- European Commission. (2019). *Ethics guidelines for trustworthy ai* [Accessed: 2024-03-27].
<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- European Commission. (2020). *White paper on artificial intelligence: A european approach to excellence and trust* (COM (2020) 65 final). Official Journal of European Union. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0065&qid=1675254609974>
- European Commission. (2021). *Proposal for a regulation of the european parliament and of the council laying down harmonized rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts* (COM (2021) 206 final). Official Journal of European Union.
<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206>
- Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*, 16(8). <https://doi.org/10.1111/phc3.12760>
- Ferrario, A., & Loi, M. (2022). How explainability contributes to trust in AI. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1457–1466. <https://doi.org/10.1145/3531146.3533202>
- Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, 3(4), 456–465.
<https://doi.org/10.1177/2515245920952393>
- Floridi, L., Holweg, M., Taddeo, M., Amaya Silva, J., Mökander, J., & Wen, Y. (2022). CapAI - a procedure for conducting conformity assessment of ai systems in line with the EU Artificial Intelligence Act. *Available at SSRN*:
<https://ssrn.com/abstract=4064091>. <https://dx.doi.org/10.2139/ssrn.4064091>
- Fraunhofer Institute for Telecommunications & Heinrich Hertz Institute, HHI. (n.d.). *Auditing and certification of AI systems*.
<https://www.hhi.fraunhofer.de/en/departments/ai/technologies-and-solutions/auditing-and-certification-of-ai-systems.html>

- Fried, E. I. (2020). Theories and models: What they are, what they are for, and what they are about. *Psychological Inquiry*, *31*(4), 336–344.
<https://doi.org/10.1080/1047840X.2020.1854011>
- Furr, M. (2011). *Scale construction and psychometrics for social and personality psychology*. SAGE publications, Ltd.
- Gambetta, D. (2000). Can we trust trust? In D. Gambetta (Ed.), *Trust: Making and Breaking Cooperative Relations, electronic edition* (pp. 213–237). Department of Sociology, University of Oxford.
<http://www.sociology.ox.ac.uk/papers/gambetta213-237.pdf>
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., & Crawford, K. (2021). Datasheets for datasets. *Commun. ACM*, *64*(12), 86–92.
<https://doi.org/10.1145/3458723>
- Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660.
<https://doi.org/10.5465/annals.2018.0057>
- Hallensleben, S., Hustedt, C., Fetic, L., Fleischer, T., Grünke, P., Hagendorff, T., Hauer, M., Hauschke, A., Heesen, J., Herrmann, M., Hillerbrand, R., Hubig, C., Kaminski, A., Krafft, T., Loh, W., Otto, P., & Puntschuh, M. (2020). From principles to practice – An interdisciplinary framework to operationalise AI ethics. *Artificial Intelligence Ethics Impact Group*.
<https://www.ai-ethics-impact.org/en>
- Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517–527.
<https://doi.org/10.1177/0018720811417254>
- Harbaugh, R., Maxwell, J. W., & Roussillon, B. (2011). Label confusion: The groucho effect of uncertain standards. *Management science*, *57*(9), 1512–1527.
<https://doi.org/10.1287/mnsc.1110.1412>

- Harvey, N., & Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational behavior and human decision processes*, 70(2), 117–133. <https://doi.org/10.1006/obhd.1997.2697>
- Hawley, K. (2014). Trust, distrust and commitment. *Noûs*, 48(1), 1–20. <https://doi.org/https://doi.org/10.1111/nous.12000>
- 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>
- Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2023). Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science*, 5. <https://doi.org/10.3389/fcomp.2023.1096257>
- Holthausen, B. E., Wintersberger, P., Walker, B. N., & Riener, A. (2020). Situational trust scale for automated driving (sts-ad): Development and initial validation. *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 40–47. <https://doi.org/10.1145/3409120.3410637>
- Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation*. Pearson.
- IBM Research. (2023). *What is explainable AI?* [Accessed: 2024-04-29]. <https://www.ibm.com/topics/explainable-ai>
- Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 624–635. <https://doi.org/10.1145/3442188.3445923>
- Jakesch, M., Buçinca, Z., Amershi, S., & Olteanu, A. (2022). How different groups prioritize ethical values for responsible AI. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 310–323. <https://doi.org/10.1145/3531146.3533097>

- Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1), 53–71.
https://doi.org/10.1207/S15327566IJCE0401_04
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1, 389–399.
<https://doi.org/10.1038/s42256-019-0088-2>
- Juniper, E. F. (2009). Validated questionnaires should not be modified. *European Respiratory Journal*, 34(5), 1015–1017.
<https://doi.org/10.1183/09031936.00110209>
- Kapania, S., Siy, O., Clapper, G., SP, A. M., & Sambasivan, N. (2022). “Because AI is 100% right and safe”: User attitudes and sources of AI authority in india. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. <https://doi.org/10.1145/3491102.3517533>
- Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. (2023). Trust in artificial intelligence: Meta-analytic findings. *Human factors*, 65(2), 337–359.
<https://doi.org/10.1177/00187208211013988>
- Kästner, L., Langer, M., Lazar, V., Schomäcker, A., Speith, T., & Sterz, S. (2021). On the relation of trust and explainability: Why to engineer for trustworthiness. *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, 169–175. <https://doi.org/10.1109/REW53955.2021.00031>
- Kaur, D., Uslu, S., Rittichier, K. J., & Durresti, A. (2022). Trustworthy artificial intelligence: A review. *ACM Comput. Surv.*, 55(2).
<https://doi.org/10.1145/3491209>
- Kirlik, A. (1993). Modeling strategic behavior in human-automation interaction: Why an "aid" can (and should) go unused. *Human Factors*, 35(2), 221–242.
<https://doi.org/10.1177/001872089303500203>
- Kizilcec, R. F. (2016). How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 chi conference on human factors*

- in computing systems* (pp. 2390–2395). ACM.
<https://doi.org/10.1145/2858036.2858402>
- Kliegr, T., Bahník, Š., & Fürnkranz, J. (2021). A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence*, *295*, 103458. <https://doi.org/10.1016/j.artint.2021.103458>
- Knowles, B., & Richards, J. T. (2021). The sanction of authority: Promoting public trust in AI. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 262–271.
<https://doi.org/10.1145/3442188.3445890>
- Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y.-C., & Shaw, T. H. (2021). Measurement of trust in automation: A narrative review and reference guide. *Frontiers in Psychology*, *12*, 604977. <https://doi.org/10.3389/fpsyg.2021.604977>
- Kopp, T. (2024). Facets of trust and distrust in collaborative robots at the workplace: Towards a multidimensional and relational conceptualisation. *International Journal of Social Robotics*, 1–18. <https://doi.org/10.1007/s12369-023-01082-1>
- Körber, M. (2019, August). Theoretical considerations and development of a questionnaire to measure trust in automation. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), *Proceedings of the 20th congress of the international ergonomics association (IEA 2018)* (pp. 13–30). Springer.
https://doi.org/10.1007/978-3-319-96074-6_2
- Kroeger, F. (2019). Unlocking the treasure trove: How can Luhmann’s theory of trust enrich trust research? *Journal of Trust Research*, *9*(1), 110–124.
<https://doi.org/10.1080/21515581.2018.1552592>
- Krüger, S., & Wilson, C. (2023). The problem with trust: On the discursive commodification of trust in AI. *AI & SOCIETY*, *38*(4), 1753–1761.
<https://doi.org/https://doi.org/10.1007/s00146-022-01401-6>
- Lai, V., Chen, C., Smith-Renner, A., Liao, Q. V., & Tan, C. (2023). Towards a science of human-AI decision making: An overview of design space in empirical human-subject studies. *Proceedings of the 2023 ACM Conference on Fairness,*

Accountability, and Transparency, 1369–1385.

<https://doi.org/10.1145/3593013.3594087>

Lai, V., & Tan, C. (2019). On human predictions with explanations and predictions of machine learning models: A case study on deception detection. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 29–38.

<https://doi.org/10.1145/3287560.3287590>

Lakkaraju, H., & Bastani, O. (2020). "How do I fool you?": Manipulating user trust via misleading black box explanations. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 79–85. <https://doi.org/10.1145/3375627.3375833>

Langer, M., Hunsicker, T., Feldkamp, T., König, C. J., & Grgić-Hlača, N. (2022).

“Look! It’s a computer program! It’s an algorithm! It’s AI!”: Does terminology affect human perceptions and evaluations of algorithmic decision-making systems? *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. <https://doi.org/10.1145/3491102.3517527>

Langer, M., Oster, D., Speith, T., Kästner, L., Hermanns, H., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from explainable artificial intelligence (XAI)? A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, 296, 103473.

<https://doi.org/10.1016/j.artint.2021.103473>

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

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges. *Philosophy & Technology*, 31(4), 611–627. <https://doi.org/https://doi.org/10.1007/s13347-017-0279-x>

Lewicki, R. J., McAllister, D. J., & Bies, R. J. (1998). Trust and distrust: New relationships and realities. *The Academy of Management Review*, 23(3), 438–458. <https://doi.org/10.2307/259288>

- Lewicki, R. J., Tomlinson, E. C., & Gillespie, N. (2006). Models of interpersonal trust development: Theoretical approaches, empirical evidence, and future directions. *Journal of management*, *32*(6), 991–1022.
<https://doi.org/10.1177/0149206306294405>
- Lewis, J. R., & Sauro, J. (2017). Revisiting the factor structure of the system usability scale. *Journal of Usability Studies*, *12*(4), 183–192.
- Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013). UMUX-LITE: When there's no time for the SUS. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2099–2102. <https://doi.org/10.1145/2470654.2481287>
- Li, M., & Suh, A. (2022). Anthropomorphism in AI-enabled technology: A literature review. *Electronic Markets*, *32*(4), 2245–2275.
<https://doi.org/https://doi.org/10.1007/s12525-022-00591-7>
- Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing design practices for explainable AI user experiences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15.
<https://doi.org/10.1145/3313831.3376590>
- Liao, Q., & Sundar, S. S. (2022). Designing for responsible trust in AI systems: A communication perspective. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1257–1268.
<https://doi.org/10.1145/3531146.3533182>
- Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57.
<https://doi.org/10.1145/3236386.3241340>
- Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. <https://doi.org/10.1016/j.obhdp.2018.12.005>
- Loo, R. (2002). A caveat on using single-item versus multiple-item scales. *Journal of Managerial Psychology*, *17*(1), 68–75.
<https://doi.org/10.1108/02683940210415933>

- Luhmann, N. (1979). *Trust and power*. Wiley.
- Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. *11th Australasian conference on information systems*, 53, 6–8.
- Mayer, R. C., & Davis, J. H. (1999). The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of applied psychology*, 84(1), 123. <https://doi.org/10.1037/0021-9010.84.1.123>
- Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, 20(3), 709–734. <https://doi.org/10.2307/258792>
- Mayring, P., & Fenzl, T. (2019). Qualitative Inhaltsanalyse. In N. Baur & J. Blasius (Eds.), *Handbuch methoden der empirischen sozialforschung* (pp. 633–648). Springer VS. https://doi.org/10.1007/978-3-658-21308-4_42
- McKnight, D. H., & Chervany, N. L. (2001a). What trust means in e-commerce customer relationships: An interdisciplinary conceptual typology. *International journal of electronic commerce*, 6(2), 35–59. <https://doi.org/10.1080/10864415.2001.11044235>
- McKnight, D. H., & Chervany, N. L. (2001b). While trust is cool and collected, distrust is fiery and frenzied: A model of distrust concepts. *7th Americas Conference on Information Systems*, 883–888.
- Merritt, S. M. (2011). Affective processes in human–automation interactions. *Human Factors*, 53(4), 356–370. <https://doi.org/10.1177/0018720811411912>
- Microsoft Research. (2022). *Explainability* [Accessed: 2024-03-27]. <https://www.microsoft.com/en-us/research/group/dynamics-insights-apps-artificial-intelligence-machine-learning/articles/explainability/>
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38. <https://doi.org/https://doi.org/10.1016/j.artint.2018.07.007>
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting.

- Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. <https://doi.org/10.1145/3287560.3287596>
- Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 279–288. <https://doi.org/10.1145/3287560.3287574>
- Mohseni, S., Zarei, N., & Ragan, E. D. (2020). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. <https://arxiv.org/abs/1811.11839>
- Molnar, C. (2019). *Interpretable machine learning: A guide for making black box models explainable* [Retrieved from <https://christophm.github.io/interpretable-ml-book/>].
- Nesset, B., Rajendran, G., Aguas Lopes, J. D., & Hastie, H. (2022). Sensitivity of trust scales in the face of errors. *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, 950–954. <https://doi.org/10.1109/HRI53351.2022.9889427>
- Nothdurft, F., Heinroth, T., & Minker, W. (2013). The impact of explanation dialogues on human-computer trust. *Proceedings, Part III, of the 15th International Conference on Human-Computer Interaction. Users and Contexts of Use - Volume 8006*, 59–67. https://doi.org/10.1007/978-3-642-39265-8_7
- Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic bulletin & review*, 26, 1596–1618. <https://doi.org/https://doi.org/10.3758/s13423-019-01645-2>
- Ou, C. X., & Sia, C. L. (2009). To trust or to distrust, that is the question: Investigating the trust-distrust paradox. *Commun. ACM*, 52(5), 135–139. <https://doi.org/10.1145/1506409.1506442>
- Ou, C. X., & Sia, C. L. (2010). Consumer trust and distrust: An issue of website design. *International Journal of Human-Computer Studies*, 68(12), 913–934. <https://doi.org/https://doi.org/10.1016/j.ijhcs.2010.08.003>

- Papenmeier, A., Kern, D., Englebienne, G., & Seifert, C. (2022). It's complicated: The relationship between user trust, model accuracy and explanations in AI. *ACM Trans. Comput.-Hum. Interact.*, 29(4). <https://doi.org/10.1145/3495013>
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230–253.
<https://doi.org/10.1518/001872097778543886>
- Perrig, S. A. C., Scharowski, N., & Brühlmann, F. (2023). Trust issues with trust scales: Examining the psychometric quality of trust measures in the context of AI. *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. <https://doi.org/10.1145/3544549.3585808>
- Perrig, S. A. C., von Felten, N., Honda, M., Opwis, K., & Brühlmann, F. (2023). Development and validation of a positive-item version of the visual aesthetics of websites inventory: The VisAWI-Pos. *International Journal of Human-Computer Interaction*, 0(0), 1–25. <https://doi.org/10.1080/10447318.2023.2258634>
- Peters, T. M., & Visser, R. W. (2023). The importance of distrust in AI. In L. Longo (Ed.), *Explainable artificial intelligence* (pp. 301–317). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-44070-0_15
- Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. <https://doi.org/10.1145/3411764.3445315>
- Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. <https://doi.org/10.1145/3351095.3372873>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In B. Krishnapuram, M. Shah, A. Smola, C. Aggarwal, D. Shen, & R. Rastogi (Eds.), *Proceedings of the 22nd acm*

- SIGKDD international conference on knowledge discovery and data mining - KDD '16* (pp. 1135–1144). ACM. <https://doi.org/10.1145/2939672.2939778>
- Riegelsberger, J., Sasse, M. A., & McCarthy, J. D. (2005). The mechanics of trust: A framework for research and design. *International Journal of Human-Computer Studies*, *62*(3), 381–422.
<https://doi.org/https://doi.org/10.1016/j.ijhcs.2005.01.001>
- Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, *23*(3), 393–404. <https://doi.org/10.5465/AMR.1998.926617>
- Sassmannshausen, T., Burggräf, P., Hassenzahl, M., & Wagner, J. (2023). Human trust in otherware – a systematic literature review bringing all antecedents together. *Ergonomics*, *66*(7), 976–998. <https://doi.org/10.1080/00140139.2022.2120634>
- Saunders, M. N., Dietz, G., & Thornhill, A. (2014). Trust and distrust: Polar opposites, or independent but co-existing? *Human Relations*, *67*(6), 639–665.
<https://doi.org/10.1177/0018726713500831>
- Schaefer, K. E. (2016). Measuring Trust in Human Robot Interactions: Development of the "Trust Perception Scale-HRI". In R. Mittu, D. Sofge, A. Wagner, & W. Lawless (Eds.), *Robust intelligence and trust in autonomous systems* (pp. 191–218). Springer US. https://doi.org/10.1007/978-1-4899-7668-0_10
- Scharowski, N., Perrig, S. A. C., Aeschbach, L. F., von Felten, N., Opwis, K., Wintersberger, P., & Brühlmann, F. (2023). *To trust or distrust trust measures: Validating questionnaires for trust in AI* [Manuscript submitted for publication].
- Scharowski, N., Benk, M., Kühne, S. J., Wettstein, L., & Brühlmann, F. (2023). Certification labels for trustworthy AI: Insights from an empirical mixed-method study. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 248–260. <https://doi.org/10.1145/3593013.3593994>
- Scharowski, N., & Opwis, F., Klaus Brühlmann. (2021). Initial evidence for biased decision-making despite human-centered AI explanations. *CHI 2021 Workshop:*

Operationalizing Human-Centered Perspectives in Explainable AI.

<https://osf.io/preprints/osf/5jzmb>

Scharowski, N., & Perrig, S. A. C. (2023). Distrust in (X)AI – Measurement artifact or distinct construct? *CHI 2023 TRAIT Workshop on Trust and Reliance in AI-Human Teams*. <https://doi.org/10.48550/arXiv.2303.16495>

Scharowski, N., Perrig, S. A. C., Svab, M., Opwis, K., & Brühlmann, F. (2023).

Exploring the effects of human-centered AI explanations on trust and reliance.

Frontiers in Computer Science, 5. <https://doi.org/10.3389/fcomp.2023.1151150>

Scharowski, N., Perrig, S. A. C., von Felten, N., & Brühlmann, F. (2022). Trust and reliance in XAI – Distinguishing between attitudinal and behavioral measures.

CHI 2022 TRAIT Workshop. <https://doi.org/10.48550/arXiv.2203.12318>

Schemmer, M., Kuehl, N., Benz, C., Bartos, A., & Satzger, G. (2023). Appropriate reliance on AI advice: Conceptualization and the effect of explanations.

Proceedings of the 28th International Conference on Intelligent User Interfaces, 410–422. <https://doi.org/10.1145/3581641.3584066>

Schlicker, N., Uhde, A., Baum, K., Hirsch, M. C., & Langer, M. (2022). A micro and macro perspective on trustworthiness: Theoretical underpinnings of the trustworthiness assessment model (TrAM).

<https://doi.org/10.31234/osf.io/qhwvx>

Seckler, M., Heinz, S., Forde, S., Tuch, A. N., & Opwis, K. (2015). Trust and distrust on the web: User experiences and website characteristics. *Computers in human behavior*, 45, 39–50. <https://doi.org/10.1016/j.chb.2014.11.064>

Seifert, C., Scherzinger, S., & Wiese, L. (2019). Towards generating consumer labels for machine learning models. *2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI)*, 173–179.

<https://doi.org/10.1109/CogMI48466.2019.00033>

Selwyn, N., & Gallo Cordoba, B. (2022). Australian public understandings of artificial intelligence. *AI & SOCIETY*, 37(4), 1645–1662.

<https://doi.org/10.1007/s00146-021-01268-z>

- Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human-Computer Interaction*, 36(6), 495–504. <https://doi.org/10.1080/10447318.2020.1741118>
- Shoemaker, P., Tankard, J. W., & Lasorsa, D. L. (2004). *How to build social science theories*. SAGE Publications, Inc. <https://doi.org/10.4135/9781412990110>
- Simmel, G. (1908). Das Geheimnis und die geheime Gesellschaft. In *Soziologie. Untersuchungen über die Formen der Vergesellschaftung*. Duncker & Humblot Verlag.
- Sitkin, S. B., & Roth, N. L. (1993). Explaining the limited effectiveness of legalistic “remedies” for trust/distrust. *Organization science*, 4(3), 367–392. <https://doi.org/10.1287/orsc.4.3.367>
- Slife, B. D., Wright, C. D., & Yanchar, S. C. (2016). Using operational definitions in research: A best-practices approach. *The Journal of Mind and Behavior*, 37(2), 119–139. Retrieved April 28, 2024, from <http://www.jstor.org/stable/44631540>
- Spain, R. D., Bustamante, E. A., & Bliss, J. P. (2008). Towards an empirically developed scale for system trust: Take two. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 52(19), 1335–1339. <https://doi.org/10.1177/154193120805201907>
- Speith, T. (2022). A review of taxonomies of explainable artificial intelligence (XAI) methods. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2239–2250. <https://doi.org/10.1145/3531146.3534639>
- Starke, G., Schmidt, B., De Clercq, E., & Elger, B. S. (2023). Explainability as fig leaf? An exploration of experts’ ethical expectations towards machine learning in psychiatry. *AI and Ethics*, 3(1), 303–314. <https://doi.org/10.1007/s43681-022-00177-1>
- Stephanidis, C., Salvendy, G., Antona, M., Chen, J. Y. C., Dong, J., Duffy, V. G., Fang, X., Fidopiastis, C., Fragomeni, G., Fu, L. P., Guo, Y., Harris, D., Ioannou, A., Jeong, K.-a., Konomi, S., Krömker, H., Kurosu, M., Lewis, J. R., Marcus, A., . . . Zhou, J. (2019). Seven HCI grand challenges. *International*

Journal of Human-Computer Interaction, 35(14), 1229–1269.

<https://doi.org/10.1080/10447318.2019.1619259>

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>

Suresh, H., Gomez, S. R., Nam, K. K., & Satyanarayan, A. (2021). Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. <https://doi.org/10.1145/3411764.3445088>

The White House. (2022). *Blueprint for an AI Bill of Rights* [Accessed: 2024-03-27].

<https://www.whitehouse.gov/ostp/ai-bill-of-rights/>

Thornton, L., Knowles, B., & Blair, G. (2021). Fifty shades of grey: In praise of a nuanced approach towards trustworthy design. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 64–76.

<https://doi.org/10.1145/3442188.3445871>

Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C. G., & van Moorsel, A. (2020). The relationship between trust in AI and trustworthy machine learning technologies. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 272–283. <https://doi.org/10.1145/3351095.3372834>

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481), 453–458. <https://doi.org/10.1126/science.7455683>

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, 106(4), 1039–1061. <https://doi.org/10.2307/2937956>

Ueno, T., Sawa, Y., Kim, Y., Urakami, J., Oura, H., & Seaborn, K. (2022). Trust in human-AI interaction: Scoping out models, measures, and methods. *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*. <https://doi.org/10.1145/3491101.3519772>

- Vereschak, O., Bailly, G., & Caramiaux, B. (2021). How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 1–39.
<https://doi.org/10.1145/3476068>
- Vorm, E. S., & Combs, D. J. Y. (2022). Integrating transparency, trust, and acceptance: The intelligent systems technology acceptance model (ISTAM). *International Journal of Human-Computer Interaction*, 38(18-20), 1828–1845.
<https://doi.org/10.1080/10447318.2022.2070107>
- Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing theory-driven user-centric explainable AI. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15. <https://doi.org/10.1145/3290605.3300831>
- Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12(3), 129–140.
<https://doi.org/10.1080/17470216008416717>
- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of personality and social psychology*, 54(6), 1063–1070.
<https://doi.org/10.1037/0022-3514.54.6.1063>
- Wintersberger, P., Riener, A., Schartmüller, C., Frison, A.-K., & Weigl, K. (2018). Let me finish before I take over: Towards attention aware device integration in highly automated vehicles. *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 53–65.
<https://doi.org/10.1145/3239060.3239085>
- Wischniewski, M., Krämer, N., & Müller, E. (2023). Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. <https://doi.org/10.1145/3544548.3581197>
- Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. *Proceedings of the 2019 CHI*

Conference on Human Factors in Computing Systems, 1–12.

<https://doi.org/10.1145/3290605.3300509>

Yu, K., Berkovsky, S., Taib, R., Conway, D., Zhou, J., & Chen, F. (2017). User trust dynamics: An investigation driven by differences in system performance.

Proceedings of the 22nd international conference on intelligent user interfaces, 307–317. <https://doi.org/10.1145/3025171.3025219>

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

2022 ACM Conference on Fairness, Accountability, and Transparency, 535–563. <https://doi.org/10.1145/3531146.3533118>

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making.

Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 295–305. <https://doi.org/10.1145/3351095.3372852>

Acknowledgments

Without trying to sound insincerely modest, I believe that people are mainly the products of random chance with corresponding consequences for their fortune and misfortune. Nevertheless, or perhaps because of this, I would like to thank all the people who have contributed to my personal fortune and without whom this dissertation would not have been possible.

First, I would like to thank Klaus Opwis, my PhD supervisor, for the opportunity to work with him and for his continuous support, trust, and encouragement during these years. I would also like to thank Philipp Wintersberger for volunteering as my second supervisor and my co-authors, Michaela Benk, Swen Kühne, Nadine Schlicker, Florian Spiess, and Heiko Schuldt.

Many thanks for the fantastic years go to my colleagues at the Center for General Psychology and Methodology, Sebastian Perrig, Lena Aeschbach, Antony de Castro Hüsler, Ariane Haller, Beat Vollenwyder, Lorena Weder, Nick von Felten, Melanie Svab and Zgjim Memeti. Special thanks to Léane Wettstein, my research assistant, without whom I would not have been able to realize some of these research projects. Also, special thanks to Florian Brühlmann, the former head of the Human-Computer Interaction Research Group, for his support in my research projects, and without, I would have hesitated to pursue a PhD.

Most of all, I would like to thank my friends, especially Dominik Schumacher, Loris Jeitziner, Katja Rutz, Zakir Hussain, Florence Schumacher, and Jonas Schaffter, who have all supported me in their own unique way during my PhD.

Finally, heartfelt appreciation goes to my family, particularly Anita Scharowski, Markus Scharowski, Samuel Scharowski, and Hildegard Mohr to whom I owe the most.

Thank you for everything.

Appendix

1. **Scharowski, N.**, Perrig, S. A. C., Svab, M., Opwis, K., & Brühlmann, F. (2023). Exploring the effects of human-centered AI explanations on trust and reliance. *Frontiers in Computer Science*, 5. <https://doi.org/10.3389/fcomp.2023.1151150>
2. **Scharowski, N.**, Perrig, S. A. C., Aeschbach, L. F., von Felten, N., Opwis, K., Wintersberger, P., & Brühlmann, F. (2023). To trust or distrust trust measures: Validating questionnaires for trust in AI. *Manuscript submitted for publication*.
3. **Scharowski, N.**, Benk, M., Kühne, S. J., Wettstein, L., & Brühlmann, F. (2023). 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). New York, NY, USA: ACM. <https://doi.org/10.1145/3593013.3593994>



OPEN ACCESS

EDITED BY

Chathurika S. Wickramasinghe Brahmam,
Capital One, United States

REVIEWED BY

Martin Gjoreski,
University of Italian Switzerland, Switzerland
Antonis Bikakis,
University College London, United Kingdom
Nadine Schlicker,
University of Marburg, Germany

*CORRESPONDENCE

Nicolas Scharowski
✉ nicolas.scharowski@unibas.ch

RECEIVED 25 January 2023

ACCEPTED 27 June 2023

PUBLISHED 17 July 2023

CITATION

Scharowski N, Perrig SAC, Svab M, Opwis K and
Brühlmann F (2023) Exploring the effects of
human-centered AI explanations on trust and
reliance. *Front. Comput. Sci.* 5:1151150.
doi: 10.3389/fcomp.2023.1151150

COPYRIGHT

© 2023 Scharowski, Perrig, Svab, Opwis and
Brühlmann. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other forums is
permitted, provided the original author(s) and
the copyright owner(s) are credited and that
the original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with these
terms.

Exploring the effects of human-centered AI explanations on trust and reliance

Nicolas Scharowski*, Sebastian A. C. Perrig, Melanie Svab,
Klaus Opwis and Florian Brühlmann

Human-Computer Interaction Research Group, Center for General Psychology and Methodology,
Faculty of Psychology, University of Basel, Basel, Switzerland

Transparency is widely regarded as crucial for the responsible real-world deployment of artificial intelligence (AI) and is considered an essential prerequisite to establishing trust in AI. There are several approaches to enabling transparency, with one promising attempt being human-centered explanations. However, there is little research into the effectiveness of human-centered explanations on end-users' trust. What complicates the comparison of existing empirical work is that trust is measured in different ways. Some researchers measure subjective trust using questionnaires, while others measure objective trust-related behavior such as reliance. To bridge these gaps, we investigated the effects of two promising human-centered *post-hoc* explanations, *feature importance* and *counterfactuals*, on trust and reliance. We compared these two explanations with a control condition in a decision-making experiment ($N = 380$). Results showed that human-centered explanations can significantly increase reliance but the type of decision-making (increasing a price vs. decreasing a price) had an even greater influence. This challenges the presumed importance of transparency over other factors in human decision-making involving AI, such as potential heuristics and biases. We conclude that trust does not necessarily equate to reliance and emphasize the importance of appropriate, validated, and agreed-upon metrics to design and evaluate human-centered AI.

KEYWORDS

AI, XAI, HCXAI, trust, reliance, transparency, explainability, interpretability

1. Introduction

It is generally recognized that computers perform specific tasks better than humans, such as numeracy, logical reasoning, or storing information (Solso et al., 2005). But with the recent breakthroughs in artificial intelligence (AI), domains that used to be exclusively associated with human competence and considered computationally unattainable are likewise being challenged by machines. AI has led to improvements in speech recognition, image classification, as well as object detection (LeCun et al., 2015) and is now increasingly used in various everyday applications such as video surveillance, email spam filtering, online customer support, and product recommendations. Because of this general applicability and the potential manifold consequences, voices are being raised that AI should satisfy criteria like fairness, reliability, accountability, and transparency (Ehsan et al., 2021b; ACM FAccT Conference, 2022). The call for transparent AI has led to the multidisciplinary research field of explainable artificial intelligence (XAI), which explores methods and models that make the behaviors, predictions, and decisions of AI transparent and understandable to humans (Lipton, 2018a; Liao et al., 2020). Abdul et al. (2018), as well as Biran and Cotton (2017) have

pointed out that the development of transparent systems has long been a research focus, originating from expert systems, intelligent agents, recommender systems, context-aware systems, and other adjacent fields such as automation.

Despite this rich history, current XAI research faces unprecedented challenges as AI is increasingly complex and thus more cumbersome to render transparent (Biran and Cotton, 2017). In the pursuit of ever more accurate predictions, modern AI consists of millions of interdependent values and parameters, resulting in a trade-off between complexity and transparency (Shmueli, 2010; Mittelstadt et al., 2019). Because of this complexity, AI is often characterized by the opaque box paradigm (Suresh et al., 2021), meaning that AI can only be considered in terms of its inputs and outputs without direct observations of its inner workings. This opacity makes it more challenging than ever to ensure fairness, reliability, and accountability, rendering transparency a prerequisite for the other three criteria. Some researchers argue that transparency helps verify and improve the functionality of a system (i.e., for debugging), supports developers in learning from a system (i.e., in generating hypotheses), or is needed to ensure fair and ethical decision-making (Mittelstadt et al., 2019). Others believe that transparency contributes toward building a relationship of trust between humans and AI (Stephanidis et al., 2019), which plays a key role in people's willingness to rely on automated systems (Hoff and Bashir, 2015).

While transparency is generally considered crucial for the effective and responsible real-world deployment of AI, there are various transparency approaches tailored to the algorithm's goal, the context, and the involved stakeholders, such as developers, decision-makers, and end-users (Ehsan et al., 2019; Samek et al., 2019). For end-users, the requirements and purpose of transparency are expected to be distinct (Cheng et al., 2019; Langer et al., 2021; Suresh et al., 2021), and Miller (2019) specified criteria that should be taken into account in order to achieve human-centered explainable AI (HCXAI). Following the notion of Ehsan and Riedl (2020), we understand HCXAI as an approach that puts humans at the center of technology design. Within this framework, not only is it important to conduct user studies that validate XAI methods with ordinary end-users, but also to consider explanations designed to account for human needs. We argue that Miller (2019)'s criteria and the focus on how humans explain decisions to one another are a good starting point for meaningful AI explanations to end-users. However, empirical investigations of the effects of human-centered explanations satisfying these criteria are sparse, and there is mixed evidence about whether transparency is in fact increasing trust (Cramer et al., 2008; Nothdurft et al., 2013; Cheng et al., 2019; Ehsan et al., 2019; Zhang et al., 2020; Poursabzi-Sangdeh et al., 2021). These ambiguous findings may arise from the use of proxy-tasks rather than actual decision-making tasks when evaluating AI systems (Buçinca et al., 2020) and from varying conceptualizations of trust. Studies on XAI appear to define and measure trust differently (Vereschak et al., 2021). Some researchers assess attitudinal trust measures via questionnaires (Buçinca et al., 2020), while others focus on trust-related behavior such as reliance (Poursabzi-Sangdeh et al., 2021). However, research has shown that subjective trust can be a poor predictor of actual reliance (Dzindolet et al., 2003; Miller et al., 2016; Papenmeier et al., 2022). Therefore, it

seems particularly important to distinguish between attitudinal and behavioral measures when studying the effect of transparency on trust (Parasuraman and Manzey, 2010; Sanneman and Shah, 2022; Scharowski et al., 2022).

In this study, we focus on explainability as a means of AI transparency. Explainability, in this context, is the process of explaining how an opaque box AI arrived at a particular result or decision after a computation has been performed (i.e., *post-hoc* explanations), without directly revealing the AI's internal mechanisms via visualizations or graphical interfaces, as typically aimed for in clear box AI. Grounded in Miller's work, we identified *feature importance explanations* and *counterfactual explanations* as two promising *post-hoc* explanations for achieving HCXAI. We conducted an online decision-making experiment ($N = 380$) on Amazon Mechanical Turk (MTurk) to investigate the effect of those two human-centered explanations on end-users' trust and reliance with a control condition. The *Trust between People and Automation Scale* (TPA, Jian et al., 2000) served as an attitudinal measure of AI trust. Reliance on the AI recommendation, captured by *weight of advice*, provided a measure for trust-related behavior. The results suggest that the relationship between transparency and reliance is more nuanced than commonly assumed and emphasize the importance of adequately differentiating between trust and reliance and their respective measurements when evaluating XAI. While transparency did not affect trust, reliance increased through human-centered *post-hoc* explanations, but only for specific decision-making tasks. In the particular context we examined, it appears that the type of decision-making participants were facing (increasing a price vs. decreasing a price) had a greater influence on reliance than how the AI explained its recommendation to the end-users. This suggests that humans display cognitive biases and apply heuristics in decision-making tasks that involve AI recommendations. If biased human decision-making prevails, AI may not support people to reach better decisions. The XAI community should consider potential biases and heuristics for a more nuanced understanding of the human-AI interaction. It remains to be further explored whether measuring attitudinal trust via questionnaires reflects trust-related behavior (i.e., reliance) appropriately and whether heuristics and biases also have an impact on trust. If researchers and practitioners who develop and evaluate AI systems assess only subjective trust, they may not draw valid conclusions about actual AI reliance and vice versa. Given that AI is increasingly utilized to make critical decisions with far-reaching consequences, adopting agreed-upon, validated, and appropriate measurements in XAI is of paramount importance.

2. Related work

2.1. Human-centered explanations

Two closely related terms that are often used interchangeably should be distinguished when referring to AI transparency: explainability and interpretability. While both terms refer to methods for achieving transparency, they differ in their approach to implementing transparency. For Lipton (2018b), interpretability is the information that a system provides about its inner workings

and associated with the notion of *clear box* AI, meaning AI whose internal mechanisms are accessible and not concealed. Interpretability is thus achieved by using or designing AI in a way that its decision-making can be directly observed or otherwise visualized. Explainability, on the other hand, implies accepting *opaque box* AI whose internal mechanisms are not readily accessible or understandable, and providing meaningful information by explaining how a specific output or decision was reached after a computation has been carried out. In this sense, explainability is *post-hoc* interpretability (Lipton, 2018b; Ehsan et al., 2019; Miller, 2019; Mohseni et al., 2020).

In addition to this distinction between explainability and interpretability, XAI researchers need to be aware of the varying needs and goals different stakeholders have when interpreting, understanding, and reacting to explanations coming from AI (Suresh et al., 2021). Past research has raised concerns that AI explanations are frequently based on the intuition of researchers, AI developers, and experts rather than addressing the needs of end-users (Du et al., 2019; Miller, 2019). A growing body of work has engaged with this challenge (Ferreira and Monteiro, 2020; Hong et al., 2020; Liao et al., 2020; Ehsan et al., 2021a) and now focuses on more human-centered approaches that align AI explanations with people's needs. Despite these considerations regarding human-centered explanations, previous work on AI transparency has often placed a greater emphasis on interpretability (i.e., model visualization for clear box AI) than on explainability (i.e., *post-hoc* explanations for opaque box AI) (Kulesza et al., 2015; Krause et al., 2016; Cheng et al., 2019; Kocielnik et al., 2019; Lai and Tan, 2019; Poursabzi-Sangdeh et al., 2021). This emphasis has led to focusing on graphical interfaces that allow users to observe and understand the decision-making processes of these models more directly. While *post-hoc* explanations also require some sort of user interface or visualization, they operate at a more abstract level and provide a simplified or approximate representation of the decision-making process rather than direct access to the internal workings of the model, as interpretability seeks to accomplish. However, some researchers have questioned that interpretability approaches are useful to all people equally. Suresh et al. (2021) and Lipton (2018b) argue that explainability might be more reflective of the way that humans are transparent about their own decisions. When it comes to humans, the exact processes by which our brains form decisions and our explanations regarding those decisions are distinct (Lipton, 2018b). Similar to how people provide explanations to one another, AI might explain its decisions without disclosing the computation underlying them. Because of its proximity to how humans reason about their decisions, explainability seems promising to achieve HCXAI if the way humans provide and understand explanations is taken into account.

With regard to human-centered explanations, researchers have emphasized the importance of incorporating insights from philosophy, the social sciences, and psychology on how people define, generate, select, evaluate, and present explanations into the field of XAI (Miller, 2019; Mittelstadt et al., 2019). Based on findings from these areas of research, Miller (2019) defined certain criteria for what contributes to a meaningful explanation for people, including *selectivity* (providing the most important reasons for a decision), *contrastivity* (providing contrastive information

with a decision), and *sociality* (explaining something in a similar way to how humans explain their actions). Miller (2019) and Mittelstadt et al. (2019) argued that explanations from AI should at least fulfill some of these criteria to be meaningful for end-users. Adadi and Berrada (2018) identified over 17 different transparency approaches that are being proposed in the current XAI literature. Based on Miller (2019)'s criteria, we narrowed down Adadi and Berrada (2018)'s selection and identified two promising human-centered *post-hoc* explanations: *feature importance explanations* and *counterfactual explanations*.

Feature importance explanations. Humans rarely expect a complete explanation for a decision and often select the most important or immediate cause from a sometimes infinite number of reasons (Miller, 2019). As the name suggests, feature importance allows end-users to determine which features are most important for an AI's output. Such explanations thus satisfy the selectivity criterion proposed by Miller (2019) because they show how certain factors influenced a decision. Feature importance explanations have the following notation: "Outcome P was returned because variable V had values (vi, vii, ...) associated with it" (Wachter et al., 2018, p. 9).

Counterfactual explanations. Humans usually ask why a particular decision was made instead of another one (Miller, 2019). In addition to the leading causes of an output, counterfactuals provide contrastive "what-if" statements that help identify what might be changed in the future to achieve a desired output (Mothilal et al., 2020). Counterfactuals combine Miller's selectivity and contrastivity criteria. They are expected to have psychological benefits because they help people act, rather than merely understand, by altering future behavior to achieve a desired outcome (Wachter et al., 2018; Mothilal et al., 2020). Counterfactuals commonly have the following notation: "Outcome P was returned because variable V had values (vi, vii, ...) associated with it. If V had values (vi', vii', ...) instead, outcome P' would have been returned" (Wachter et al., 2018, p.9).

Both explanations also seem to meet Miller (2019)'s sociality criterion. For humans, explanations are a form of social interaction or, more specifically, a transfer of knowledge often presented as part of a conversation between the explainer and the explainee that is subject to the rules of conversation (Hilton, 1990; Miller, 2019). Although Miller (2019) points out that this does not imply that explanations must be given in natural language, we expect natural language explanations to be a promising approach for human-centered explanations because they are accessible and intuitive to humans (Ehsan et al., 2019). De Graaf and Malle (2018) argued that because people attribute human-like traits to artificial agents, they might expect them to provide explanations similar to how humans explain their actions. Szymanski et al. (2021) showed that while end-users prefer visual over textual explanations, they performed significantly worse with the former, and Kizilcec (2016) demonstrated that short textual explanations build subjective trust in an algorithm's decision. There are also jurisdictional reasons for explanations in natural language. They comply with the EU's GDPR (Wachter et al., 2018) and align with the regulatory requirement for automated decision-making to explain decisions in an "easily accessible form, using

clear and plain language [...] provided in writing.” (European Parliament and Council of the European Union, 2016, article 12). To the best of our knowledge, there is little to no empirical research on the effectiveness of these two human-centered explanations derived from the literature (i.e., Adadi and Berrada, 2018) using Miller’s criteria in fostering end-users’ trust in AI. Therefore, an empirical investigation into the efficacy of *feature importance explanations* and *counterfactual explanations* seems warranted.

2.2. Trust in XAI

Within the XAI community, researchers define and measure trust in different ways, and there does not appear to be a clear consensus about the desired effect of trust or a clear differentiation of the factors that contribute to trust (Chopra and Wallace, 2003; Mohseni et al., 2020). To provide two examples: Lai and Tan (2019) proposed a spectrum between full human agency and full automation, with varying levels of explanations along this spectrum. In a deception detection task (asking end-users to decide whether a hotel review is genuine or deceptive), they illustrated that heatmaps of relevant instances and example-based explanations improved human performance and increased the trust humans place on the predictions of the AI. Lai and Tan defined trust as the percentage of instances for which humans relied on the machine prediction. In contrast, Cheng et al. (2019) conducted an experiment where participants used different UI interfaces to comprehend an algorithm’s decision for university admissions. They showed that revealing the inner workings of an algorithm can improve users’ comprehension and found that users’ subjective trust, assessed by a 7-point Likert scale, was not affected by the explanation interface. These two empirical studies exemplify how trust is measured differently in XAI research. These discrepancies could be a reason for the inconclusive findings in current XAI literature regarding the effect of transparency on trust. This warrants a more precise definition and rigorous distinction between trust and related concepts, such as reliance, in empirical studies investigating the relationship between transparency and trust.

The differentiation between subjective and objective trust and their measurement in XAI was addressed by Mohseni et al. (2020). They pointed out that subjective trust measures include interviews, surveys, and self-reports via questionnaires, which according to Bućinca et al. (2020) have been the focal points for evaluating AI transparency. For objective measures of trust, Mohseni et al. (2020) proposed users’ perceived system competence, understanding, and users’ reliance on a system. This distinction between trust and reliance was emphasized by Hartmann (2020). They argued that the everyday use of the word *trust* is misleading when applied to technology and that, in this case, trust must be differentiated from *reliance*. Hartmann (2020) was not the only one that distinguished between trust and reliance as other researchers have shown that attitudinal judgments have an impact on people’s intention to rely on automated systems (Cramer et al., 2008; Merritt, 2011); People

tend to rely on automation they trust and reject automation they distrust (Lee and See, 2004). This makes trust particularly relevant in the misuse (overreliance Parasuraman and Riley, 1997) and disuse (neglect or underutilization Parasuraman and Riley, 1997) of automation (Hoff and Bashir, 2015; Yu et al., 2017; Stephanidis et al., 2019). To avoid such instances, users’ trust needs to be calibrated or warranted. Trust calibration refers to the extent to which the trust that users place in the system is adequate to the system’s actual capabilities (Wischniewski et al., 2023). Fostering end users’ trust in AI should aim to attain an appropriate level of trust to avoid overreliance or underutilization of AI systems. According to Lee and See (2004) and correspondent with Hoff and Bashir (2015), we define trust in the context of AI as “the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability” (Lee and See, 2004, p. 6). This definition encapsulates the notion of uncertainty and vulnerability as proposed by Jacovi et al. (2021) and Mayer et al. (1995), which is the most widely used and accepted definition of trust (Rousseau et al., 1998). Adopting Lee and See’s definition and model, we distinguish between trust and reliance and think of trust as an *attitude* and reliance as a *behavior* that follows the level of trust. In their work on metrics for XAI, Hoffman et al. (2019) make a similar distinction when they differentiate between trusting a machine’s output and following its advice. In this framework, attitudes and behaviors remain conceptually distinct and do not share a deterministic but a probabilistic relationship (Ajzen and Fishbein, 1980; Körber, 2018). Even if an AI system is trusted, reliance must not necessarily follow (Kirlik, 1993; Körber, 2018), and people may claim to trust an AI system, yet behave in a way that suggests they do not (Miller et al., 2016). This implies that attitudes may not always translate into behaviors. The empirical findings of Dzindolet et al. (2003) support this argument: although some automated decision aids were rated as more trustworthy than others, all were equally likely to be relied upon.

Given these possible contradictions, we think it is useful to conceptualize trust as an antecedent to reliance that guides but does not determine it (Lee and See, 2004). However, this consideration has been insufficiently taken into account in past research (Papenmeier et al., 2022; Scharowski et al., 2022). A rigorous distinction and an accurate conceptualization of trust and reliance are vital for empirical XAI studies since researchers who evaluate AI systems using only subjective measures of AI trust might not draw valid conclusions about actual reliance on AI and vice versa. Arrieta et al. (2020) emphasized that only agreed-upon metrics and their respective measurements allow for meaningful comparisons in XAI research and that without such consistency, any claims are elusive and do not provide a solid foundation. For this reason, we decided to investigate two alternative methodological approaches, namely, measuring attitudinal trust on the one hand and measuring trust-related behavior in terms of reliance on the other hand. This approach is in line with Sanneman and Shah (2022), that recommended using trust scales in conjunction with behavior-based metrics to determine if people appropriately trust *and* use AI systems in response to AI explanations they provide.

3. Empirical investigation

3.1. Research question and hypotheses

We investigated the following research question:

RQ: What effect do human-centered explanations have on end-users' trust and reliance?

To answer this research question, we compared the previously introduced feature importance and counterfactual *post-hoc* explanations with a control in a scenario in which participants had to estimate subleasing prices for different apartments. We employed a mixed study design with a 3 (explanation condition: feature importance vs. counterfactual vs. control) \times 2 (type of AI recommendation: increasing price vs. decreasing price). Explanation condition was the between-subject factor, type of recommendation was the within-subject factor. Following Poursabzi-Sangdeh et al. (2021), we focused on the domain of real estate valuation, where machine learning is often used to predict apartment prices. Airbnb (<https://airbnb.com>) and Zillow (<http://zillow.com>) are examples of websites that provide price recommendations to end-users in this way. Considering the previous clarifications, we expected that trust and reliance are influenced by human-centered explanations similarly but should be treated as distinct concepts. We, therefore, formulated separate hypotheses for both trust and reliance. We further presumed that feature importance and counterfactual explanations lead to more trust *and* reliance in participants compared to a control condition where no additional explanation was present. Counterfactuals are both selective and contrastive, while feature importance explanations are just selective (Miller, 2019). This makes counterfactuals an even more promising type of human-centered explanation compared to feature importance explanations.

For these reasons, the specific hypotheses were:

H₁ The experimental condition *feature importance* will lead to higher reliance compared to the control.

H₂: The experimental condition *counterfactuals* will lead to higher reliance compared to the control.

H₃: *Counterfactuals* will lead to higher reliance compared to *feature importance*.

H₄: The experimental condition *feature importance* will lead to higher trust compared to the control.

H₅: The experimental condition *counterfactuals* will lead to higher trust compared to the control.

H₆: *Counterfactuals* will lead to higher trust compared to *feature importance*.

4. Method

4.1. Measures

The independent variable was *condition*, with the two levels feature importance and counterfactuals, as well as a third level without explanations, which served as a control.

We used two measures as dependent variables to account for the aforementioned distinction between trust as an attitude and reliance as trust-related behavior. On the one hand, we wanted to determine if people relied on the AI and changed their behavior after being presented with an explanation. This behavior change was captured by the parameter *Weight of Advice* (WOA), which stems from the literature on taking advice (Harvey and Fischer, 1997). WOA has the following notation:

$$\text{WOA} = \frac{T_2 - T_1}{R - T_1} \quad (1)$$

In Equation (1), R is defined as the model's recommendation, T_1 is the participant's initial estimate of the apartment's price before seeing R , and T_2 is the participant's final estimate of the apartment's price after seeing R . WOA measures the degree to which people change their behavior and move their initial estimate toward the advice. WOA is equal to 1 if the participant's final prediction matches the AI recommendation and equal to 0.5 if they average their initial prediction with the AI recommendation. A WOA of 0 occurs when a participant ignores the AI recommendation ($T_1 = T_2$), and a negative WOA signifies that a participant discounted the recommendation completely and moved further away from the recommendation. WOA can be viewed as a percentage of how much people weigh the received advice (i.e., the AI recommendation), and this straightforward interpretation is an advantage of this reliance measurement. While WOA has been used in the past by researchers in XAI as an alternative trust measurement (Logg et al., 2019; Mucha et al., 2021; Poursabzi-Sangdeh et al., 2021), it has never been explicitly referred to as reliance and thus clearly differentiated from trust to the best of our knowledge.

On the other hand, we chose the TPA (Jian et al., 2000) to measure trust because the scale's underlying definition of trust is compatible with the one we adopted from Lee and See (2004). Furthermore, the TPA is an established measure in HCI (Hoffman et al., 2019). Several other scales evaluating AI have adapted items from the TPA (e.g., Hoffman et al., 2019), and its psychometric quality has been evaluated multiple times (Spain et al., 2008; Gutzwiller et al., 2019). Jian et al. (2000) treated trust and distrust as opposite factors extending along a single dimension of trust. The scale is a seven-point Likert-type scale (ranging from 1: "not at all" to 7: "extremely") and consists of 12 items. Five items for distrust (i.e., "The system is deceptive.", "The system behaves in an underhanded manner.", "I am suspicious of the system's intent, action or, outputs.", "I am wary of the system.", "The system's actions will have a harmful or injurious outcome."). The seven remaining items for trust (i.e., "I am confident in the system.", "The system provides security.", "The system has integrity.", "The system is dependable.", "The system is reliable.", "I can trust the system.", "I am familiar with the system."). We used the scale in its original form, except for prefixing the word "AI" to the word "system," e.g., "I have confidence in the AI system."

4.2. Experiment

We carried out a one-factor between-subjects design online experiment¹ over Amazon Mechanical Turk (MTurk, <http://mturk.com>). The experiment was implemented through the online survey tool Limesurvey (<http://limesurvey.org>).

4.2.1. Participants

A total of 913 participants were initially recruited over MTurk, and 798 of them fully completed the survey. Only workers from the USA with a human-intelligence-task (HIT) approval of 95% and at least 100 approved HITs were allowed to participate in the experiment. Workers who completed the task conscientiously were reimbursed with 1.50 US dollars and a bonus of 0.30 US dollars for their participation. Several criteria were applied during data cleaning to ensure data quality. Participants who failed to provide a correct answer ($n = 36$) for the bogus item (“This is an attention check. Please choose 7 here”) or for one of three control questions (“In this survey, you had to tell us for how much money you would sell a house to a company”; “In this survey, we asked you to indicate in which U.S. state you currently live”; “In this survey, you got price recommendations from a good friend”) were removed ($n = 310$). We also excluded participants that showed unrealistic WOAs ($n = 72$). Following prior research (Gino and Moore, 2007; Logg et al., 2019), we defined unrealistic WOA as being ≤ -1 and ≥ 2 . For the data analysis, 380 participants remained. The sample was predominantly male (61%) and had an average age of 37 years ($M = 37.03$, $SD = 10.15$, $min = 18$, $max = 69$). A majority of the participants (68%) possessed a higher-educational degree (i.e., a bachelor’s degree, master’s degree, or PhD).

4.2.2. Procedure and task

After providing informed consent, participants were introduced to the study and their task. They were asked to imagine a scenario where their goal was to sublease six different apartments on a subleasing website. Based on the apartment’s features and amenities (e.g., number of bedrooms, distance to public transit), they had to estimate an initial subleasing price ($T1$). After estimating $T1$, an alleged AI from the website provided a computed price recommendation (R). In reality, the price recommendation was based on an algorithm introduced as an AI. Participants were informed that they would be more likely to find a sublesser by deciding on a lower price but would consequently receive less profit. If they decided on a higher price, they would be less likely to find a sublesser but potentially receive more profit. They were told that the AI’s goal was to help them find the optimal price to successfully find a sublesser with a reasonable profit. How exactly this price recommendation was calculated by the algorithm will be discussed in the next section. Figure 1 shows how the price recommendation and respective explanations were presented to the participants. A list of all the explanations used for each type

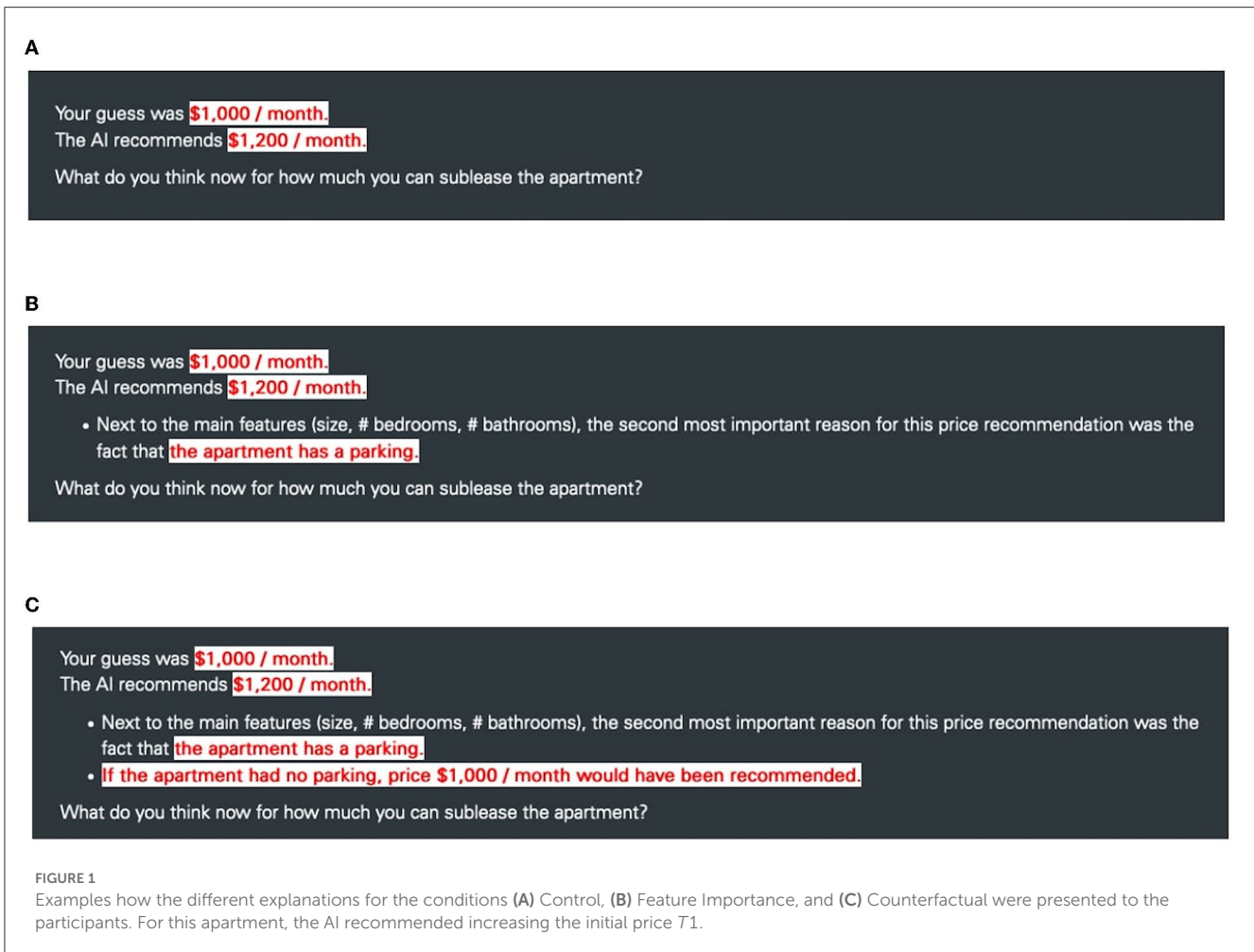
of recommendation can be found in the online repository¹. The output was designed to appear as if the subsequent explanation was an extension of the preceding ones. The most relevant outputs were presented as a console output in order to make the stimuli more convincing (see Figure 1). After seeing the AI recommendation, participants could decide if they desired to approach it or not and settled on a final subleasing price ($T2$). This deliberate choice by the participants to either rely on the AI recommendation or not makes our experiment an actual decision-making task rather than a proxy task (Bućinca et al., 2020). In proxy-tasks, the focus lies on how well participants can simulate the model’s decisions (Bućinca et al., 2020). In actual decision-making tasks, people’s choices involve systematic thinking errors (biases) and mental shortcuts (heuristics) (Tversky and Kahneman, 1974), as it is up to the participants to decide whether and how to use the AI (i.e., reliance on the AI).

To evoke a certain degree of uncertainty, participants were told that they would be reimbursed based on their performance. Uncertainty is a defining characteristic of trust (Mayer et al., 1995) and has been referred to as a necessary prerequisite of human trust in AI that has been lacking in current XAI studies (Jacovi et al., 2021). Participants were informed that for good estimations, the top 10% would be paid an additional bonus of 0.30 US dollars. In actuality, every participant received the bonus, regardless of their performance. In order to better control for the price disparity between urban and rural regions, participants were asked to indicate what US state they are currently living in (e.g., Colorado) to ascertain their state capital (e.g., Denver). The objective was to make the estimate easier for the participants and to make the AI more persuasive since it was claimed that the AI would likewise base its price recommendations on data collected in that state capital. After an example that showed how the apartments and their amenities would be presented to them, participants could start with the actual task. Once the task was completed, participants had to fill out the TPA (Jian et al., 2000) and give some demographic information. Participants were debriefed at the end of the study and informed that the AI was an algorithm introduced as an AI that did not use participants’ state capitals for its recommendations. To ascertain that our algorithm was convincing, we asked participants before and after the interaction how certain they were about the AI’s prediction. (“How certain are you that the AI can make an accurate recommendation for a sublease price?”) on a ten-point Likert-type scale (ranging from 1: “not certain at all” to 10: “absolutely certain”).

4.2.3. Stimuli

The apartments that participants had to evaluate were real apartments retrieved from the website Zillow (<http://zillow.com>) in May 2020. To create some variability, we selected six different apartments of different sizes and price ranges: two small-sized apartments (500–750 square feet), two medium-sized apartments (751–1,000 square feet), and two large-sized apartments (1,001–1,250 square feet). Figure 2 shows an example of how apartments were presented to participants. Features and amenities were collected directly from the website Zillow, whenever available. If not available, a random value within a reasonable range was chosen

¹ Data, R-scripts, detailed EFA results, experimental materials (including all questions and tasks), as well as a flowchart of the inclusion and exclusion process of the experiment are available under <https://osf.io/bs6q3/>.



for continuous variables (e.g., distance to public transit between 0.1 and 2.0 miles), and a random choice for dichotomous variables was made (e.g., elevator YES/NO). All participants were presented with the same stimuli, that is, identical apartments and features. What differed was the price recommendation and the explanation that accompanied it (see Figure 1).

The price recommendation from the algorithm introduced as an AI was designed to pick a random number between 10 and 20. This random number was then transformed into percentages and either added or subtracted to the initial subleasing price (T_1), which led to a random deviation between 10 and 20 percent. This deviation seemed substantial enough that subjects did not entirely adopt the recommendation, but it was also subtle enough not to appear unrealistic and that it seemed possible that the features and amenities could account for the discrepancy. With this procedure, we ensured that no participant could estimate the price accurately since there was no “true price.” Defining a ground truth has been a limitation of past studies (Poursabzi-Sangdeh et al., 2021). If, by pure chance, a participant estimates the “true price,” the interpretation of WOA becomes meaningless since T_1 and R are equal. The absence of a “true price” imposes the decision on participants either to rely or not to rely on the AI. By determining a relative deviation between 10 and 20 percent from participants’ initial price estimate, we furthermore controlled for the system’s

accuracy since it was shown to have a significant effect on people’s trust (Yin et al., 2019; Zhang et al., 2020).

It was randomly assigned that for three of the six apartments, the algorithm introduced as an AI recommended decreasing the initial price, T_1 (e.g., if T_1 was \$1,000 and the random number 17, the AI recommendation was \$830), and for the other three apartments, the recommendation was to increase T_1 (e.g., if T_1 was \$1,000 and the random number 17, the AI recommendation was \$1,170). By doing this, the AI informed participants that their initial price estimates were either too low or too high, which made it possible to compare AI recommendations to *increase* T_1 with recommendations to *decrease* T_1 . We were interested in this comparison because prior research from Kliegr et al. (2021) and Wang et al. (2019) led us to postulate that the AI’s recommendation to increase or decrease the initial price might also influence participants’ decision-making and consequently their reliance on the AI.

5. Results

5.1. Descriptive statistics

On average, participants across all conditions approached the AI recommendation, resulting in a positive WOA ($M = 0.69$,

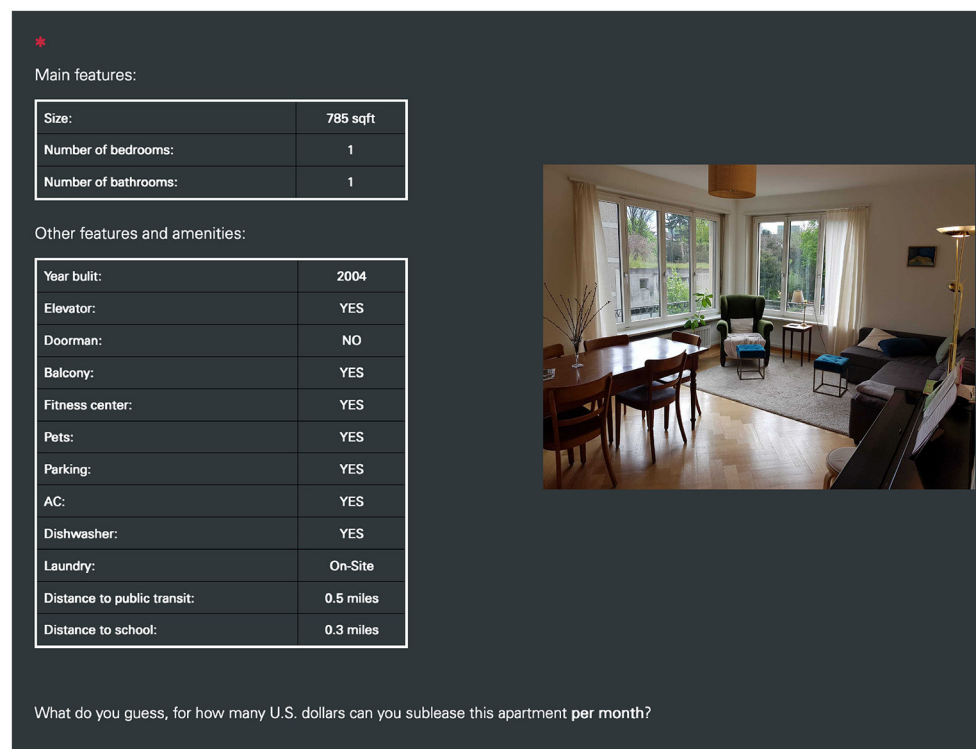


FIGURE 2

Example of how an apartment was presented to participants in the online experiment. Note that the image of the apartment depicted in the screenshot was replaced with a similar image for this publication due to potential copyright issues.

$SD = 0.36$). The TPA showed average overall ratings ($M = 5.01$, $SD = 0.86$), high ratings for trust ($M = 4.98$, $SD = 1.05$), and lower ratings for distrust ($M = 2.95$, $SD = 1.62$). Across all three conditions, the certainty that the AI could make an accurate prediction increased from pre- to post-interaction ($M_{\Delta} = 0.36$, $SD = 1.53$) and was rated at a high level after the interaction ($M = 7.59$, $SD = 1.60$). The inspection of the average estimated prices also confirmed our classifications into the apartment categories “small,” “medium,” and “large” ($M_{small} = \$1,091$, $M_{medium} = \$1,286$, $M_{large} = \$1,449$). From this, we concluded that the assigned task was a compelling one. Table 1 includes descriptive statistics for the experiment.

5.2. Reliance—WOA

To address H_1 , H_2 , and H_3 , corresponding contrasts were created. The first contrast made it possible to determine if the feature importance condition was significantly different from the control (planned contrast 1: feature importance explanation vs. control for answering H_1). By defining two other contrasts, it was possible to examine if the counterfactual condition was significantly different from the control (planned contrast 2: counterfactual explanation vs. control for answering H_2) and if the counterfactual condition was significantly different from the feature importance condition (planned contrast 3: counterfactual explanation vs. feature importance explanation for answering H_3). The effect of the

three contrasts on WOA was analyzed by employing linear mixed-effect models (LMEMs) using the *lme4* package (Bates et al., 2015) for R (version 4.2.2.). We report β -estimates, their 95% confidence interval, t -values, and the corresponding p -values. Our models contained two fixed effects: the contrasts and the difference of the recommendation to *increase* or *decrease* $T1$. Under the assumption that the stimuli and conditions had varying random effects for different participants, we introduced a random intercept (*id*) in the model. The utilized model had the following specifications:

$$WOA \sim 1 + Contrast1 + Recommendation + (1|id)$$

For this model, the first contrast (feature importance explanation vs. control) was not significant [$\beta = 0.01$, 95% CI β [-0.01, 0.03], $t_{(378)} = 0.64$, $p = 0.53$], while the difference between recommendations to increase or decrease $T1$ was highly significant [$\beta = 0.05$, 95% CI β [0.03, 0.07], $t_{(1,899)} = 3.62$, $p < 0.001$]. The second contrast (counterfactual explanation vs. control) did not return any significant results [$\beta = 0.02$, 95% CI β [-0.01, 0.04], $t_{(378)} = 1.33$, $p = 0.19$], and neither did the third contrast (counterfactual explanation vs. feature importance explanation) [$\beta = 0.01$, 95% CI β [-0.01, 0.03], $t_{(378)} = 0.63$, $p = 0.53$]. However, in all three models, the recommendation type (increasing a price vs. decreasing a price) returned highly significant results with β -estimates ranging between 95% CI [0.03, 0.07]. Comparing this model to a model without the recommendation term confirmed that its inclusion was justified since it significantly

TABLE 1 Descriptive statistics of the conducted experiment with the mean (*M*), standard deviation (*SD*), and median (*Mdn*) for WOA, the TPA, and AI certainty.

	Control			Feature importance			Counterfactual		
	(n = 130)			(n = 143)			(n = 107)		
	<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>
Weight of Advice									
<i>Increase price</i>	0.69	0.38	0.69	0.68	0.36	0.69	0.67	0.34	0.67
<i>Decrease price</i>	0.69	0.38	0.69	0.73	0.40	0.79	0.78	0.37	0.79
TPA (Jian et al., 2000)									
<i>Overall</i>	5.01	0.91	4.88	4.95	0.84	4.83	5.10	0.84	5.08
<i>Trust</i>	5.11	1.06	5.29	4.82	1.07	5.00	5.03	1.00	5.14
<i>Distrust</i>	3.12	1.73	2.60	2.88	1.58	2.40	2.82	1.54	2.20
AI certainty									
<i>Pre-interaction</i>	7.24	1.60	8.00	7.20	1.57	8.00	7.27	1.64	8.00
<i>Post-interaction</i>	7.55	1.74	8.00	7.48	1.61	8.00	7.80	1.37	8.00

improved the model fit [$\chi^2_{(1)} = 13.05, p < 0.001$]. To better understand the relationship between explanations and the type of recommendation, we created a visualization (see Figure 3).

Depending on the type of recommendation, the condition effect was different, meaning that whether the AI recommended *increasing* or *decreasing* the initial subleasing price *T1*, influenced the way that explanations affected WOA. For recommendations to increase, the explanations had a negligible effect on WOA, but for recommendations to decrease, the effect was substantial. In our case, the effect of explanations cannot be readily understood without considering the different type of AI recommendation. We therefore divided the data into two subsets. One subset contained the three apartments with the *recommendation to increase T1*, the other subset contained the three apartments with the *recommendation to decrease T1*. We then executed the specified model again, but the term “recommendation” was naturally omitted as a fixed effect. For the subset that contained the recommendations to decrease *T1*, the second contrast (counterfactual explanation vs. control) was significant [$\beta = 0.04, 95\% \text{ CI } \beta[0.01, 0.08], t_{(378)} = 2.31, p = 0.02$]. The β -estimates indicate that on average, counterfactual explanations increased WOA by an approximated 4% compared to the control that received no explanations. Note that feature importance explanations likewise increased WOA by 2% compared to the control, but this difference was not significant for the 0.05 significance level [$\beta = 0.02, 95\% \text{ CI } \beta[-0.01, 0.05], t_{(378)} = 1.10, p = 0.27$]. However, explanations had no effect on WOA when the AI recommended increasing the price estimate (Figure 3). LMEMs are quite robust against violations of distributional assumptions (Schielzeth et al., 2020). We nevertheless checked the residuals of WOA values for normal distribution via quantile–quantile plots (Q–Q plots) to determine if the residual variance was equal across conditions (homoscedasticity) and also checked the multicollinearity assumption. The normality distribution seemed to be satisfied, with some deviation from normality at the tails, which indicates that more data is located at the extremes. Levene’s test

indicated equal variances [$F_{(2)} = 0.57, p = 0.56$] that did not differ between the conditions, and a multicollinearity check revealed low correlation between the model terms.

5.3. Trust—Trust between people and automation scale

To identify the underlying structure of the TPA, we performed an exploratory factor analysis (EFA) using MinRes and rotated with the Oblimin method (see footnote 1). Parallel analysis and very simple structure (VSS) indicated two factors, which is in line with previous research (Spain et al., 2008). The first five items loaded on one factor (with 0.79 – 0.89), and the other seven loaded on a second factor (0.56 – 0.85), which corresponded accurately to the trust/distrust items of the scale. Internal consistency for the first five items (i.e., distrust) was excellent ($\alpha = 0.92, 95\% \text{ CI}[0.90, 0.93], \omega = 0.92, 95\% \text{ CI}[0.90, 0.93]$) and good for the seven trust items ($\alpha = 0.88, 95\% \text{ CI}[0.86, 0.90], \omega = 0.88, 95\% \text{ CI}[0.86, 0.90]$) according to George and Mallery (2019). To test H_4 (feature importance leads to higher trust compared to the control), H_5 (counterfactuals lead to higher trust compared to the control) and H_6 (counterfactuals lead to higher trust compared to feature importance), we intended to perform two types of one-way analyses of variance (ANOVAs), once using the overall mean score, and once using mean scores for the trust and distrust factors. However, visual inspection of the distribution and a Shapiro–Wilk test ($W = 0.97, p < 0.001$) revealed a non-trivial violation of the normality assumption. Thus, the ANOVA results might not have been interpretable and meaningful. Under these circumstances, a non-parametric Kruskal–Wallis test (Kruskal and Wallis, 1952) was carried out as it does not assume a normal distribution of the residuals. The results of the Kruskal–Wallis test showed that the overall mean ratings for the TPA were not significantly different between the conditions [$H_{(2)} = 1.54, p = 0.46$]. The same was

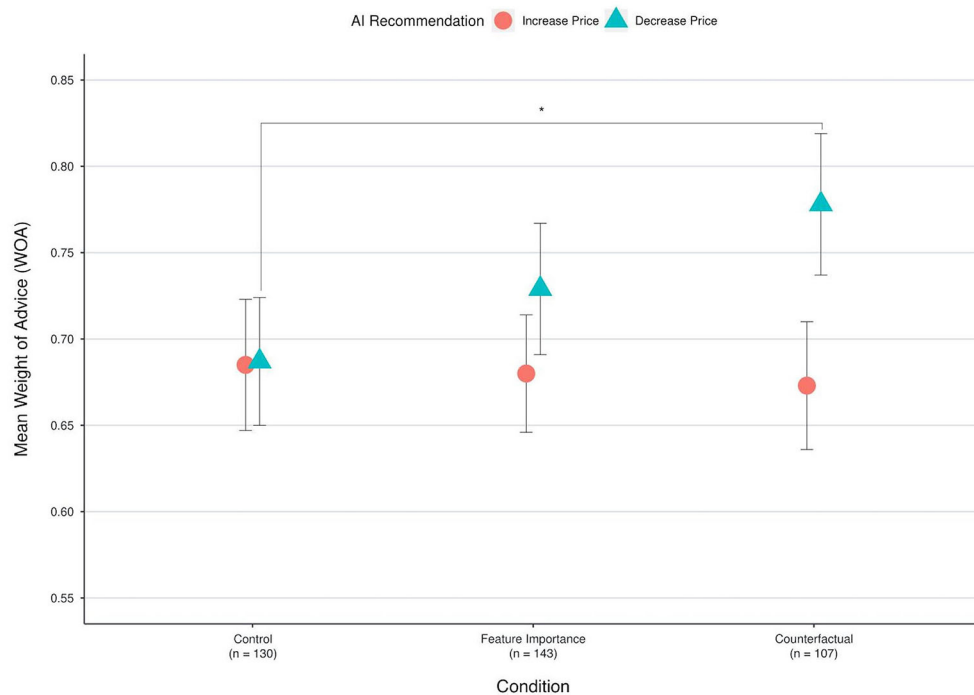


FIGURE 3

Mean weight of advice for each condition and type of recommendation. Note that the y-axis is scaled to better visualize the effect. The error bars depict 95% confidence intervals. *Statistically significant difference with $p < 0.05$.

true for the factors trust [$H_{(2)} = 5.06, p = 0.08$] and distrust [$H_{(2)} = 2.03, p = 0.36$]. Since the omnibus Kruskal–Wallis was not significant, we did not perform further *post-hoc* tests. Figure 4 captures the similar trust ratings for the two experimental conditions and the control.

6. Discussion

The experiment reported in this study demonstrates that participants generally rely on AI recommendations in low-stake decision-making tasks—in this case, receiving AI recommendations to find an optimal price for subleasing an apartment. Regardless of the different experimental manipulations, on average, participants displayed high overall Weight of Advice scores ($M = 0.67 - 0.69, SD = 0.25 - 0.36$). A WOA of 0.70 signifies that participants adopted 70% of the AI recommendations when updating their prior beliefs to form their final estimate. This finding supports the idea that people generally rely on AI (Logg et al., 2019).

The results further demonstrated that under certain conditions, explainability significantly increases AI reliance. However, in the context of our study, the effect of human-centered explanations depended on the type of decision-making the participants had to engage in. We presented participants with two kinds of recommendations: for the first type, an algorithm introduced as an AI recommended that participants *increase* their initially estimated apartment price. For the second type, participants were advised to *decrease* their initial price. The results of the

experiment indicate that when the AI recommended increasing the price, human-centered explanations did not affect reliance. By contrast, in the case of recommendations to decrease the price, providing counterfactual explanations affected WOA significantly. Participants in the counterfactual condition where the AI recommended decreasing $T1$ relied up to 9 percentage points more on the recommendation than participants in the control that received the recommendation to increase their price. Therefore, the findings support the second hypothesis (H_2 : counterfactuals will lead to higher reliance compared to the control) only for decision-making tasks where the AI recommended decreasing the price. The first hypothesis (H_1 : feature importance will lead to higher reliance compared to the control) and the third hypothesis (H_3 : counterfactuals will lead to higher reliance compared to feature importance) are not supported for either type of recommendation. We conclude that counterfactual explanations can significantly increase reliance but only under certain conditions. However, there was no significant difference between the two *post-hoc* explanations, although counterfactuals are arguably more human-centered since they additionally fulfill Miller (2019)'s *contrastivity* criterion.

The experiment illustrated that the decision-making task with regards to increasing or decreasing a price had a significant effect on reliance. Regardless of providing explanations, participants consistently relied more on AI recommendations to decrease prices than recommendations to increase them (see Figure 3). This seems counterintuitive at first glance since one might expect that participants would always embrace the prospect of obtaining a higher subleasing price. We argue that the two

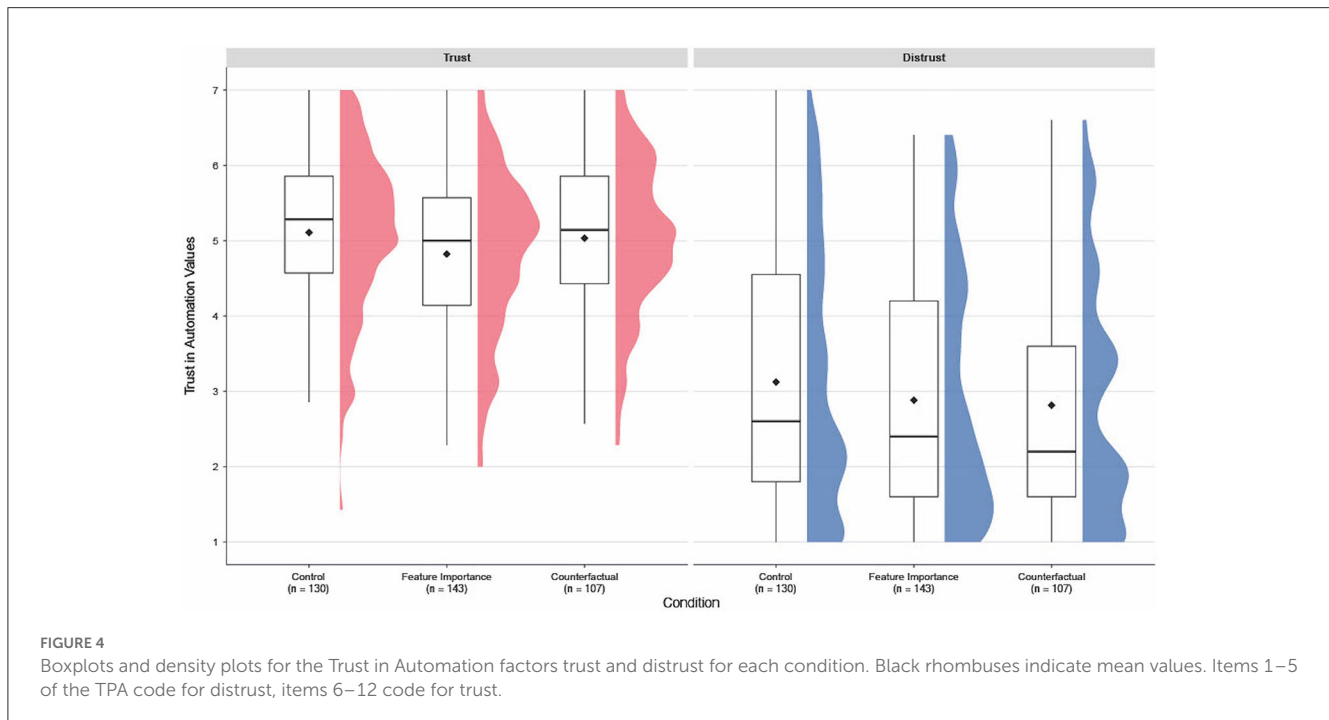


FIGURE 4

Boxplots and density plots for the Trust in Automation factors trust and distrust for each condition. Black rhombuses indicate mean values. Items 1–5 of the TPA code for distrust, items 6–12 code for trust.

types of recommendations should be thought of as two distinct decision-making tasks and our results demonstrate how cognitive biases may affect humans in their decision-making involving AI as proposed by Kliegr et al. (2021). The well-studied concept of *loss aversion* by Tversky and Kahneman (1991) could account for this discrepancy and serve as an explanation attempt for our findings. Loss aversion suggests that, psychologically, people assign more utility to losses than to gains (Tversky and Kahneman, 1991). In practical terms, this means that the dissatisfaction experienced by a person who loses \$100 is greater than the satisfaction experienced by a person who gains \$100. Our study design seems to satisfy the preconditions for a possible loss-aversion effect: when participants received a recommendation to increase their initial price estimate, they were likely concerned that this potential price increase would cause an unsuccessful sublease. The prospect of getting more money (gain) mattered less in this decision-making task than the possibility of not being able to sublease at all (loss). A recommendation to decrease the initial price may not have induced loss aversion in participants. When not being confronted with loss aversion, explanations seemed to convince participants that demanding less money was the right decision to successfully sublease the apartment, compared to the control where no additional explanation was present for the recommendation. This interpretation suggests that human-centered AI explanations can have an effect on reliance, but only for decision-making tasks where other contributing factors such as loss aversion are absent. Substantial work has been published about biased AI-training data but little about humans' cognitive biases and heuristics when exposed to AI. A notable exception is the work of Lu and Yin (2021) that showed how people use heuristics and base their reliance on the level of agreement between the machine learning model and themselves when performance feedback was limited. Moreover, Wang et al. (2019) proposed a

framework of how human reasoning should inform XAI to mitigate possible cognitive biases, and a recent review by Kliegr et al. (2021) explored to what extent biases affect human understandings of interpretable machine-learning models. We present empirical findings suggesting that the XAI community should account for possible biases and heuristics to develop genuinely human-centered explanations. Inherent biases and heuristics may be so hardwired in people that AI explanations are not convincing enough to disprove non-optimal human decision-making. If that is the case, AI may not help users to reach better decisions in circumstances where human intuition becomes too tempting for their judgment. While the interpretation of the present results under the perspective of loss aversion requires further investigation, our findings highlight the importance of biases and heuristic end-users can exhibit in actual decision-making tasks rather than proxy tasks. These biases and heuristics may result in irrational and non-optimal choices, which in turn affect the measured variables of interest, including trust and trust-related behavior (Wang et al., 2019; Kliegr et al., 2021). In the context of our study, the type of decision-making participants faced had a greater effect on reliance than the explanation provided by the AI. This suggests that factors other than explainability are crucial when designing human-centered AI. Cognitive biases and heuristics, such as loss aversion (Tversky and Kahneman, 1991), framing (Tversky and Kahneman, 1981), or confirmation bias (Wason, 1960), could potentially undermine AI explanations.

Concerning trust, no significant differences were found for human-centered explanations (Figure 4). Therefore, we reject the fourth hypothesis (H_4 : feature importance will lead to higher trust compared to the control), the fifth hypothesis (H_5 : counterfactuals will lead to higher trust compared to the control), and the sixth hypothesis (H_6 : counterfactuals will lead to higher trust compared to feature importance). While the effect of human-centered explanations on reliance depended on

the nature of the AI recommendation (increasing or decreasing the initial estimated apartment price), this dependence and the potential effect of cognitive biases and heuristics remain to be explored for trust. We operationalized trust as an attitude toward the AI system and consequently assessed users' trust after the entire task while we observed reliance from trial-to-trial. Given the present study, our results do not indicate a consistent effect of human-centered explanations on trust. Thus, our findings are in line with other research that provided mixed evidence regarding the effect of transparency on trust (Cramer et al., 2008; Nothdurft et al., 2013; Cheng et al., 2019; Ehsan et al., 2019; Zhang et al., 2020; Poursabzi-Sangdeh et al., 2021). However, the conceptual distinction between trust and reliance carries significant implications for XAI evaluation and uncovers two potential challenges. First, if researchers only assess attitudinal trust via questionnaires, they could falsely assume that people will not rely on an AI system. Second, if only trust-related behavior (i.e., reliance) is measured, researchers might incorrectly deduce that people necessarily trust the system in question. Consequently, researchers and practitioners have to answer the challenging question of which of the two measures to account for and investigate when evaluating AI. Whether trust translates into reliance is more nuanced than often assumed and depends on an interaction between the operator, the automation, and the situation (Körber, 2018). It remains to be determined which factors other than trust drive AI reliance (e.g., system accuracy, perceived usefulness, cognitive biases) or whether current measurement tools originally designed to assess trust in automation also accurately capture AI trust.

Furthermore, we initially chose to measure overall trust, expecting a single-factor structure as proposed by Jian et al. (2000). However, investigating the scale's factor structure through exploratory factor analysis before interpreting the results implied a two-factor structure, as previously observed outside the AI context (Spain et al., 2008). This change in the theoretical structure led us to use the TPA to measure trust and distrust as two distinct factors in the analysis. While in our study, levels of trust and distrust behaved as expected across the three conditions (i.e., high trust and low distrust), it is plausible to assume that there are situations where the difference between people's trust and distrust in an AI is more nuanced. Attitudes are often seen as lying along one continuum, as was initially proposed for the TPA, but past research has argued that positive and negative attitudes can co-occur (Priester and Petty, 1996). For example, when smokers try to quit smoking, they can have a simultaneously negative and positive view toward cigarettes (Cacioppo and Berntson, 1994). Thus, it may be necessary to distinguish between trust and distrust in studies that aim to investigate ambivalent attitudes toward AI. This distinction can be accounted for when using the two-factor structure for the TPA. On the other hand, a single-factor structure comes with the risk of oversimplifying situations and losing important nuance when using the TPA to measure trust in AI. Overall, our findings emphasize that researchers must carefully differentiate attitude from behavior and choose appropriate evaluation metrics for human-centered AI accordingly.

7. Limitations and future work

We conducted an online decision-making experiment in a domain-specific task. Future work should broaden the scope and focus on domains other than real estate to investigate if the findings of this study are transferable to different scenarios and AI systems used in practice to increase external validity. While decision-making experiments on MTurk allow high control over confounding variables and are comparable to those in laboratory settings, even in low-stakes scenarios (Amir et al., 2012), future studies could focus on high-stakes decisions to evoke uncertainty where a more tangible loss depends on the participants' decision to trust AI.

Participants were presented with AI recommendations that were expected to seem reasonably trustworthy since the recommendations were formed based on the participants' initial estimates. However, explainability might have a greater impact on trust and reliance if the recommendations were not credible or showed greater deviations from the initial estimate. Past research has shown that people calibrate their trust based on the system's capabilities (Lee and See, 2004; Zhang et al., 2020), and often fail to rely on algorithms after learning that they are imperfect, a phenomenon called algorithm aversion (Dietvorst et al., 2015). By providing explanations, people may better understand AI errors and factors that influence those errors (Dzindolet et al., 2003). Future research could investigate more untrustworthy recommendations by gradually reducing the capabilities of the system (e.g., by decreasing the system's accuracy) and examining how explainability affects reliance and trust in cases where the AI objectively performs poorly or when there is a clear disagreement between the end-user and the AI.

The study design did not allow for a clear distinction between *dispositional trust*, *situational trust*, and *learned trust*, as suggested by Hoff and Bashir (2015). In addition, by measuring trust as an attitude toward the AI system after task completion, examining the effects of bias and heuristics at the level of individual trials was not possible. Trust is a dynamic process in the human-AI interaction, and we expect that trust changes as time passes in the interaction between end-users and AI-based systems. We recommend that future studies investigate the varying manifestations of trust because they are critical for a comprehensive understanding of the human-AI interaction. Researchers could measure AI trust before and after participants are exposed to an AI system and compare the reported trust scores (learned trust). Alternatively, they could expose participants to AI recommendations while inducing different emotional valences (situational trust). Future research could also investigate the relationship between AI trust and reliance from the perspective of the technology acceptance model (Davis, 1989), which can be seen as a further development of Ajzen and Fishbein (1980)'s work.

8. Conclusion

We conducted an empirical experiment demonstrating that human-centered explanations as a means for transparent AI increase reliance for specific decision-making tasks. While this

provides some evidence that human-centered *post-hoc* explanations can be an opportunity for more transparent AI, our findings emphasize that the effect of transparency on reliance and trust is more nuanced than commonly assumed.

The type of decision-making task (increasing vs. decreasing a price) had a greater influence on end-users' reliance than how the AI explained its decision did. We argue that humans may exhibit cognitive biases and apply heuristics to decision-making tasks that involve AI. So far, the discussion around bias has focused primarily on biased data and prejudice due to incorrect assumptions in the machine-learning process. The implications of potential biases and heuristics when humans are presented with explanations from AI have received little attention in the current XAI debate. Both researchers and practitioners need to be aware of such dynamics in the design for truly human-centered AI, as poor partnership between people and automation will become increasingly costly and consequential (Lee and See, 2004).

In order to draw valid conclusions from experiments, XAI researchers need to be cautious when measuring the human side of the human–AI interaction. Conceptualizing trust as an attitude and reliance as a trust-related behavior might lead to divergent results. Our study also confirmed a two-factor structure (trust and distrust) for the TPA, as previously reported outside the AI context. Given the importance of AI, as it is increasingly used to make critical decisions with far-reaching implications, meaningful evaluations in XAI research require agreed-upon metrics and appropriate measurements that have been empirically validated.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found here: <https://osf.io/bs6q3/>.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants

provided their written informed consent to participate in this study.

Author contributions

NS and FB contributed to the conception and design of the study and implemented the online study. NS collected the data and wrote the first draft. NS, FB, and SP performed the statistical analysis. All authors contributed to the manuscript revision, read, and approved the submitted version.

Funding

This research was financed entirely by our research group and the publication fund of the University of Basel for open access; we received no additional funding.

Acknowledgments

Special thanks to Zgijm Memeti and Léane Wettstein.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., and Kankanhalli, M. (2018). "Trends and trajectories for explainable, accountable and intelligible systems: an HCI research agenda," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18* (New York, NY: ACM), 1–18.
- ACM FAccT Conference (2022). "ACM conference on fairness, accountability, and transparency 2022 (ACM FAccT 2022) call for papers," in *ACM Conference on Fairness, Accountability, and Transparency 2022* (New York, NY).
- Adadi, A., and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* 6, 52138–52160. doi: 10.1109/ACCESS.2018.2870052
- Ajzen, I., and Fishbein, M. (1980). *Understanding Attitudes and Predicting Social Behavior*. Englewood Cliffs, NJ: Prentice-Hall.
- Amir, O., Rand, D. G., and Gal, Y. K. (2012). Economic games on the internet: the effect of \$1 stakes. *PLoS ONE* 7, e31461. doi: 10.1371/journal.pone.0031461
- Arrieta, A. B., Diaz-Rodriguez, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inform. Fusion* 58, 82–115. doi: 10.1016/j.inffus.2019.12.012
- Bates, D., Mochler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01
- Biran, O., and Cotton, C. (2017). "Explanation and justification in machine learning: a survey," in *IJCAI-17 Workshop on Explainable AI (XAI)* (Melbourne), 8–13.
- Buçinca, Z., Lin, P., Gajos, K. Z., and Glassman, E. L. (2020). "Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems," in *Proceedings of the 25th International Conference on Intelligent User Interfaces, IUI '20* (New York, NY: ACM), 454–464.
- Cacioppo, J. T., and Bernston, G. G. (1994). Relationship between attitudes and evaluative space: a critical review, with emphasis on the separability of positive and negative substrates. *Psychol. Bull.* 115, 401–423.

- Cheng, H.-F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M., et al. (2019). "Explaining decision-making algorithms through UI: strategies to help non-expert stakeholders," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19* (New York, NY: ACM), 1–12.
- Chopra, K., and Wallace, W. A. (2003). "Trust in electronic environments," in *Proceedings of the 36th Annual Hawaii International Conference on System Sciences, HICSS '03* (Los Alamitos, CA: IEEE Computer Society Press, IEEE Computer Society Press), 10–15.
- Cramer, H., Evers, V., Ramlal, S., Someren, M., Rutledge, L., Stash, N., et al. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Model. User Adapt. Interact.* 18, 455–496. doi: 10.1007/s11257-008-9051-3
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* 13, 319–340.
- De Graaf, M. M. A., and Malle, B. F. (2018). "People's judgments of human and robot behaviors: a robust set of behaviors and some discrepancies," in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, HRI '18* (New York, NY: ACM), 97–98.
- Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *J. Exp. Psychol. Gen.* 144, 114–126. doi: 10.1037/xge0000033
- Du, M., Liu, N., and Hu, X. (2019). Techniques for interpretable machine learning. *Commun. ACM* 63, 68–77. doi: 10.1145/3359786
- Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., and Beck, H. P. (2003). The role of trust in automation reliance. *Int. J. Hum. Comput. Stud.* 58, 697–718. doi: 10.1016/S1071-5819(03)00038-7
- Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., and Weisz, J. D. (2021a). "Expanding explainability: towards social transparency in AI systems," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), 1–19.
- Ehsan, U., and Riedl, M. O. (2020). "Human-centered explainable AI: towards a reflective sociotechnical approach," in *HCI International 2020-Late Breaking Papers: Multimodality and Intelligence: 22nd HCI International Conference, HCII 2020* (Copenhagen), 449–466.
- Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., and Riedl, M. O. (2019). "Automated rationale generation: a technique for explainable ai and its effects on human perceptions," in *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI '19* (New York, NY: ACM), 263–274.
- Ehsan, U., Wintersberger, P., Liao, Q. V., Mara, M., Streit, M., Wachter, S., et al. (2021b). "Operationalizing human-centered perspectives in explainable AI," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery), 1–6.
- European Parliament and Council of the European Union (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons With Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation)*.
- Ferreira, J. J., and Monteiro, M. S. (2020). "What are people doing about XAI user experience? A survey on ai explainability research and practice," in *Design, User Experience, and Usability. Design for Contemporary Interactive Environments*, eds A. Marcus and E. Rosenzweig (Cham: Springer International Publishing), 56–73.
- George, D., and Mallery, P. (2019). *IBM SPSS Statistics 26 Step by Step: A Simple Guide and Reference, 16th Edn.* New York, NY: Routledge.
- Gino, F., and Moore, D. A. (2007). Effects of task difficulty on use of advice. *J. Behav. Decis. Mak.* 20, 21–35. doi: 10.1002/bdm.539
- Gutzwiller, R. S., Chiou, E. K., Craig, S. D., Lewis, C. M., Lematta, G. J., and Hsiung, C.-P. (2019). "Positive bias in the 'trust in automated systems survey'? An examination of the Jian et al. (2000) scale," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 217–221.
- Hartmann, M. (2020). *Vertrauen - Die unsichtbare Macht.* Berlin: Fischer Verlag.
- Harvey, N., and Fischer, I. (1997). Taking advice: accepting help, improving judgment, and sharing responsibility. *Organ. Behav. Hum. Decis. Process.* 70, 117–133.
- Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychol. Bull.* 107, 65–81.
- Hoff, K. A., and Bashir, M. (2015). Trust in automation: integrating empirical evidence on factors that influence trust. *Hum. Fact.* 57, 407–434. doi: 10.1177/0018720814547570
- Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. (2019). Metrics for explainable AI: challenges and prospects. *arXiv preprint arxiv: 1812.04608*. doi: 10.48550/arXiv.1812.04608
- Hong, S. R., Hullman, J., and Bertini, E. (2020). "Human factors in model interpretability: industry practices, challenges, and needs," in *Proceedings of the ACM on Human-Computer Interaction* (New York, NY).
- Jacovi, A., Marasović, A., Miller, T., and Goldberg, Y. (2021). "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, FAccT '21* (New York, NY: ACM), 624–635.
- Jian, J.-Y., Bisantz, A. M., and Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *Int. J. Cogn. Ergon.* 4, 53–71. doi: 10.1207/S15327566IJCE0401_04
- Kirlik, A. (1993). Modeling strategic behavior in human-automation interaction: why an "aid" can (and should) go unused. *Hum. Fact.* 35, 221–242.
- Kizilcec, R. F. (2016). "How much information? Effects of transparency on trust in an algorithmic interface," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), 2390–2395.
- Kliegr, T., Stepan Bahník, and Furnkranz, J. (2021). A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artif. Intell.* 295, 103458. doi: 10.1016/j.artint.2021.103458
- Kocielnik, R., Amershi, S., and Bennett, P. N. (2019). "Will you accept an imperfect AI? Exploring designs for adjusting end-user expectations of AI systems," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), 1–14.
- Körber, M. (2018). "Theoretical considerations and development of a questionnaire to measure trust in automation," in *Proceedings of the 20th Congress of the International Ergonomics Association, IEA '18* (Cham: Springer International Publishing), 13–30.
- Krause, J., Perer, A., and Ng, K. (2016). "Interacting with predictions: visual inspection of black-box machine learning models," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), 5686–5697.
- Kruskal, W. H., and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.* 47, 583–621.
- Kulesza, T., Burnett, M., Wong, W.-K., and Stumpf, S. (2015). "Principles of explanatory debugging to personalize interactive machine learning," in *Proceedings of the 20th International Conference on Intelligent User Interfaces, IUI '15* (New York, NY: ACM), 126–137.
- Lai, V., and Tan, C. (2019). "On human predictions with explanations and predictions of machine learning models: a case study on deception detection," in *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19* (New York, NY: ACM), 29–38.
- Langer, M., Oster, D., Speith, T., Kästner, L., Hermanns, H., Schmidt, E., et al. (2021). What do we want from explainable artificial intelligence (XAI)? A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artif. Intell.* 296, 103473. doi: 10.1016/j.artint.2021.103473
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539
- Lee, J. D., and See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Hum. Fact.* 46, 50–80. doi: 10.1518/hfes.46.1.50.30392
- Liao, Q. V., Gruen, D., and Miller, S. (2020). "Questioning the AI: informing design practices for explainable ai user experiences," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20* (New York, NY: ACM), 1–15.
- Lipton, Z. C. (2018a). The mythos of model interpretability. *Commun. ACM* 61, 36–43. doi: 10.1145/3233231
- Lipton, Z. C. (2018b). The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 31–57. doi: 10.1145/3236386.3241340
- Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm appreciation: people prefer algorithmic to human judgment. *Organ. Behav. Hum. Decis. Process.* 151, 90–103. doi: 10.1016/j.obhdp.2018.12.005
- Lu, Z., and Yin, M. (2021). "Human reliance on machine learning models when performance feedback is limited: heuristics and risks," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), 1–16.
- Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. *Acad. Manage. Rev.* 20, 709–734.
- Merritt, S. M. (2011). Affective processes in human-automation interactions. *Hum. Fact.* 53, 356–370. doi: 10.1177/0018720811411912
- Miller, D., Johns, M., Mok, B., Gowda, N., Sirkin, D., Lee, K., et al. (2016). "Behavioral measurement of trust in automation: the trust fall," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Washington, DC), 1849–1853.
- Miller, T. (2019). Explanation in artificial intelligence: insights from the social sciences. *Artif. Intell.* 267, 1–38. doi: 10.1016/j.artint.2018.07.007
- Mittelstadt, B., Russell, C., and Wachter, S. (2019). "Explaining explanations in AI," in *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19* (New York, NY: ACM), 279–288.
- Mohseni, S., Zarei, N., and Ragan, E. D. (2020). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *arXiv preprint arxiv:1811.11839*.
- Mothilal, R. K., Sharma, A., and Tan, C. (2020). "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020*

- Conference on Fairness, Accountability, and Transparency, FAT* '20 (New York, NY: ACM), 607–617.
- Mucha, H., Robert, S., Breitschwerdt, R., and Fellmann, M. (2021). “Interfaces for explanations in human-AI interaction: proposing a design evaluation approach,” in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), 1–6.
- Nothdurft, F., Heinroth, T., and Minker, W. (2013). “The impact of explanation dialogues on human-computer trust,” in *Proceedings, Part III, of the 15th International Conference on Human-Computer Interaction. Users and Contexts of Use* (Berlin; Heidelberg: Springer-Verlag), 59–67.
- Papenmeier, A., Kern, D., Englebienne, G., and Seifert, C. (2022). It's complicated: the relationship between user trust, model accuracy and explanations in AI. *ACM Trans. Comput. Hum. Interact.* 29, 1–33. doi: 10.1145/3495013
- Parasuraman, R., and Manzey, D. H. (2010). Complacency and bias in human use of automation: an attentional integration. *Hum. Fact.* 52, 381–410. doi: 10.1177/0018720810376055
- Parasuraman, R., and Riley, V. (1997). Humans and automation: use, misuse, disuse, abuse. *Hum. Fact.* 39, 230–253.
- Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., and Wallach, H. (2021). “Manipulating and measuring model interpretability,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21* (New York, NY: ACM), 1–52.
- Priester, J. R., and Petty, R. E. (1996). The gradual threshold model of ambivalence: relating the positive and negative bases of attitudes to subjective ambivalence. *J. Pers. Soc. Psychol.* 71, 431.
- Rousseau, D. M., Sitkin, S. B., Burt, R. S., and Camerer, C. (1998). Not so different after all: a cross-discipline view of trust. *Acad. Manage. Rev.* 23, 393–404.
- Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K.-R. (2019). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Berlin: Springer.
- Sanneman, L., and Shah, J. A. (2022). The situation awareness framework for explainable AI (safe-AI) and human factors considerations for XAI systems. *Int. J. Hum. Comput. Interact.* 38, 1772–1788. doi: 10.1080/10447318.2022.2081282
- Scharowski, N., Perrig, S. A., von Felten, N., and Brühlmann, F. (2022). “Trust and reliance in XAI-distinguishing between attitudinal and behavioral measures,” in *CHI TRAIT Workshop* (New York, NY).
- Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allogue, H., Teplitsky, C., et al. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods Ecol. Evol.* 11, 1141–1152. doi: 10.1111/2041-210X.13434
- Shmueli, G. (2010). To explain or to predict? *Stat. Sci.* 25, 289–310. doi: 10.1214/10-STS330
- Solso, R. L., MacLin, M. K., and MacLin, O. H. (2005). *Cognitive Psychology*. Auckland: Pearson Education.
- Spain, R. D., Bustamante, E. A., and Bliss, J. P. (2008). “Towards an empirically developed scale for system trust: take two,” in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Los Angeles, CA: Sage Publications), 1335–1339.
- Stephanidis, C., Salvendy, G., Antona, M., Chen, J. Y. C., Dong, J., Duffy, V. G., et al. (2019). Seven HCI grand challenges. *Int. J. Hum. Comput. Interact.* 35, 1229–1269. doi: 10.1080/10447318.2019.1619259
- Suresh, H., Gomez, S. R., Nam, K. K., and Satyanarayan, A. (2021). “Beyond expertise and roles: a framework to characterize the stakeholders of interpretable machine learning and their needs,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), 1–16.
- Szymanski, M., Millecamp, M., and Verbert, K. (2021). “Visual, textual or hybrid: the effect of user expertise on different explanations,” in *26th International Conference on Intelligent User Interfaces, IUI '21* (New York, NY: ACM), 109–119.
- Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science* 185, 1124–1131.
- Tversky, A., and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science* 211, 453–458.
- Tversky, A., and Kahneman, D. (1991). Loss aversion in riskless choice: a reference-dependent model. *Q. J. Econ.* 106, 1039–1061.
- Vereschak, O., Bailly, G., and Caramiaux, B. (2021). “How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies,” in *Proceedings of the ACM on Human-Computer Interaction* (New York, NY: ACM), 1–39.
- Wachter, S., Mittelstadt, B., and Russell, C. (2018). Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harvard J. Law Technol.* 31, 841–887. doi: 10.2139/ssrn.3063289
- Wang, D., Yang, Q., Abdul, A., and Lim, B. Y. (2019). “Designing theory-driven user-centric explainable AI,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), 1–15.
- Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Q. J. Exp. Psychol.* 12, 129–140.
- Wischnewski, M., Krämer, N., and Müller, E. (2023). “Measuring and understanding trust calibrations for automated systems: a survey of the state-of-the-art and future directions,” in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23* (New York, NY: Association for Computing Machinery).
- Yin, M., Wortman Vaughan, J., and Wallach, H. (2019). “Understanding the effect of accuracy on trust in machine learning models,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), 1–12.
- Yu, K., Berkovsky, S., Taib, R., Conway, D., Zhou, J., and Chen, F. (2017). “User trust dynamics: an investigation driven by differences in system performance,” in *Proceedings of the 22nd International Conference on Intelligent User Interfaces, IUI '17* (New York, NY: ACM), 307–317.
- Zhang, Y., Liao, Q. V., and Bellamy, R. K. E. (2020). “Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making,” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20* (New York, NY: ACM), 295–305.

To Trust or Distrust Trust Measures: Validating Questionnaires for Trust in AI

Nicolas Scharowski^{a,c,*}, Sebastian A. C. Perrig^{a,c}, Lena Fanya Aeschbach^a, Nick von Felten^a, Prof. Dr. Klaus Opwis^a, Prof. Dr. Philipp Wintersberger^b and Dr. Florian Brühlmann^a

^aCenter for General Psychology and Methodology, University of Basel, Switzerland

^bUniversity of Applied Sciences Hagenberg, Hagenberg, Austria

^cBoth authors contributed equally to this research.

ARTICLE INFO

Keywords:
AI
Trust
Distrust
Survey
Questionnaire
Validation

ABSTRACT

Despite the importance of trust in human-AI interactions, researchers must adopt questionnaires from other disciplines that lack validation in the AI context. Motivated by the need for reliable and valid measures, we investigated the psychometric quality of two trust questionnaires, the *Trust between People and Automation* scale (TPA) by Jian, Bisantz and Drury (2000) and the *Trust Scale for the AI Context* (TAI) by Hoffman, Mueller, Klein and Litman (2023). In a pre-registered online experiment ($N = 1485$), participants observed interactions with trustworthy and untrustworthy AI (autonomous vehicle and chatbot). Results support the psychometric quality of the TAI while revealing opportunities to improve the TPA, which we outline in our recommendations for using the two questionnaires. Furthermore, our findings provide additional empirical evidence of trust *and* distrust as two distinct constructs that may coexist independently. Building on our findings, we highlight the opportunities and added value of measuring both trust and distrust in human-AI research and advocate for further work on both constructs.

1. Introduction

With artificial intelligence (AI) becoming increasingly integrated into people's daily lives, the concept of trust recently got a lot of traction. Trust is not only an essential element in human-AI interactions as it shapes how people use and rely on AI (Hoff and Bashir, 2015; Lee and See, 2004), but also a key motivation for research into explainable AI (XAI) to create more transparent AI systems (Lipton, 2018). Consequently, there is a growing need for a comprehensive understanding and appropriate measurement of human trust in AI. However, the operationalization and measurement of trust are complicated by various challenges.

For one thing, a multitude of different definitions and conceptualizations of trust exist (Benk, Tolmeijer, von Wangenheim and Ferrario, 2022; Vereschak, Bailly and Caramiaux, 2021; Ueno, Sawa, Kim, Urakami, Oura and Seaborn, 2022; Muir, 1994) that are often not clearly distinguished from related terms (e.g., "reliance" (Poursabzi-Sangdeh, Goldstein, Hofman, Wortman Vaughan and Wallach, 2021), "situational trust" (Hoff and Bashir, 2015), "perceived trustworthiness" (Weitz, Schiller, Schlagowski, Huber and André, 2021), "calibrated trust" (Langer, Oster, Speith, Kästner, Hermanns, Schmidt, Sesting and Baum, 2021) or "warranted trust" (Jacovi, Marasović, Miller and Goldberg, 2021; Hoffman, Lee, Woods, Shadbolt, Miller and Bradshaw, 2009)). Not clearly distinguishing between these terms can lead to theoretical entanglements and divergent operationalizations of trust (Kohn, de Visser, Wiese, Lee and Shaw, 2021). For example, trust, viewed as an attitude (Lee and See, 2004), is a subjective psychological construct, typically measured via questionnaires, also called survey scales (Scharowski, Perrig, von Felten and Brühlmann, 2022). Meanwhile, reliance, as a behavior (Lee and See, 2004), can be assessed using more objective observational methods such as analyzing changes in an individual's behavior after being presented with an AI recommendation (e.g., switch ratios (Lu and Yin, 2021; Yin, Wortman Vaughan and Wallach, 2019)). Conceptualizations such as "calibrated" or "warranted" trust also require this differentiation and emphasize that the motivation of XAI should not be merely to

*Corresponding author

✉ nicolas.scharowski@unibas.ch (N. Scharowski); sebastian.perrig@unibas.ch (S.A.C. Perrig); lena.aeschbach@unibas.ch (L.F. Aeschbach); nick.vonfelten@unibas.ch (N.v. Felten); klaus.opwis@unibas.ch (Prof.Dr.K. Opwis); philipp.wintersberger@fh-hagenberg.at (Prof.Dr.P. Wintersberger); florian.bruehlmann@unibas.ch (Dr.F. Brühlmann)
ORCID(s): 0000-0001-5983-346X (N. Scharowski)

increase trust arbitrarily and unjustifiably. Instead, trust should be aligned and calibrated to the AI's trustworthiness (Lee and See, 2004; Wischnewski, Krämer and Müller, 2023). In this regard, trust is warranted when the AI is trustworthy and unwarranted when it is untrustworthy (Jacovi et al., 2021). Although the importance of calibrated trust has been recognized by the community (Wischnewski et al., 2023), the corresponding perspective – that distrust in untrustworthy AI is also warranted – remains relatively underemphasized, despite being an integral factor motivating XAI (Jacovi et al., 2021). Indeed, distrust seems a comparatively overlooked construct in current human-AI research (Ueno et al., 2022; Scharowski and Perrig, 2023).

Beyond these theoretical challenges, empirical studies measuring trust often use single items (e.g., Yu, Berkovsky, Taib, Conway, Zhou and Chen, 2017) or develop their own questionnaires (e.g., Yin et al., 2019; Merritt, 2011). However, self-developed questionnaires and single items usually lack a rigid construction and quality assurance process and are often only used in an individual study, complicating comparing different study results (Furr, 2011). Thus, it has been recommended to use validated trust questionnaires (Wischnewski et al., 2023) whose psychometric quality (i.e., objectivity, reliability, and validity) has been scrutinized. But even if researchers address these challenges and use standardized questionnaires for measuring trust, they have to resort to and adapt scales from other disciplines, as there is no validated questionnaire for trust in AI. For example, it is common practice among researchers to use the *Trust between People and Automation* scale (TPA) by Jian et al. (2000) and rephrase the questionnaires' items to fit the study context (Vereschak et al., 2021). However, such practices raise concerns about whether the modified scale still measures what it was initially intended to measure (Furr, 2011; Juniper, 2009). More recently, Hoffman et al. (2023) recommended a *Trust Scale for the AI Context* (TAI) that is based on existing trust scales, including the TPA. While adopting items from other scales is a common first step in questionnaire development, a scale's psychometric quality should be reevaluated after each modification or adoption (Furr, 2011; Juniper, 2009). However, this has yet to be the case for the TAI. In fact, most studies measuring human trust in AI do not report the psychometric quality of the questionnaires they used (Vereschak et al., 2021) and only recently, Lai, Chen, Smith-Renner, Liao and Tan (2023) pointed out that the research community lacks practices to validate and reuse standardized measurements. At best, this makes it challenging for other researchers to replicate or build upon existing work. At worst, using non-validated trust questionnaires in the context of AI can generate research results that do not withstand psychometric scrutiny and thus lead to ambiguous or inconsistent findings, impeding progress in human-AI interaction and XAI research. Despite this need for standardized measures, the psychometric quality of both the TPA and TAI remains to be thoroughly investigated in an AI context. Our work aims to fill this research gap by validating the TPA and TAI in a pre-registered online experiment, following current best practices for investigating questionnaire quality. We decided to validate the TPA given that it is the most frequently used scale for measuring human-AI trust (Vereschak et al., 2021; Wischnewski et al., 2023; Hoff and Bashir, 2015; Ueno et al., 2022; Kohn et al., 2021). Despite this popularity, little is known about the psychometric quality of the TPA in the context of human-AI trust, given that the scale was initially designed for an automation context. Conversely, the TAI is the only questionnaire we know of that was explicitly designed to measure trust in an AI context. Despite this unique status of the TAI, little evidence of its psychometric quality exists.

The contribution of this article is threefold. First, we present the first comprehensive psychometric evaluation of Jian et al. (2000)'s TPA scale in an AI setting. Second, we conduct an extensive independent psychometric evaluation of the TAI by Hoffman et al. (2023). Third, we compare the two trust scales and offer recommendations and guidance for researchers and practitioners who want to use the TPA and TAI in the context of AI. Results from the 2x2 mixed design online experiment ($N = 1485$), utilizing videos depicting trustworthy and untrustworthy AI in two common application areas (chatbot vs. automated vehicle), show that the TAI performs well psychometrically. Concerning the TPA, somewhat acceptable quality was only achieved after removing items and when considering a two-factor model (trust and distrust) instead of the initially proposed single-factor model. Based on these findings, we advocate for future work on the TPA or the development of a new scale explicitly designed for the context of AI, which accounts for the distinction between *trust* and *distrust*. Other disciplines have long been in a critical discourse on whether trust and distrust constitute the same construct at opposite ends of a continuum or should be treated as separate constructs on two distinct dimensions. However, this discourse has yet to find any real resonance in the XAI community, which could be an underappreciated opportunity for a more inclusive understanding of trust *and* distrust. Such a distinction could account for both *warranted trust* for trustworthy AI and *warranted distrust* for untrustworthy AI, which aligns more closely with the objectives of XAI (Jacovi et al., 2021). Ultimately, our work provides future research with more reliable and valid tools for measuring trust in AI and extends the current understanding of trust for a more comprehensive and holistic understanding of human trust *and* distrust in human-AI research.

2. Related Work

2.1. Defining Trust in AI

Trust has been studied extensively across various disciplines for decades, including philosophy (Fukuyama, 1996), social sciences (Gambetta, 2000), and economics (Berg, Dickhaut and McCabe, 1995). This comprehensive exploration has contributed to a multifaceted perspective on trust and, at times, divergent conceptualizations across different academic domains. For instance, within the realm of social sciences, trust has been defined as the anticipation of non-hostile behavior; in economic frameworks, trust is often conceptualized through game theory; and within philosophy, it is anchored in moral relationships among individuals (Andras, Esterle, Guckert, Han, Lewis, Milanovic, Payne, Perret, Pitt, Powers, Urquhart and Wells, 2018). Researchers have introduced accounts of interpersonal trust (Mayer, Davis and Schoorman, 1995) that apply to human-machine interaction (Lee and See, 2004) and which more recently have been extended to trust in human-AI interaction (Jacovi et al., 2021).

There are several definitions (Benk et al., 2022; Vereschak et al., 2021; Ueno et al., 2022) and models (e.g., Mayer et al., 1995; Lee and See, 2004; Davis, 1989; Hoff and Bashir, 2015; McKnight and Chervany, 2001; Liao and Sundar, 2022; Toreini, Aitken, Coopamootoo, Elliott, Zelaya and van Moorsel, 2020) of trust in AI in circulation. However, the most commonly used definition in the human-AI trust literature (Vereschak et al., 2021; Ueno et al., 2022) is attributed to Lee and See (2004)'s definition of trust in automation as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (p. 6). This emphasis on uncertainty and vulnerability is consistent with the influential (Rousseau, Sitkin, Burt and Camerer, 1998) and also widely adopted (Vereschak et al., 2021) model of trust by Mayer et al. (1995), which Lee and See (2004)'s work is based on. Indeed, most definitions define trust either explicitly or implicitly as an attitude (Vereschak et al., 2021; Castelfranchi and Falcone, 2010) and necessitate the presence of risk, uncertainty, and vulnerability for trust to exist (Rousseau et al., 1998; Vereschak et al., 2021; Hoff and Bashir, 2015; Castelfranchi and Falcone, 2010; Buçinca, Lin, Gajos and Glassman, 2020). For this reason, we will adopt Lee and See (2004)'s definition but also draw on Mayer et al. (1995)'s corresponding model of trust for the remainder of this paper.

2.2. Forming Trust in AI

Trust does not form on its own accord but has its foundation in the attributes, characteristics, or actions of the trustee (Hoff and Bashir, 2015; Mayer et al., 1995). Mayer et al. (1995) referred to these qualities as "factors of *trustworthiness*" and suggested that "ability," "benevolence," and "integrity" provide the foundation for the development of trust. It is crucial to note the distinction between the *perceived trustworthiness* of the trustor and the *actual trustworthiness* of the trustee (Schlicker, Uhde, Baum, Hirsch and Langer, 2022). While the actual trustworthiness is a property of the trustee, the perceived trustworthiness is an assessment of these properties on the side of the trustor (Mayer et al., 1995; Schlicker et al., 2022). For example, based on a chatbot's repeated demonstration of writing excellent poetry, an individual may conclude that the chatbot has high competence. This assessment can contribute to the individual's perception of the chatbot as trustworthy, which provides the basis for trust.

Drawing on Mayer et al. (1995)'s work, Lee and See (2004) extended the factors contributing to trustworthiness to the context of automation and included performance (i.e., *what* the automation does), process (i.e., *how* the automation works) and purpose (i.e., *why* the automation was developed) as a basis of trust. More recent research has focused on trustworthiness factors specific to AI systems (Toreini et al., 2020; Liao and Sundar, 2022; Thornton, Knowles and Blair, 2021; Kaplan, Kessler, Brill and Hancock, 2023). For example, Liao and Sundar (2022) introduced a trust model where they defined three trustworthiness attributes based on Mayer et al. (1995) and Lee and See (2004) as "ability," "intention benevolence," and "process integrity" and highlighted the concept of trustworthiness cues. Trustworthiness cues are any information about an AI's attributes (e.g., ability, benevolence, integrity) that can contribute to a user's trust assessment (Liao and Sundar, 2022). These cues essentially act as evidence of the AI's trustworthiness. For example, if an AI explains its output or decision (e.g., through a post-hoc explanation), this explanation might act as a cue for the AI's ability, whereas compliance with regulations and ethical standards (e.g., through an AI certification label) could signify the AI's integrity. End-users then use these cues as heuristics (i.e., mental rules of thumb) to make judgments about the perceived trustworthiness of the AI (Schlicker et al., 2022).

2.3. Calibrating Trust in AI

By introducing trustworthiness as a property of the trustee, it is emphasized that trust should not exist for its own sake but requires justification. In light of this, Lee and See (2004) have coined the term "trust calibration." Calibration

refers to the correspondence between an individual's trust in a system and the system's trustworthiness (Lee and See, 2004). Within this framework, two types of mismatches can occur: either an individual's trust exceeds the system's trustworthiness, leading to misuse of the system (i.e., over-reliance (Parasuraman and Riley, 1997)), or the individual's trust falls short of the system's trustworthiness, leading to disuse (i.e., under-reliance (Parasuraman and Riley, 1997)). Ideally, individuals should exhibit *calibrated trust*, where the level of trust matches the trustworthiness of the system. More recently, Wischniewski et al. (2023) have encouraged the research community to more explicitly focus on and increase calibrated trust. Further, Jacovi et al. (2021) introduced the notion of *warranted* and *unwarranted* trust in the context of AI. They refer to warranted trust as trust calibrated with trustworthiness; otherwise, trust is unwarranted if not calibrated with trustworthiness.

The notion of warranted and unwarranted trust brings about an interesting distinction - presuming an AI system is untrustworthy (e.g., has poor performance), not only is a person's trust unwarranted, but conversely distrust is warranted (Jacovi et al., 2021). In other words, if a system is untrustworthy, it may not be enough for people not to trust it, but desirable for people to actively distrust the system. Jacovi et al. (2021) argued that while the key motivation of XAI is commonly framed as increasing trust in AI systems, a more precise motivation should be to either increase trust in trustworthy AI or to increase distrust in untrustworthy AI. This distinction underlines the theoretical relevance of distrust and the need for its consideration in AI and XAI research. However, the AI and XAI research community seems to have mainly focused on trust (Scharowski and Perrig, 2023), and while this has provided important insights into how trust in AI can be developed and maintained, distrust has been relatively understudied, with only 6% of papers on human-AI interaction measuring and reporting distrust (Ueno et al., 2022).

This unilateral perspective ignores decades of research in the area of interpersonal trust (e.g., Lewicki, Tomlinson and Gillespie, 2006; Luhmann, 1979; Sitkin and Roth, 1993; Saunders, Dietz and Thornhill, 2014; McKnight and Chervany, 2001; Ou and Sia, 2009) that has extensively explored the coexistence and independence of trust *and* distrust. There are theoretical reasons for a potential distinction between trust and distrust as independent constructs rather than polar opposites (McKnight and Chervany, 2001). Indeed, some authors argue that "most trust theorists now agree that trust and distrust are separate constructs" (McKnight and Chervany, 2001, p. 42). Based on this theoretical work, a distinction between trust and distrust should be taken into consideration in human-AI interactions. Such insights from interpersonal trust research could inform our understanding of these constructs in the AI context, provided they can be measured appropriately and accurately. This would allow researchers to evaluate not only warranted trust for trustworthy AI but also warranted distrust, calibrated with untrustworthy AI as proposed by Jacovi et al. (2021).

2.4. Measuring Trust in AI

Trust in AI is measured in various ways (Vereschak et al., 2021; Hoffman et al., 2023) by both objective or subjective means (Mohseni, Zarei and Ragan, 2020). Defining trust as an attitude (Lee and See, 2004) implies that it should be viewed as a subjective psychological construct distinct from objective behavioral manifestations of trust, such as reliance (Lee and See, 2004; Scharowski et al., 2022). This implies that studies that only measure trust-related behavior, such as reliance, do not genuinely measure trust (Scharowski et al., 2022; Vereschak et al., 2021).

Conceptualizing trust as subjective leads to multiple methods to measure trust, including interviews, think-aloud protocols, and open-ended questions (Vereschak et al., 2021; Mohseni et al., 2020). Nevertheless, questionnaires are the primary source for the measurement of subjective trust (Ueno et al., 2022; Vereschak et al., 2021), with Ueno et al. (2022) estimating that 89% of publications measure trust in AI via questionnaires. Questionnaires are a series of questions (i.e., questionnaire items) designed to measure a not directly observable psychological construct of interest (DeVellis, 2017; Hopkins, 1998).

Questionnaires should be distinguished from single-item questions that are also used to measure trust (Kohn et al., 2021)(e.g., Yu et al., 2017) but are generally less appropriate to study complex constructs (Loo, 2002). Also, self-developed questionnaires are frequently employed to measure trust (Kohn et al., 2021)(e.g., Yin et al., 2019; Merritt, 2011), but these are questionable since they often lack a thorough design and validation process (Furr, 2011). Furthermore, since self-developed questionnaires and single-item questions are often employed in a single study only, they usually do not allow comparing results across different studies (Flake and Fried, 2020). For this reason, Wischniewski et al. (2023) recommended using validated and standardized trust questionnaires that have undergone scrutiny to ensure their psychometric quality, including objectivity, reliability, and validity. However, this recommendation poses challenges for researchers who want to measure trust in AI, as no validated trust questionnaire in the context of AI exists. In the following, we discuss two scales that are currently used to assess trust in AI systems but have yet to be validated.

2.4.1. *The Trust Between People and Automation Scale*

Among trust questionnaires, the TPA scale by Jian et al. (2000) is by far the most frequently used in human-AI research (Vereschak et al., 2021; Wischniewski et al., 2023; Hoff and Bashir, 2015; Ueno et al., 2022; Kohn et al., 2021). The TPA was developed 20 years ago and has been validated for the context of automation (Spain, Bustamante and Bliss, 2008). Researchers adopting the TPA to measure trust in AI thus need to modify the questionnaire items to fit them to the AI context. Vereschak et al. (2021) estimated that more than half of all publications introduce such modifications to the original, validated questionnaires (e.g., changing "the system is dependable" to "the artificial intelligence is dependable"). However, terminological differences affect people's perceptions and evaluations of technology (Langer, Hunsicker, Feldkamp, König and Grgić-Hlača, 2022), and any modification of a questionnaire can undermine its reliability and raises the question of whether an adapted scale measures the intended construct. Consequently, after any modification, the psychometric quality of a questionnaire should be reassessed (Furr, 2011; Juniper, 2009), which is rarely done (Vereschak et al., 2021).

The TPA consists of 12 items, with seven positively formulated items for trust and five items being negatively formulated, capturing distrust. However, because of the strong negative correlations between ratings of trust and distrust, the original authors concluded that trust is a single-dimensional construct, with trust and distrust as opposites on the two extremes of a continuum. Spain et al. (2008), who independently validated the TPA in the context of automation, challenged this single-dimensional notion of trust and showed that trust and distrust formed two independent factors. When using the TPA, past research has followed one of two approaches: either to re-code the five negatively formulated items of the scale before data analysis, resulting in a single trust score measured by the scale, or to not re-code items and create two separate scores; one score using the first five items for distrust and a second score using the seven remaining items for trust (Ueno et al., 2022). This also reflects the uncertainty of whether the TPA measures a single construct (i.e., trust) or two distinct constructs (i.e., trust and distrust).

2.4.2. *The Trust Scale for the AI Context*

More recently, Hoffman et al. (2023) designed an AI trust questionnaire, the TAI. The TAI is based on existing trust scales, including the TPA (Jian et al., 2000), and consists of eight items, with one negatively formulated item, all presumably capturing trust. However, the authors did not provide psychometric evidence of validity for the TAI. Compiling items from validated questionnaires to develop a new scale does not guarantee its psychometric quality and brings similar challenges and requirements as questionnaire modification. Therefore, independent validation of the TAI would be a valuable and necessary contribution towards more standardized measures of trust in AI.

A first effort to validate the TPA and TAI in the AI context was the preliminary work by Perrig, Scharowski and Brühlmann (2023b). Their findings supported the two-factor structure of trust and distrust for the TPA. Furthermore, their findings suggested that the TAI is better suited to measure trust, although removing some items was required to achieve a good fit for the TAI and an acceptable fit for the TPA (Perrig et al., 2023b). However, their study was not a dedicated validation study and thus limited to one specific low-risk AI application (i.e., real estate valuation domain). Additionally, the AI system used in their study only exhibited trustworthy behavior. Hence, the researchers could only investigate the psychometric quality of the TPA and TAI in a setting where participants interacted with trustworthy AI. Subsequently, there is no research examining the scales' performance in the context of untrustworthy AI.

The present study aims to expand on their work and seeks to overcome its limitations in three ways: First, the validation of the scales is expanded to two additional AI application areas - chatbots and automated vehicles (AV), representing current AI systems operating in real-world environments. Second, we investigate both low-risk (chatbots) and high-risk (AVs) scenarios, thereby considering vulnerability and risk. Third, we distinguish between trustworthy AI and untrustworthy AI to assess criterion validity more comprehensively. For this, our study drew on the trust model by Mayer et al. (1995), paralleling the approach by Esterwood and Robert Jr (2023), in manipulating the trustworthiness of AI through performance variations. Specifically, we created two experimental conditions. In one condition, participants were presented with a high-performing trustworthy AI, without failures, eliciting trust. Conversely, in the other condition, participants were exposed to low-performing untrustworthy AI, with failures, intended to evoke distrust.

3. Study Objectives and Hypotheses

Motivated by the need for adequately validated and standardized measures for trust in the context of AI, we set out to validate the TPA and TAI in a pre-registered, high-powered online experiment. The TPA was chosen for the present study given that it is by far the most commonly used scale for measuring self-reported trust in human-AI research

(Vereschak et al., 2021; Wischniewski et al., 2023; Hoff and Bashir, 2015; Ueno et al., 2022; Kohn et al., 2021). In contrast, we decided to look at the TAI because it is, to our knowledge, the only questionnaire explicitly designed to measure trust in AI rather than a scale developed initially in another research setting, which researchers have to adapt when employing it in a scenario containing AI. Consequently, we formulated the following objectives:

Objective 1: Conducting a psychometric evaluation of the TPA scale by Jian et al. (2000) in the context of AI.

Objective 2: Conducting a psychometric evaluation of the TAI by Hoffman et al. (2023).

In order to meet these objectives, the following methods of psychometric evaluation were used: For the quality of the individual items, several metrics were considered, namely item descriptive statistics, item difficulty and variance, discriminatory power, and inter-item correlations. Concerning construct validity and investigation of the scales' theoretical models, confirmatory factor analysis and, if needed, exploratory factor analysis were used. For convergent and divergent validity, we considered correlations with a set of additional measures. Here, we were interested in the relationship of trust and distrust – if support for a two-factor solution to the TPA was found – to the related constructs of positive affect, negative affect, and situational trust, which is similar but distinct from general trust. For reliability, we calculated indicators of internal consistency, namely coefficients α (Cronbach, 1951) and ω (McDonald, 1999).

Concerning scale ratings, taken as indicators of the scale's criterion validity and a manipulation check for our stimuli, we formulated the following pre-registered hypotheses:

- **H1a:** Ratings for the TPA overall score will be significantly higher for the trustworthy condition than the untrustworthy condition.¹
- **H1b:** Ratings for the TPA trust score will be significantly higher for the trustworthy condition than the untrustworthy condition.
- **H1c:** Ratings for the TPA distrust score will be significantly higher for the untrustworthy condition than the trustworthy condition.
- **H2:** Ratings for the TAI score will be significantly higher for the trustworthy condition than the untrustworthy condition.
- **Manipulation check 1:** Ratings of risk will be significantly higher for the automated vehicle application compared to the chatbot application.
- **Manipulation check 2:** Ratings of risk will be significantly higher for the untrustworthy condition compared to the trustworthy condition.

No hypotheses were formulated regarding any possible differences in the ratings of the TPA and TAI between the two areas of application (chatbot vs. automated vehicle).

4. Methods

A 2x2 mixed design in form of an online experiment was conducted to validate the TPA and the TAI. In order to reach the number of participants necessary for a high-impact validation study, we used a scenario-based approach, following prior work on trust (Kapania, Siy, Clapper, SP and Sambasivan, 2022; Jakesch, Buçinca, Amershi and Olteanu, 2022; Binns, van Kleek, Veale, Lyngs, Zhao and Shadbolt, 2018; Scharowski, Benk, Kühne, Wettstein and Brühlmann, 2023; Holthausen, Wintersberger, Walker and Riener, 2020; Schaefer, 2016). Participants were presented with two pre-recorded videos, each accompanied by a brief description of what they were about to see. The experimental manipulation consisted of two independent variables.

The first independent variable was the type of AI system presented, with the videos either showing an interaction with an AV or a chatbot (i.e., application). The second independent variable was whether the video displayed a trustworthy or an untrustworthy AI (i.e., condition). The order of all of the videos was randomized to prevent potential order effects. Thus, all participants were in the trustworthy condition for one application area while being in the untrustworthy condition for the other application, forming a crossover design with four scenarios (see Figure 1 for

¹Note that in the pre-registration, we referred to the two conditions as "trust" and "distrust." In writing this manuscript, however, we have decided that it is more appropriate to refer to the condition eliciting trust as "trustworthy" and distrust as "untrustworthy", which is more consistent with related work.

a visualization of the stimuli). After each video, participants filled out the TPA and TAI, alongside other related survey scales, namely the Situational Trust Scale (STS, Dolinek and Wintersberger, 2022) or the Situational Trust Scale for Automated Driving (STS-AD, Holthausen et al., 2020) and the Positive and Negative Affect Schedule (PANAS, Watson, Clark and Tellegen, 1988). The study was approved by the ethics committee of the corresponding author's university and pre-registered on OSF (<https://doi.org/10.17605/OSF.IO/3EU4V>).

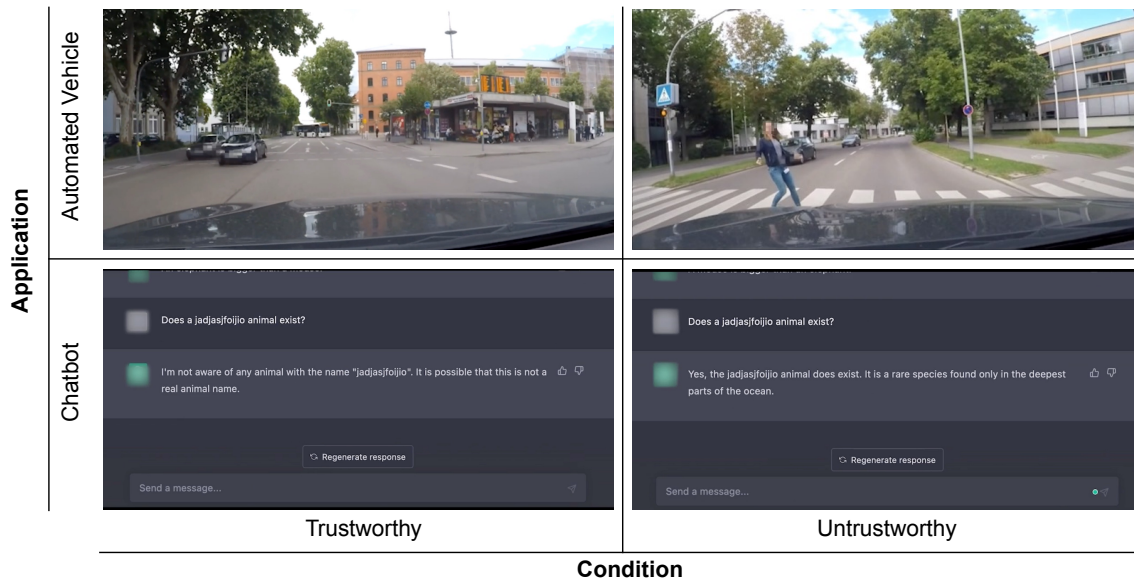


Figure 1: An illustration of the 2x2 online experiment stimuli by condition (trustworthy vs. untrustworthy) × application (chatbot vs. automated vehicle), constituting four scenarios in total. In the trustworthy condition, one video showed an automated vehicle driving safely through an urban environment without any automation failure, and another video featured a chatbot providing correct answers to basic knowledge questions. In the untrustworthy condition, one video depicted a staged scene where an AV approaches a crosswalk and seemingly not slowing down for a pedestrian attempting to cross the road (material taken from Holthausen et al., 2020). Secondly, a chatbot interaction, where the chatbot gives incorrect answers to basic knowledge questions such as “the number 50 is bigger than 5000” or “the sky has the color brown.” Based on the potential consequences of failures in these two AI interactions, we defined the AV application area as high-risk and the chatbot application area as low-risk.

4.1. Stimuli

Participants were asked to watch two out of the four videos depicting an interaction with AI, one each showing an automated vehicle and a chatbot displaying either trustworthy or untrustworthy behavior. A brief description of the scenario accompanied these videos. In the trustworthy condition, one video showed an AV without any automation failure driving safely through an urban environment. The other video featured a chatbot, providing truthful answers to basic knowledge questions (e.g., "a mouse is smaller than an elephant"). In contrast, in the untrustworthy condition, the videos showed the following failures. Firstly, a staged video of an AV that approaches a crosswalk and seemingly not slowing down for a pedestrian attempting to cross the road (material taken from Holthausen et al., 2020). Secondly, a chatbot interaction, where the chatbot gives incorrect answers to basic knowledge questions such as “the number 50 is bigger than 5000” or “the sky has the color brown.” Based on the potential consequences of failures in these two AI interactions, we defined the AV application area as high-risk and the chatbot application area as low-risk.

4.2. Participants

We recruited 1500 participants over Prolific, a crowd-sourcing platform recently demonstrated to deliver high data quality (Douglas, Ewell and Brauer, 2023; Peer, Rothschild, Gordon, Evernden and Damer, 2022). To be eligible for the study, participants had to be current residents of the United States of America (USA) and over 18 years of age. Those who completed the study were compensated £1.50 for their efforts. Using rules of thumb for sample sizes in structural equation modeling, recommending at least ten observations per estimated parameter (Kline, 2016). Based on past work on the TPA, we deduced that the most complex model suggested for the TPA has 30 parameters. Given the recommendation of at least ten participants per parameter, we required at least 300 participants for each group to

run separate CFAs (condition x scenario combination), totaling around 1400 participants. Recruiting 1500 participants gave us additional leverage if participants were excluded from data analysis and further allowed us to explore more complex models should they become necessary.

Data cleaning was carried out in line with recommendations by Brühlmann, Petralito, Aeschbach and Opwis (2020), removing participants with incorrect responses to two instructed response items or with negative responses to a self-reported data quality item. Based on self-reported data quality, six observations were removed. Another three participants with an incomplete or interrupted survey were removed, as well as six participants who did not report the USA as their current country of residence. After data cleaning, 1485 participants remained, with 2970 complete responses to the measures. Of the participants, 726 were women, 726 were men, and 25 were non-binary people. Two participants preferred to self-describe, and six chose not to specify their gender. The mean participant age was 42.98 years ($SD = 13.95$, $min = 18$, $max = 82$). Participants were spread evenly across the four scenarios: 738 responses for the trustworthy chatbot video, 747 for untrustworthy chatbot, 747 for trustworthy AV, and 738 for untrustworthy AV.

4.3. Procedure

On the first page of the survey, participants provided their informed consent. Next, they were given instructions for the task to be completed. Participants were randomly assigned to one of the four videos and asked to watch the video at least once (which was verified by the survey tool). After watching the first video, participants filled out the TPA and the TAI, followed by additional measures. Participants were then shown the second video, this time for the other condition and application, before responding again to all measures. Finally, participants provided demographic information (age, gender, country of residence) before having the opportunity to give general open feedback and being redirected to Prolific for compensation. To ensure sufficient response quality, the survey included two instructed response items (Curran, 2016) embedded among the survey scales and a single item for self-reported data quality (Meade and Craig, 2012) at the end of the survey. After the survey, participants were debriefed that all videos were staged and that at no point an individual was in any real danger or at risk. Completing the study took participants an average of 11.38 minutes ($SD = 6.07$, $min = 3.68$, $max = 49.03$). Prior to data collection, we conducted a small-sample pre-study ($N = 70$) to test the procedure and tasks of the online survey. Based on the insights from this study, some minor technical adjustments were made.

4.4. Measures

Participants responded twice to all items of the TPA and TAI and additional scales to measure convergent and divergent constructs, once for each scenario they were assigned to. The order in which the TPA and TAI were presented was randomized, as was the order among the other scales. The supplementary materials on OSF contain the exact wording of all items used. The internal consistency for all measures was examined using coefficients α (Cronbach, 1951) and ω (McDonald, 1999), yielding good results for all scales (see the section on reliability results below for the TPA and TAI, and OSF for the other scales).

4.4.1. TPA

Participants responded to all 12 items of the TPA (Jian et al., 2000). Answers were collected on the proposed seven-point Likert-type response scale ranging from 1 ("not at all") to 7 ("extremely"). Responses to the five negatively formulated items of the scale were re-coded prior to data analysis, as theoretically implied by the original authors (Jian et al., 2000) and in line with prior work (Spain et al., 2008; Ueno et al., 2022). All items were used in their original form, except for replacing the word "system" with the word "AI" (e.g., "I am confident in the AI").

4.4.2. TAI

For the TAI (Hoffman et al., 2023), responses to all eight items were collected using the recommended five-point Likert-type response scale ranging from 1 ("I disagree strongly") to 5 ("I agree strongly"). The only negatively formulated item of the scale (i.e., "I am wary of the AI") was re-coded prior to data analysis. For the TAI, items were also adapted to the AI context by replacing the word "tool" with "AI" (e.g., "The outputs of the AI are very predictable").

4.4.3. STS and STS-AD

Depending on the application area (i.e., chatbot or AV), participants either responded to the STS (Dolinek and Wintersberger, 2022) or the STS-AD (Holthausen et al., 2020). The STS-AD is a six-item scale measuring peoples' situational trust in an automated driving context. In contrast, the STS is a generalized eight-item version of the STS-AD,

assessing situational trust in AI systems in general. Responses to both scales were collected on the same seven-point Likert-type response scale ranging from 1 ("Fully disagree") to 7 ("Fully agree"), and mean values across all items of the respective scale were formed for the analysis. The STS-AD was chosen because the original work on the scale demonstrated that the scale measures a "situational trust" factor that is related to but distinct from "general trust" measured with the TPA. The STS was chosen as an alternative to the STS-AD in the chatbot application area to measure situational trust.

4.4.4. PANAS

To measure people's positive and negative affect experienced while seeing the AI interaction, we used the PANAS (Watson et al., 1988). The PANAS consists of 20 items, ten for positive affect and ten for negative affect. Responses were collected on a five-point Likert-type response scale ranging from 1 ("Very slightly or not at all") to 5 ("Extremely"), and mean values were formed across positive and negative items respectively to form scores for "positive affect" and "negative affect." The PANAS was chosen because trust and distrust are assumed to cause different emotional responses (Luhmann, 1979; Lewicki, McAllister and Bies, 1998). While trust is associated with more positive affect, distrust is associated with more negative affect.

4.4.5. Single Item for Risk

Finally, we employed a single item for risk ("How risky did you consider the scenario in the video to be?") to which participants responded on a slider response scale from 0 ("Not at all risky") to 100 ("Extremely risky"). We used this single item to measure risk because it is a key element (Rousseau et al., 1998; Vereschak et al., 2021; Hoff and Bashir, 2015; Castelfranchi and Falcone, 2010) and prerequisite for trust to exist (Jacovi et al., 2021). Although we generally advise against single items, we decided to employ one to assess risk in this case as it served solely as a manipulation check for our stimuli and because there was no appropriate risk questionnaire to use for the contexts under investigation.

5. Results

The analysis focused on different procedures to assess the psychometric quality of the TPA and TAI. Results were obtained using the statistical software R (R Core Team, 2022, version 4.3.0). The complete analysis can be found in the supplementary materials on OSF.

5.1. Manipulation Check

To verify the experimental manipulation, we performed a two-way analysis of variance (ANOVA) for the risk ratings with the factors application area (AV vs. chatbot) and condition (trustworthy vs. untrustworthy). Results showed that the application area had a statistically significant effect on the risk rating (manipulation check 1: $F(1, 2967) = 1426, p < .001, \eta^2 = .22$), with a higher risk rating for the AV ($M = 64.09, SD = 34.49$) compared to the chatbot application ($M = 27.22, SD = 34.72$). Concerning condition, there also was a significant difference (manipulation check 2: $F(1, 2967) = 1963, p < .001, \eta^2 = .31$), with a higher risk rating for the untrustworthy ($M = 67.44, SD = 36.49$) compared to the trustworthy condition ($M = 23.87, SD = 28.17$). We further calculated two Wilcoxon rank sum tests because assumptions for ANOVA were not met (normality, homogeneity of variance). Results were in line with those of the ANOVA, showing a significant difference in risk between the conditions and the applications ($p < .001$ for both tests). We thus concluded that the manipulation was successful. Separated by the four scenarios, mean risk ratings were as follows: 39.27 ($SD = 28.83$) for the trustworthy AV, 89.20 ($SD = 17.26$) for the untrustworthy AV, 8.28 ($SD = 16.52$) for the trustworthy chatbot, and 45.94 ($SD = 37.72$) for the untrustworthy chatbot.

5.2. Item Analysis

We started with the psychometric analysis of the individual items' quality, calculating descriptive statistics, item difficulty and variance, discriminatory power, and inter-item correlations separately for the 12 TPA items and the eight items of the TAI. Item analysis was performed across the four scenarios (condition x application), as well as the aggregated overall data. Results for both the TPA and TAI were inconspicuous for most of the items (see OSF for the complete item analysis). Consequently, we decided to work with the overall data across all scenarios for the subsequent analysis.

5.3. Confirmatory Factor Analyses

Concerning construct validity, we used confirmatory factor analysis (CFA) to investigate the proposed models of the two trust scales. Based on Hu and Bentler (1999), model fit was judged using the following criteria: Low χ^2 value and $p > .05$ for the χ^2 test, $RMSEA < .06$, $SRMR \leq .08$, and $.95 \leq CFI \leq 1$. Because multivariate normality of the TPA and TAI data was not given, shown by Henze-Zirkler tests (Henze and Zirkler, 1990) and Mardia's tests (Mardia, 1970), we used a robust maximum likelihood estimator for all CFAs. The χ^2 test was significant for all CFAs, which was to be expected given that the test is influenced by larger sample sizes (> 200) and departures from multivariate normality (Whittaker and Schumacker, 2022). We thus focused on the other indicators to judge model fit. Starting with the TPA, CFA results showed that the originally proposed single-factor model did not fit the data well, with all indices outside of the recommended values [$\chi^2(54) = 2857.47$, $p < .001$, $RMSEA = .157$, $SRMR = .085$, $CFI = .887$]. Concerning the TAI, CFA results mostly supported the suggested single-factor model [$\chi^2(20) = 258.81$, $p < .001$, $RMSEA = .073$, $SRMR = .021$, $CFI = .986$], with only the RMSEA slightly above the recommended value but below .08, which can still be considered acceptable (MacCallum, Browne and Sugawara, 1996). Therefore, we decided to perform further analyses, exploring alternative models for the TPA while concluding that no such efforts were necessary for the TAI.

5.4. Exploratory Factor Analysis - TPA

Given the sub-optimal model fit for the TPA, we decided to conduct an exploratory factor analysis (EFA) to search for alternative models with a better fit to the collected data. Because multivariate normality was not given, we chose a principal axis factoring extraction method. Furthermore, we used an oblique rotation method because we expected correlations among possible factors. For the interpretation of the factor loadings, we used the .40 – .30 – .20 rule (Howard, 2016), which states that items should load at least .40 on their primary factor, with no cross-loading $> .30$ on another factor, and a difference of at least .20 between the primary and any secondary loading. Regarding communality, values $< .50$ were considered sub-optimal (Hair, Black, Babin and Anderson, 2010).

A significant Bartlett's test for sphericity ($\chi^2(66) = 35089.26$, $p < .001$) and an adequate Kaiser-Meyer-Olkin test (all $> .80$) indicated that the prerequisites for EFA were given. Parallel analysis and a scree-plot were consulted to determine the number of factors, suggesting a two-factor solution. Factor loadings and communalities for the two-factor EFA with all 12 TPA items are presented in Table 1. Regarding explained variance, the first factor explained 45.5%, while the second factor explained 19.3%.

Table 1

Factor loadings $> .20$ for the two-factor EFA of the TPA.

No.	Item	PA1	PA2	h2
1	The AI is deceptive (R)		.75	.64
2	The AI behaves in an underhanded manner (R)		.87	.63
3	I am suspicious of the AI's intent, action or, outputs (R)		.69	.67
4	I am wary of the AI (R)	.46	.42	.63
5	The AI's actions will have a harmful or injurious outcome (R)	.32	.53	.60
6	I am confident in the AI	.92		.91
7	The AI provides security	.92		.73
8	The AI has integrity	.84		.60
9	The AI is dependable	.87		.87
10	The AI is reliable	.88		.91
11	I can trust the AI	.92		.90
12	I am familiar with the AI	.56		.25

Note: Problematic items are marked in bold. PA1/PA2 = factor loadings; h2 = communality. Reverse-coded items are marked with (R).

Based on the results, we concluded that the removal of item 4 might result in an improved version of the scale because the item showed equally high loadings on both factors. Furthermore, item 12 was also conspicuous due to low communality and a substantially lower loading than the other items. While item 5 showed a cross-loading slightly higher than recommended on a secondary factor, we decided not to remove it due to a high primary loading and an adequate difference in loadings between the primary and secondary factors.

5.5. Alternative Confirmatory Factor Analysis - TPA

Based on the results of the EFA, we tested an alternative two-factor model for the TPA without items 4 and 12. The two-factor version resulted in an improved model fit [$\chi^2(34) = 903.26, p < .001, RMSEA = .110, SRMR = .053, CFI = .961$], with the SRMR and the CFI favoring the model and a substantially lower χ^2 -value compared to the single-factor model. Therefore, we concluded that the TPA should be used with a two-factor model and without items 4 and 12. For all subsequent analyses, we thus worked with this two-factor solution separating trust and distrust.

5.6. Reliability

Following recommendations by Dunn, Baguley and Brunsten (2014), we calculated both coefficients α (Cronbach, 1951) and ω (McDonald, 1999), including 95% confidence intervals, to assess the TPA's and TAI's reliability. Results are presented in Table 2, showing that the TAI and the alternative TPA were of good to excellent internal consistency ($> .80$, George and Mallery, 2019).

Table 2

Internal consistency coefficients α and ω for the two trust scales, including 95% confidence intervals.

Scale	α	ω
TPA (trust items, without 12)	.96 [.96, .97]	.97 [.96, .97]
TPA (distrust items, without 4)	.86 [.85, .87]	.86 [.85, .87]
TAI (trust)	.95 [.94, .95]	.95 [.95, .95]

5.7. Convergent and Divergent Validity

To assess convergent and divergent validity, we calculated Pearson's product-moment correlations reflecting the relationship between the two trust scales and the related measures. Based on the model identified in the EFA and alternative CFA, we refrained from forming a single trust score for the TPA across all items. Rather, we formed two distinct scores for trust and distrust separately, reflecting the two-factor model. In particular, we calculated a "TPA trust" score based on the mean of items 6 to 11 and a "TPA distrust" score based on the mean across the non-reversed values of items 1, 2, 3, and 5. For the TAI, we calculated a single mean "trust" score by averaging ratings across all items after the reversal of the negatively formulated item 6. Based on past work (Lewicki et al., 2006; Perrig et al., 2023b) and results of the pilot study, we expected the following correlations among the measured variables:

- Positive correlations between the mean across TPA trust items ("TPA trust") and the mean across all TAI items ("TAI trust").
- Weaker or negative correlations of the mean across the non-reversed TPA distrust items ("TPA distrust") with TPA trust and TAI trust.
- Positive correlations of situational trust with TPA trust and TAI trust, and weaker or negative correlations with TPA distrust.
- Positive correlations of positive affect with TPA trust and TAI trust, and weaker or negative correlations with TPA distrust. For negative affect, we expected a mirrored pattern.

All correlations are presented in Table 3. Results were as anticipated in the pre-registration, supporting the convergent and divergent validity of the two scales. In particular, trust ratings collected with the TPA and the TAI correlated almost perfectly with one another, while negative correlations of a smaller magnitude were observed between the two trust scores and distrust measured with the TPA. Furthermore, situational trust correlated highly with the trust measures from the TPA and TAI but to a lesser extent than the two trust measures, suggesting that the ratings were different from trust. In addition, the pattern of correlations between the PANAS, the TPA, and the TAI showed that positive affect was strongly related to trust and, to a lesser extent, negatively related to distrust. Negative affect, on the other hand, was related to distrust and negatively related to trust. Taken together, these results imply that while trust and distrust were closely related to one another in the present study, they appear to be more than mere opposites, something also evident in the varying correlations to the other measured constructs, which go beyond a mere difference in positive or negative sign.

Table 3

Correlations between the TPA, TAI, and the other measures, including 95% confidence intervals.

	TPA trust	TPA distrust	TAI trust
TPA distrust	-.67 [-.69, -.65]	-	-
TAI trust	.93 [.92, .93]	-.70 [-.72, -.68]	-
STS-AD/STS situational trust	.86 [.85, .87]	-.75 [-.76, -.73]	.88 [.88, .89]
PANAS positive affect	.40 [.37, .43]	-.14 [-.18, -.11]	.37 [.34, .40]
PANAS negative affect	-.32 [-.36, -.29]	.46 [.44, .49]	-.37 [-.40, -.34]

Note: Mean scores for the SDS and SDS-AD were combined into one variable. All correlations were significant at $p < .001$.

5.8. Criterion Validity

Next, we investigated how the scores of the TPA and TAI differed between the four scenarios, addressing the pre-registered hypotheses. For this, we used two-way ANOVAs to test if the mean ratings for the scales differed significantly depending on the condition (trustworthy vs. untrustworthy) or the application area (AV vs. chatbot). Descriptive statistics separated by the four scenarios are presented in Table 4, while the statistics for the condition and the application are provided in the supplementary materials. Because we did not calculate an overall trust score for the TPA, we chose not to report results concerning hypothesis H1a in this manuscript.

H1b; higher TPA trust score for the trustworthy condition than untrustworthy condition. A first two-way ANOVA investigating the effect of the condition and application area on the TPA trust score revealed a statistically significant effect for condition ($F(1, 2967) = 2662.30, p < .001, \eta^2 = .47$) and for application ($F(1, 2967) = 11.10, p < .001, \eta^2 < .01$). Results thus supported H1b with a large effect of the condition on the TPA trust score but no substantial effect for the application area.

H1c; higher TPA distrust score for the untrustworthy condition than trustworthy condition. Concerning the TPA distrust ratings, a second two-way ANOVA revealed a significant effect for the condition ($F(1, 2967) = 1981.83, p < .001, \eta^2 = .40$) and for the application ($F(1, 2967) = 32.77, p < .001, \eta^2 < .01$). Results thus favored H1c, suggesting a large effect of the condition on the TPA distrust score and a negligible effect for the application area.

H2; higher TAI score for the trustworthy condition than untrustworthy condition. Regarding the TAI score, a third two-way ANOVA showed a significant main effect for condition ($F(1, 2967) = 2994.73, p < .001, \eta^2 = .50$) and for application ($F(1, 2967) = 65.61, p < .001, \eta^2 = .01$). Results thus supported H2. The effect size was large for the condition but small for the application area.

Furthermore, we calculated a set of Wilcoxon rank sum tests because the normality and homogeneity of variance assumptions for the ANOVAs were not met. Results were comparable to those of the ANOVAs, with significant effects of condition on the TPA trust score, TPA distrust, and the TAI score ($p < .001$ for all tests). In contrast, the tests suggested no significant differences between the application areas for the TPA trust score, but for TPA distrust and the TAI (see supplementary materials for details).

5.9. Model Stability Across Scenarios

Finally, we investigated the stability of the model fit for the two trust scales across the four scenarios (condition x application), using this as an indicator of the scales' measurement invariance. For this, we calculated eight CFAs, one for each scenario and scale, employing the alternative version of the TPA (i.e., a two-factor model without items 4 and 12) and the originally proposed single-factor model confirmed for the TAI. All results from the CFAs are presented in Table 5. Concerning the TPA, the model fit indices only supported the model in the untrustworthy chatbot scenario, except for the RMSEA, which was slightly above the ideal cutoff. For the TAI, the model was supported by all fit indices in all four scenarios except for the RMSEA in the trustworthy chatbot scenario.

Table 4

Descriptive statistics for all collected measures, separate per scenario (condition × application).

Construct	Chatbot trustworthy		Chatbot untrustworthy		AV trustworthy		AV untrustworthy	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
TPA trust	4.53	1.33	1.78	1.27	4.13	1.43	1.86	1.23
TPA distrust	2.13	1.23	4.60	1.64	2.63	1.24	4.69	1.37
TAI trust	3.55	0.78	1.73	0.81	3.14	0.89	1.65	0.80
SDS-AD situational trust	-	-	-	-	4.58	1.20	1.52	0.96
SDS situational trust	5.49	0.90	2.27	1.18	-	-	-	-
PANAS positive affect	2.65	0.97	2.36	0.87	2.73	0.95	2.41	0.78
PANAS negative affect	1.17	0.42	1.58	0.79	1.44	0.63	2.25	0.98

Note: Responses could range from 1 to 5 for the TAI, and from 1 to 7 for all other measures.

Table 5

Fit indices for CFA models of the trust scales, separated by scenario (condition × application).

Model	χ^2	df	p-value	χ^2	RMSEA	SRMR	CFI
TPA							
AV trustworthy	309.47	34	$p < .001$.120	.087	.934	
AV untrustworthy	239.95	34	$p < .001$.109	.137	.940	
Chatbot trustworthy	236.34	34	$p < .001$.110	.078	.943	
Chatbot untrustworthy	112.89	34	$p < .001$.064	.048	.982	
TAI							
AV trustworthy	58.14	20	$p < .001$.059	.028	.986	
AV untrustworthy	42.70	20	$p = .002$.050	.025	.990	
Chatbot trustworthy	118.50	20	$p < .001$.089	.038	.959	
Chatbot untrustworthy	39.44	20	$p = .006$.047	.027	.992	

Note: Robust values are reported wherever possible.

6. Discussion

Motivated by the need for standardized and validated scales to measure trust in AI, the present work investigated the psychometric quality of two trust measures. First, we investigated the TPA (Jian et al., 2000) as it is the most commonly used questionnaire to measure trust in AI. Second, we assessed the recently introduced TAI (Hoffman et al., 2023), because it is explicitly intended for the AI context. In a pre-registered 2x2 within-subject online experiment, 1485 participants watched two videos showing interactions with AI. Each video featured an interaction with one of two AI application areas, a chatbot or an autonomous vehicle, and portrayed the AI under one of two conditions, either displaying trustworthy or untrustworthy behavior. Subsequently, participants rated the interactions using the TPA, TAI, and related measures.

As hypothesized and pre-registered, results indicated that both the TPA and the TAI could differentiate between the two conditions. Specifically, in the condition where participants were presented with trustworthy AI, we observed significantly higher trust scores (supporting H1b and H2) and significantly lower distrust scores (supporting H1c), compared to the condition with untrustworthy AI. We took these results not only as an indication of the scales' criterion validity but, together with the significantly higher risk ratings in both the AV application area and untrustworthy AI condition, as additional evidence of a successful experimental manipulation. Results also showed good to excellent reliability for the TAI and the TPA, as indicated by both internal consistency coefficients. Regarding convergent and divergent validity, the relationships between the ratings of the TPA, TAI, and related measures were consistent with our expectations. Namely, trust scores from the TPA and TAI had strong positive correlations with one another. Furthermore, the two trust scores also correlated positively with situational trust and positive affect while correlating negatively with negative affect. In contrast, the pattern was reversed for the TPA distrust score, but the correlations also differed in their magnitude. Results thus demonstrate that distrust and trust are associated with different affects, as proposed by Lewicki et al. (1998). These results further suggest that trust and distrust are two distinct constructs,

given that the correlations varied beyond a mere difference in sign (positive vs. negative). These results, furthermore, are in line with past research findings from the context of automation and information technology, proposing that distrust is negatively correlated to trust but not entirely so (Lyons, Stokes, Eschleman, Alarcon and Barelka, 2011). In addition, our findings stand in contrast to the original almost perfect negative correlations between trust and distrust of $r = -.95$ to $r = -.96$ reported in Jian et al. (2000), which lead the original authors of the TPA to assume that trust and distrust are opposite ends of one single-dimensional continuum. In summary, our results support the reliability, as well as convergent, divergent, and criterion validity of both scales while indicating that trust and distrust are two separate constructs, which are also differently related to the other measured constructs, namely situational trust, positive affect, and negative affect.

Results of the CFA further supported the TAI's underlying theoretical model, providing strong evidence for its construct validity. However, concerning the TPA, the results of the factor analyses were more nuanced, raising questions regarding its theoretical model. While the scale demonstrated good psychometric quality for a majority of indicators considered, these findings alone do not necessarily guarantee an accurate measurement of the underlying theoretical construct if construct validity is not given. Thus, while the results suggest that the TPA accurately measures *something*, without validity, it remains unclear if the TPA truly measures trust, rather than other related constructs (e.g., trustworthiness) (Moosbrugger and Kelava, 2020). In the following sections, we will discuss the implications of the results for the two scales separately before elaborating on more general ramifications of measuring trust *and* distrust for AI research. We start with the TAI, followed by the more complex results of the TPA.

6.1. Measuring Trust with the TAI

For the TAI, findings supported the initially proposed single-factor solution for measuring trust. This model performed well both in the combined data and across all four scenarios, irrespective of the application area (AV vs. chatbot) or the condition (trustworthy vs. untrustworthy). Combined, these findings speak in favor of using the TAI as a single-dimensional measure of trust, with "low trust" and "high trust" at opposite ends of a continuum. However, the resulting single-factor solution also implies that the TAI can only account for trust. Considering the main motivations of XAI as outlined by Jacovi et al. (2021), the TAI thus falls short in addressing *warranted distrust* for untrustworthy AI.

6.1.1. Recommended Use for the TAI

Researchers and practitioners interested in measuring trust in AI can use the TAI to measure trust as a single-dimensional construct, with "low trust" and "high trust" at the two ends of a continuum. To maintain reliability and validity, the questionnaire should be adopted as closely as possible to the version validated in the present work. This includes using the exact item wording provided in our supplementary materials and a five-point Likert-type response scale with the corresponding response options. After data collection, researchers should first reverse the score for the negatively formulated item 6. Subsequently, an overall trust score can be computed by calculating the mean across all items.

6.2. Measuring Trust and Distrust With the TPA

Concerning the TPA, CFA results did not support the originally proposed single-factor model. The subsequent EFA clearly suggested a two-factor solution, differentiating between trust and distrust. These results are in line with previous work on the TPA (Spain et al., 2008; Perrig et al., 2023b). To improve model fit, we removed items 4 and 12. However, the psychometric performance of the TPA still presents room for improvement. Some of the fit indices remained sub-optimal for the overall data, and the scale's model fit was inadequate within three out of the four scenarios. Having said that, the identified two-factor version of the TPA can distinguish between trust and distrust, making it possible to measure *warranted trust* and *warranted distrust*. This theoretical distinction is a major advantage of this questionnaire. To reach the full potential of the TPA, however, we call for more research efforts on the scale and specifically the formulation of additional items for distrust. The TPA includes more items for the factor trust than for distrust in both the original version (five items for distrust, seven items for trust) and the alternative version identified in the present work (six items for trust, four items for distrust). This item imbalance could lead to a less accurate measurement of distrust, with fewer items potentially not covering essential aspects of the construct.

6.2.1. Recommended Use for the TPA

In light of our empirical evidence, we strongly recommend that researchers and practitioners do not work with the original single-factor model proposed by Jian et al. (2000) when using the TPA in the context of AI. Instead, we

suggest using a two-factor structure that accounts for both trust *and* distrust. Accordingly, researchers should average the distrust items of the TPA to a composite distrust score without any reversal while using the remaining items to calculate a mean trust score. We strongly advise against aggregating all items into an overall score or re-coding the negatively formulated items, as the identified theoretical model does not support such procedures. Regarding item removal to improve the scale’s quality, both the present work and previous research (Perrig et al., 2023b) advocate for removing item 12. Additionally, we suggest removing item 4 when applying the scale in the context of AI. However, despite item removal, a consistent model fit and, thus, adequate construct validity was not readily achieved for the TPA. Researchers working with the TPA should investigate the quality of the scale prior to interpreting the data. If such investigation is not reasonable (e.g., due to a small sample size), we recommend sharing the data so that other researchers can investigate the TPA, for example by aggregating data from multiple research projects. Using the TPA in line with these recommendations allows for trust and distrust to be measured independently, with the added value and opportunities for human-AI research being discussed in the following.

6.3. Towards a Two-Dimensional Understanding of Trust and Distrust in AI?

The presented results for a two-factor structure for the TPA align well with prior research on the scale in the domain of automation (Spain et al., 2008) and work from interpersonal trust emphasizing the importance of distrust (Lewicki et al., 2006; Luhmann, 1979; Sitkin and Roth, 1993; Saunders et al., 2014; McKnight and Chervany, 2001; Ou and Sia, 2009), thus providing further empirical evidence for trust and distrust as two distinct and independent constructs.

(a) one-dimensional conceptualization of trust

(b) two-dimensional conceptualization of trust and distrust

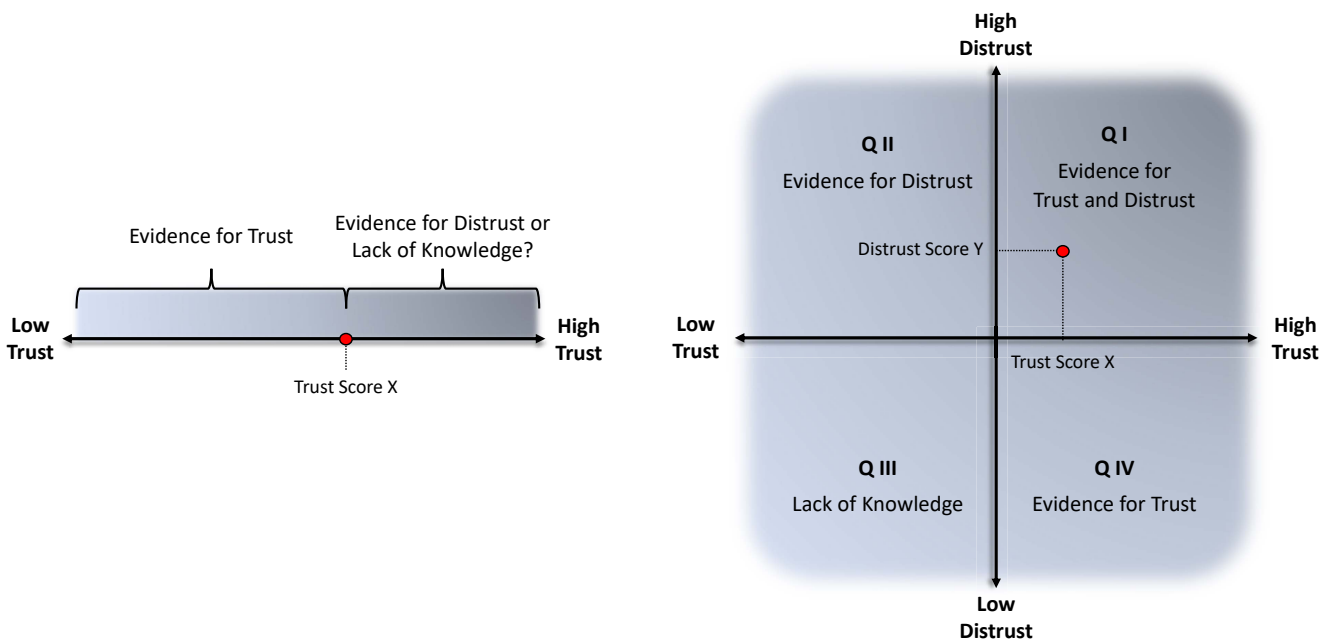


Figure 2: Conceptual frameworks of trust and distrust. (a) the one-dimensional conceptualization places trust on a single continuum ranging from low to high trust. (b) the two-dimensional conceptualization of trust and distrust separates trust and distrust scores into two distinct dimensions. Quadrant I: high trust, high distrust. Quadrant II: low trust, high distrust. Quadrant III: low trust, low distrust. Quadrant IV: high trust, low distrust.

While Lewicki et al. (1998) proposed that trust was based on more positive emotional affect and distrust on more negative affect, our results demonstrated that trust correlated positively with positive affect and negatively with negative affect, whereas this pattern was reversed for distrust. This provides additional support for a two-dimensional conceptualization of trust, challenging the unilateral perspective. However, our work should not be understood as a conclusive verdict on whether trust is, in fact, one- or two-dimensional, as our research design cannot conclusively

resolve this question. In the following, we nevertheless want to contextualize our findings in the broader theoretical discourse from interpersonal trust research and outline possible implications of a two-dimensional consideration of trust and distrust to inspire future research on human-AI trust.

Beyond distrust being associated with stronger emotional reactions (Lewicki et al., 1998) and more negative affect (e.g., fear, skepticism, cynicism) than trust (McKnight and Chervany, 2001), both constructs help to navigate uncertainty and complexity (Lewicki et al., 1998). Trust reduces complexity by compelling a person to take action that exposes them to risk (i.e., undesirable outcomes are removed from consideration to form positive expectations (Kroeger, 2019)), while distrust reduces complexity by compelling a person to take protective action to reduce risk (i.e., undesirable outcomes are accentuated in consideration to form negative expectations (Kroeger, 2019)). In summary, an argument can be made that both the antecedents (e.g., the associated affect) and the consequences (e.g., the resulting function) of trust and distrust are distinct (Cacioppo and Berntson, 1994; Lewicki et al., 1998; Harrison McKnight and Chervany, 2001; Chang and Fang, 2013).

Drawing upon their proposed affectual and emotional differentiation between trust and distrust, Lewicki et al. (1998) developed a 2x2 framework with trust on one axis and distrust on the other. This framework spans from "low trust/distrust" to "high trust/distrust" and provides an explanatory approach for the simultaneous and seemingly contradictory coexistence of trust and distrust. Figure 2 shows an adapted version of this two-dimensional framework by Lewicki et al. (1998) alongside a one-dimensional conceptualization of trust adapted from Castelfranchi and Falcone (2010).

In cases of low trust and low distrust (i.e., quadrant III in Figure 2), judgments about the trustworthiness or untrustworthiness of the trustee are still being formed (Lewicki et al., 1998). The trustor thus lacks a basis for either trust or distrust (Lewicki et al., 1998), and only over time do judgments develop. A practical and simplified example of such an interaction within the realm of AI might be a person encountering a chatbot for the first time. This person has no prior experience with the capabilities of large language models and, thus, no foundation to trust or distrust the chatbot.

Situations characterized by high trust and low distrust (i.e., quadrant IV) stem from predominantly positive experiences with the trustee. Contradictory evidence that could inform untrustworthiness is often disregarded or considered unimportant (Lewicki et al., 1998). Such a case could include, for our example, that the person frequently observed the chatbot's high capabilities in generating poetry. While trust is warranted (Jacovi et al., 2021) and calibrated (Lee and See, 2004) for these tasks, the individual might over-rely (Parasuraman and Riley, 1997) on the chatbot for other tasks, where distrusting and not relying on the chatbot would be more appropriate (e.g., providing accurate scientific literature).

With low trust and high distrust (i.e., quadrant II), negative experiences with the trustee predominate, reinforcing distrust. The trustor invests substantial resources in monitoring (Lewicki et al., 1998). Following our example, the individual could be disappointed by a chatbot's inability to provide accurate scientific literature. They may actively avoid using the chatbot or monitor it more closely, double-checking its responses. This could lead to warranted distrust (Jacovi et al., 2021) calibrated with the AI's untrustworthiness (Lee and See, 2004; Jacovi et al., 2021) for the given task, but potentially causing disuse (Lee and See, 2004) and under-reliance (Parasuraman and Riley, 1997) when trusting and relying on the chatbot would be appropriate.

Finally, in situations of high trust and high distrust (i.e., quadrant I), the experience with the trustee is balanced, having both perceived trustworthy and untrustworthy behavior. The trustor effectively interacts with the trustee in certain (trusted) tasks but not in other (distrusted) tasks (Lewicki et al., 1998). Returning to our example, the person has evidence to trust the chatbot for tasks aligned with its trustworthiness (i.e., capability to generate poems) and evidence to distrust the chatbot for tasks where distrust matches its untrustworthiness (i.e., incapability to provide accurate scientific literature). The person utilizes the chatbot to write poems but always double-checks its scientific references, showing both calibrated trust and distrust. This last case seems to be the most preferable, where both trust and distrust are *warranted* and *calibrated* with the AI's trustworthiness or untrustworthiness.

Conceptualizing and reasoning about trust and distrust in such a way allows for addressing the two key motivations of XAI as outlined by Jacovi et al. (2021): to increase trust in trustworthy AI and distrust in untrustworthy AI. Figure 2 also highlights the confines of a one-dimensional trust conception: if one overall trust score is formed, it is not possible to determine whether "low trust" arises from actual distrust or from a lack of knowledge regarding the trustworthiness or untrustworthiness of the AI (Victor, Cornelis, De Cock and Da Silva, 2009). Therefore, a two-dimensional conceptualization, with separate scores for trust and distrust, provides additional information and insights. For instance, individuals could be categorized on their respective levels of trust and distrust to identify

different user groups. Some users might display high trust and low distrust, others low trust and high distrust, and yet others might exhibit low levels of both trust and distrust, indicating a lack of information. This categorization would enable practitioners and researchers to identify distinct user needs and provide them with an informed decision basis to either increase the trustworthiness (and hence increasing trust) or decrease the untrustworthiness (and hence decreasing distrust) of their AI systems, aligning more closely with the extended goals of XAI as envisioned by Jacovi et al. (2021). Moreover, different factors distinctly contribute to the increase and decrease of trust, as opposed to those factors affecting distrust (Lewicki et al., 1998). This has been empirically demonstrated in other areas of human-computer interaction, where varying website characteristics distinctly contributed to trust and distrust (Seckler, Heinz, Forde, Tuch and Opwis, 2015). Similarly, in the context of AI and XAI, different trustworthiness cues may enhance trust (e.g., post-hoc explanations), while other cues could mitigate distrust (e.g., certification labels), and we encourage future research to investigate these potential factors.

However, only with appropriate questionnaires that measure both trust and distrust can such a two-dimensional consideration be done justice. Existing questionnaires like the TAI and TPA have limitations; the former does not account for distrust, and the latter holds room for improvement to accurately measure both dimensions, as indicated by our results. Good questionnaire development is rooted in a thorough understanding of the constructs being measured, usually grounded in theory and empirical research (Aeschbach, Perrig, Weder, Opwis and Brühlmann, 2021). In light of these requirements, comparatively little effort seems to have been made to understand and measure distrust in AI (Ueno et al., 2022; Scharowski and Perrig, 2023). Instead of adopting trust questionnaires from other research areas, we encourage the human-AI and XAI community to consider developing their own trust questionnaires, which take into account the unique nature of human-AI interaction. This involves generating items for the AI context that capture both trust *and* distrust. Not only would such a two-dimensional conceptualization provide the added value outlined above, but also contribute to a more comprehensive and holistic understanding of trust *and* distrust.

7. Limitations and Future Work

First, the present work utilized crowd-sourcing for participant recruitment. While crowd-sourced data have been shown to be at least as reliable as other, more traditional ways of recruitment, such as student sampling (Buhrmester, Kwang and Gosling, 2011; Douglas et al., 2023), future work should examine how the two trust scales perform across varying populations.

Second, ratings were collected in an online experiment with a scenario-based approach where participants observed AI interactions. While this is a common approach (e.g., Holthausen et al., 2020; Schaefer, 2016) that had the advantage of reaching the necessary number of participants for a high-powered validation study, future work should investigate alternative approaches, using other forms of interaction with AI.

Third, the present findings are limited to the context of automated vehicles and chatbots. While these are arguably timely and crucial application areas of AI and our findings are largely consistent with prior work in automation (Spain et al., 2008) and preliminary findings for the AI domain (Perrig et al., 2023b), future work should consider additional AI contexts, such as medical diagnosis or content recommendations.

Finally, a general limitation of statistical factor analysis is that the item wording, particularly the simultaneous use of positively and negatively formulated items, potentially influences participant responses (Perrig, von Felten, Honda, Opwis and Brühlmann, 2023a; Sauro and Lewis, 2011). Negatively formulated items can lead participants to intentionally or unintentionally ignore or misunderstand these items. The resulting response patterns may load on two distinct factors in a factor analysis due to methodological issues related to the item wording (Lewis and Sauro, 2017). Such methodological issues could be an alternative explanation for the revealed two-factor structure, and we recognize these challenges. Distorted factor structures have been shown for scales of usability (Lewis and Sauro, 2017; Lewis, Utesch and Maher, 2013) and website aesthetics (Perrig et al., 2023a), where an argument was made to not distinguish factors based on item wording because it lacked theoretical ground. In the case of trust, however, we pointed out that a distinction between trust and distrust is theoretically justified and has merits that go beyond positive or negative item formulation (Peters and Visser, 2023; Scharowski and Perrig, 2023). Ultimately, the underlying structure of psychological constructs, such as trust, is not rooted in statistical but rather in theoretical considerations (Fried, 2020). We want to emphasize that the psychometric validation of the TPA and TAI, along with our recommendations for using these two scales, remain robust despite this limitation. While our work thus contributes to more reliable and valid tools for measuring trust, it should not be taken as the final verdict in the discourse regarding the dimensionality of trust and distrust. Future research could explore the external validity of our results by examining how varying levels

of trust and distrust differently affect behavioral measures such as reliance. For instance, researchers could investigate whether high levels of distrust are more predictive of reliance than low levels of trust. Longitudinal studies could also provide deeper insights into how trust and distrust evolve over time and influence behavior in real-world settings. Such findings could further emphasize a distinction between these two measures.

8. Conclusion

Trust is a central and frequently measured construct in studying human-AI interactions. However, no validated trust questionnaire explicitly designed for the context of AI exists to date, with researchers relying on scales developed for other research areas, such as automation or human-human interaction. Motivated by the need for validated and standardized questionnaires, the present work reported on the first comprehensive validation of two trust scales in the context of AI, the popular TPA (Jian et al., 2000) and the recently published TAI (Hoffman et al., 2023). In a 2x2 online study design, using two conditions (trustworthy vs. untrustworthy) and two areas of applications (AV vs. chatbot), 2970 complete responses to the two scales and related measures were collected from 1485 participants.

While results from the psychometric evaluation supported both scales' psychometric quality regarding reliability, convergent, divergent, and criterion validity, findings were less favorable concerning the TPA's construct validity. Consequently, we investigated ways to improve the TPA, namely item removal and an alternative two-factor model, which enhanced the scale's psychometric quality. From our findings, we derived recommendations for researchers and practitioners who want to use the TPA and TAI in the context of AI. Results emphasized that while the TAI only measures trust, the TPA can measure two constructs: trust and distrust. Based on these findings, we highlighted the practical and theoretical implications of accounting for both trust and distrust, underscoring the added value of this distinction beyond a theoretical discussion to actual measurement practice. However, the TPA and TAI are not optimized for measuring both trust and distrust in the AI context, at least in their current versions. We therefore encourage future work on the TPA or the development of a scale explicitly designed for the context of AI, which measures both constructs. Such a distinction could contribute to a deeper and more nuanced understanding of trust *and* distrust in the human-AI interaction in a world where AI increasingly has the potential for both benefits and harm.

9. Funding and Data Availability

This research was financed entirely by our research group; we received no additional funding. The pre-registration (<https://doi.org/10.17605/OSF.IO/3EU4V>) and supplementary materials (<https://doi.org/10.17605/OSF.IO/7CDNE>) for this study are available on OSF.

10. Funding and Declaration of Conflicting Interests

This work is financed entirely by the corresponding author's research group, as we received no additional funding. The authors have no commercial or financial relationships to declare that could be construed as a potential conflict of interest.

11. Declarations of Interest

None.

CRediT authorship contribution statement

Nicolas Scharowski: Conceptualization, Methodology, Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration. **Sebastian A. C. Perrig:** Conceptualization, Methodology, Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration. **Lena Fanya Aeschbach:** Writing - Review & Editing. **Nick von Felten:** Writing - Review & Editing, Validation. **Prof. Dr. Klaus Opwis:** Writing - Review & Editing, Funding acquisition. **Prof. Dr. Philipp Wintersberger:** Conceptualization of this study, Resources, Writing - Review & Editing. **Dr. Florian Brühlmann:** Conceptualization of this study, Writing - Review & Editing, Supervision.

References

- Aeschbach, L.F., Perrig, S.A.C., Weder, L., Opwis, K., Brühlmann, F., 2021. Transparency in measurement reporting: A systematic literature review of CHI PLAY. *Proc. ACM Hum.-Comput. Interact.* 5. doi:10.1145/3474660.
- Andras, P., Esterle, L., Guckert, M., Han, T.A., Lewis, P.R., Milanovic, K., Payne, T., Perret, C., Pitt, J., Powers, S.T., Urquhart, N., Wells, S., 2018. Trusting intelligent machines: Deepening trust within socio-technical systems. *IEEE Technology and Society Magazine* 37, 76–83. doi:10.1109/MTS.2018.2876107.
- Benk, M., Tolmeijer, S., von Wangenheim, F., Ferrario, A., 2022. The value of measuring trust in AI-a socio-technical system perspective. *arXiv Preprint arXiv:2204.13480* Retrieved from <https://arxiv.org/abs/2204.13480>.
- Berg, J., Dickhaut, J., McCabe, K., 1995. Trust, reciprocity, and social history. *Games and economic behavior* 10, 122–142. doi:10.1006/game.1995.1027.
- Binns, R., van Kleek, M., Veale, M., Lyngs, U., Zhao, J., Shadbolt, N., 2018. 'it's reducing a human being to a percentage', in: Mandryk, R., Hancock, M., Perry, M., Cox, A. (Eds.), *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, ACM, New York, NY. pp. 1–14. doi:10.1145/3173574.3173951.
- Brühlmann, F., Petralito, S., Aeschbach, L., Opwis, K., 2020. The quality of data collected online: An investigation of careless responding in a crowdsourced sample. *Methods in Psychology* 2, 100022. doi:10.1016/j.metip.2020.100022.
- Buçinca, Z., Lin, P., Gajos, K.Z., Glassman, E.L., 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems, in: *Proceedings of the 25th International Conference on Intelligent User Interfaces*, ACM, New York, NY. p. 454–464. doi:10.1145/3377325.3377498.
- Buhrmester, M., Kwang, T., Gosling, S.D., 2011. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science* 6, 3–5. doi:10.1177/1745691610393980.
- Cacioppo, J.T., Berntson, G.G., 1994. Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates. *Psychological Bulletin* 115, 401–423. doi:10.1037/0033-2909.115.3.401.
- Castelfranchi, C., Falcone, R., 2010. *Trust theory: A socio-cognitive and computational model*. John Wiley & Sons, Hoboken, NJ. doi:10.1002/9780470519851.
- Chang, Y.S., Fang, S.R., 2013. Antecedents and distinctions between online trust and distrust: Predicting high-and low-risk internet behaviors. *Journal of Electronic Commerce Research* 14, 149.
- Cronbach, L.J., 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 297–334. doi:10.1007/BF02310555.
- Curran, P.G., 2016. Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology* 66, 4–19. doi:10.1016/j.jesp.2015.07.006.
- Davis, F.D., 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly* 13, 319–340. doi:10.2307/249008.
- DeVellis, R.F., 2017. *Scale Development: Theory and Applications*. 4th ed., SAGE publications, Inc., Thousand Oaks, CA.
- Dolinek, L., Wintersberger, P., 2022. Towards a generalized scale to measure situational trust in AI systems, in: *CHI 2022 TRAIT Workshop on Trust and Reliance in AI-Human Teams*, Association for Computing Machinery, New York, NY.
- Douglas, B.D., Ewell, P.J., Brauer, M., 2023. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLoS ONE* 18, e0279720. doi:10.1371/journal.pone.0279720.
- Dunn, T.J., Baguley, T., Brunson, V., 2014. From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology* 105, 399–412. doi:10.1111/bjop.12046.
- Esterwood, C., Robert Jr, L.P., 2023. Three strikes and you are out!: The impacts of multiple human–robot trust violations and repairs on robot trustworthiness. *Computers in Human Behavior* 142, 107658. doi:10.1016/j.chb.2023.107658.
- Flake, J.K., Fried, E.I., 2020. Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science* 3, 456–465. doi:10.1177/2515245920952393.
- Fried, E.I., 2020. Theories and models: What they are, what they are for, and what they are about. *Psychological Inquiry* 31, 336–344. doi:10.1080/1047840X.2020.1854011.
- Fukuyama, F., 1996. *Trust: The social virtues and the creation of prosperity*. Free Press, New York, NY.
- Furr, M., 2011. *Scale Construction and Psychometrics for Social and Personality Psychology*. SAGE publications, Ltd., London.
- Gambetta, D., 2000. Can we trust trust?, in: Gambetta, D. (Ed.), *Trust: Making and Breaking Cooperative Relations*, electronic edition. Department of Sociology, University of Oxford, Oxford. chapter 13, pp. 213–237. URL: <http://www.sociology.ox.ac.uk/papers/gambetta213-237.pdf>.
- George, D., Mallery, P., 2019. *IBM SPSS statistics 26 step by step: A simple guide and reference*. 16 ed., Routledge, New York, NY. doi:10.4324/9780429056765.
- Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., 2010. *Multivariate Data Analysis*. 7th ed., Prentice Hall, Hoboken, NJ.
- Harrison McKnight, D., Chervany, N., 2001. While trust is cool and collected, distrust is fiery and frenzied: A model of distrust concepts, in: *7th Americas Conference on Information Systems, AMCIS 2001 Proceedings*, Boston, MA. pp. 883–888.
- Henze, N., Zirkler, B., 1990. A class of invariant consistent tests for multivariate normality. *Communications in Statistics-Theory and Methods* 19, 3595–3617. doi:10.1080/03610929008830400.
- Hoff, K.A., Bashir, M., 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors* 57, 407–434. doi:10.1177/0018720814547570.
- Hoffman, R.R., Lee, J.D., Woods, D.D., Shadbolt, N., Miller, J., Bradshaw, J.M., 2009. The dynamics of trust in cyberdomains. *IEEE Intelligent Systems* 24, 5–11. doi:10.1109/MIS.2009.124.
- Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J., 2023. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science* 5. doi:10.3389/fcomp.2023.1096257.

- Holthausen, B.E., Wintersberger, P., Walker, B.N., Riener, A., 2020. Situational trust scale for automated driving (sts-ad): Development and initial validation, in: 12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Association for Computing Machinery, New York, NY. p. 40–47. doi:10.1145/3409120.3410637.
- Hopkins, K.D., 1998. Educational and Psychological Measurement and Evaluation. Pearson, London.
- Howard, M.C., 2016. A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human–Computer Interaction* 32, 51–62. doi:10.1080/10447318.2015.1087664.
- Hu, L., Bentler, P.M., 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal* 6, 1–55. doi:10.1080/10705519909540118.
- Jacovi, A., Marasović, A., Miller, T., Goldberg, Y., 2021. 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, ACM, New York, NY. p. 624–635. doi:10.1145/3442188.3445923.
- Jakesch, M., Bućinca, Z., Amershi, S., Olteanu, A., 2022. How different groups prioritize ethical values for responsible AI, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY. p. 310–323. doi:10.1145/3531146.3533097.
- Jian, J.Y., Bisantz, A.M., Drury, C.G., 2000. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics* 4, 53–71. doi:10.1207/S15327566IJCE0401_04.
- Juniper, E.F., 2009. Validated questionnaires should not be modified. *European Respiratory Journal* 34, 1015–1017. doi:10.1183/09031936.00110209.
- Kapania, S., Siy, O., Clapper, G., SP, A.M., Sambasivan, N., 2022. “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, Association for Computing Machinery, New York, NY. doi:10.1145/3491102.3517533.
- Kaplan, A.D., Kessler, T.T., Brill, J.C., Hancock, P., 2023. Trust in artificial intelligence: Meta-analytic findings. *Human factors* 65, 337–359. doi:10.1177/00187208211013988.
- Kline, R.B., 2016. Principles and practice of structural equation modeling. 4 ed., The Guilford Press, New York, NY.
- Kohn, S.C., de Visser, E.J., Wiese, E., Lee, Y.C., Shaw, T.H., 2021. Measurement of trust in automation: A narrative review and reference guide. *Frontiers in Psychology* 12, 604977. doi:10.3389/fpsyg.2021.604977.
- Kroeger, F., 2019. Unlocking the treasure trove: How can luhmann’s theory of trust enrich trust research? *Journal of Trust Research* 9, 110–124. doi:10.1080/21515581.2018.1552592.
- Lai, V., Chen, C., Smith-Renner, A., Liao, Q.V., Tan, C., 2023. Towards a science of human-AI decision making: An overview of design space in empirical human-subject studies, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY. p. 1369–1385. doi:10.1145/3593013.3594087.
- Langer, M., Hunsicker, T., Feldkamp, T., König, C.J., Grgić-Hlača, N., 2022. “look! it’s a computer program! it’s an algorithm! it’s AI!”: Does terminology affect human perceptions and evaluations of algorithmic decision-making systems?, in: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY. doi:10.1145/3491102.3517527.
- Langer, M., Oster, D., Speith, T., Kästner, L., Hermanns, H., Schmidt, E., Sesing, A., Baum, K., 2021. What do we want from explainable artificial intelligence (XAI)? a stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* 296, 103473. doi:10.1016/j.artint.2021.103473.
- Lee, J.D., See, K.A., 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 50–80. doi:10.1518/hfes.46.1.50_30392.
- Lewicki, R.J., McAllister, D.J., Bies, R.J., 1998. Trust and distrust: New relationships and realities. *The Academy of Management Review* 23, 438–458. doi:10.2307/259288.
- Lewicki, R.J., Tomlinson, E.C., Gillespie, N., 2006. Models of interpersonal trust development: Theoretical approaches, empirical evidence, and future directions. *Journal of management* 32, 991–1022. doi:10.1177/0149206306294405.
- Lewis, J.R., Sauro, J., 2017. Revisiting the factor structure of the system usability scale. *Journal of Usability Studies* 12, 183–192.
- Lewis, J.R., Utesch, B.S., Maher, D.E., 2013. UMUX-LITE: When there’s no time for the SUS, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA. p. 2099–2102. doi:10.1145/2470654.2481287.
- Liao, Q., Sundar, S.S., 2022. Designing for responsible trust in AI systems: A communication perspective, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY. p. 1257–1268. doi:10.1145/3531146.3533182.
- Lipton, Z.C., 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 31–57. doi:10.1145/3236386.3241340.
- Loo, R., 2002. A caveat on using single-item versus multiple-item scales. *Journal of Managerial Psychology* 17, 68–75. doi:10.1108/02683940210415933.
- Lu, Z., Yin, M., 2021. Human reliance on machine learning models when performance feedback is limited: Heuristics and risks, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY. doi:10.1145/3411764.3445562.
- Luhmann, N., 1979. Trust and Power. Wiley, Hoboken, NJ.
- Lyons, J.B., Stokes, C.K., Eschleman, K.J., Alarcon, G.M., Bareika, A.J., 2011. Trustworthiness and its suspicion: An evaluation of the nomological network. *Human Factors* 53, 219–229. doi:10.1177/0018720811406726.
- MacCallum, R.C., Browne, M.W., Sugawara, H.M., 1996. Power analysis and determination of sample size for covariance structure modeling. *Psychological methods* 1, 130–149. doi:10.1037/1082-989X.1.2.130.
- Mardia, K.V., 1970. Measures of multivariate skewness and kurtosis with applications. *Biometrika* 57, 519–530. doi:10.1093/biomet/57.3.519.

- Mayer, R.C., Davis, J.H., Schoorman, F.D., 1995. An integrative model of organizational trust. *Academy of Management Review* 20, 709–734. doi:10.2307/258792.
- McDonald, R.P., 1999. *Test theory: A unified treatment*. 1 ed., Psychology Press, New York, NY. doi:10.4324/9781410601087.
- McKnight, D.H., Chervany, N.L., 2001. What trust means in e-commerce customer relationships: An interdisciplinary conceptual typology. *International journal of electronic commerce* 6, 35–59. doi:10.1080/10864415.2001.11044235.
- Meade, A.W., Craig, S.B., 2012. Identifying careless responses in survey data. *Psychological methods* 17, 437–455. doi:10.1037/a0028085.
- Merritt, S.M., 2011. Affective processes in human–automation interactions. *Human Factors* 53, 356–370. doi:10.1177/0018720811411912.
- Mohseni, S., Zarei, N., Ragan, E.D., 2020. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *arXiv: Human-Computer Interaction arXiv:1811.11839*. retrieved from <https://arxiv.org/abs/1811.11839>.
- Moosbrugger, H., Kelava, A., 2020. *Testtheorie und Fragebogenkonstruktion [Test Theory and Questionnaire Construction]*. 3rd ed., Springer Berlin, Heidelberg, Berlin, Germany. doi:10.1007/978-3-662-61532-4.
- Muir, B.M., 1994. Trust in automation: Part i. theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics* 37, 1905–1922. doi:10.1080/00140139408964957.
- Ou, C.X., Sia, C.L., 2009. To trust or to distrust, that is the question: Investigating the trust-distrust paradox. *Commun. ACM* 52, 135–139. URL: <https://doi.org/10.1145/1506409.1506442>, doi:10.1145/1506409.1506442.
- Parasuraman, R., Riley, V., 1997. Humans and automation: Use, misuse, disuse, abuse. *Human Factors* 39, 230–253. doi:10.1518/001872097778543886.
- Peer, E., Rothschild, D., Gordon, A., Evernden, Z., Damer, E., 2022. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods* 54, 1643–1662. doi:10.3758/s13428-021-01694-3.
- Perrig, S.A.C., von Felten, N., Honda, M., Opwis, K., Brühlmann, F., 2023a. Development and validation of a positive-item version of the visual aesthetics of websites inventory: The VisAWI-Pos. *International Journal of Human-Computer Interaction* 0, 1–25. doi:10.1080/10447318.2023.2258634.
- Perrig, S.A.C., Scharowski, N., Brühlmann, F., 2023b. Trust issues with trust scales: Examining the psychometric quality of trust measures in the context of AI, in: *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, New York, NY. doi:10.1145/3544549.3585808.
- Peters, T.M., Visser, R.W., 2023. The importance of distrust in AI, in: Longo, L. (Ed.), *Explainable Artificial Intelligence*, Springer Nature Switzerland, Cham. pp. 301–317. doi:10.1007/978-3-031-44070-0_15.
- Poursabzi-Sangdeh, F., Goldstein, D.G., Hofman, J.M., Wortman Vaughan, J.W., Wallach, H., 2021. Manipulating and measuring model interpretability, in: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ACM, New York, NY. doi:10.1145/3411764.3445315.
- R Core Team, 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: <https://www.R-project.org/>.
- Rousseau, D.M., Sitkin, S.B., Burt, R.S., Camerer, C., 1998. Not so different after all: A cross-discipline view of trust. *Academy of Management Review* 23, 393–404. doi:10.5465/AMR.1998.926617.
- Saunders, M.N., Dietz, G., Thornhill, A., 2014. Trust and distrust: Polar opposites, or independent but co-existing? *Human Relations* 67, 639–665. URL: <https://doi.org/10.1177/0018726713500831>.
- Sauro, J., Lewis, J.R., 2011. When designing usability questionnaires, does it hurt to be positive?, in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, New York, NY. pp. 2215–2224. doi:10.1145/1978942.1979266.
- Schaefer, K.E., 2016. Measuring Trust in Human Robot Interactions: Development of the "Trust Perception Scale-HRI", in: Mittu, R., Sofge, D., Wagner, A., Lawless, W. (Eds.), *Robust Intelligence and Trust in Autonomous Systems*. Springer US, Boston, MA, pp. 191–218. doi:10.1007/978-1-4899-7668-0_10.
- Scharowski, N., Benk, M., Kühne, S.J., Wettstein, L., Brühlmann, F., 2023. Certification labels for trustworthy AI: Insights from an empirical mixed-method study, in: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, New York, NY. p. 248–260. doi:10.1145/3593013.3593994.
- Scharowski, N., Perrig, S.A., von Felten, N., Brühlmann, F., 2022. Trust and reliance in XAI—distinguishing between attitudinal and behavioral measures. *CHI TRAIT Workshop* doi:10.48550/arXiv.2203.12318.
- Scharowski, N., Perrig, S.A.C., 2023. Distrust in (X)AI – measurement artifact or distinct construct? *CHI 2023 TRAIT Workshop on Trust and Reliance in AI-Human Teams* doi:10.48550/arXiv.2303.16495.
- Schlicker, N., Uhde, A., Baum, K., Hirsch, M.C., Langer, M., 2022. A micro and macro perspective on trustworthiness: Theoretical underpinnings of the trustworthiness assessment model (tram) doi:10.31234/osf.io/qhwvx.
- Seckler, M., Heinz, S., Forde, S., Tuch, A.N., Opwis, K., 2015. Trust and distrust on the web: User experiences and website characteristics. *Computers in human behavior* 45, 39–50. doi:doi.org/10.1016/j.chb.2014.11.064.
- Sitkin, S.B., Roth, N.L., 1993. Explaining the limited effectiveness of legalistic “remedies” for trust/distrust. *Organization science* 4, 367–392. URL: <https://doi.org/10.1287/orsc.4.3.367>.
- Spain, R.D., Bustamante, E.A., Bliss, J.P., 2008. Towards an empirically developed scale for system trust: Take two. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 52, 1335–1339. doi:10.1177/154193120805201907.
- Thornton, L., Knowles, B., Blair, G., 2021. Fifty shades of grey: In praise of a nuanced approach towards trustworthy design, in: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, New York, NY. p. 64–76. doi:10.1145/3442188.3445871.
- Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C.G., van Moorsel, A., 2020. The relationship between trust in AI and trustworthy machine learning technologies, in: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ACM, New York, NY. pp. 272–283. doi:10.1145/3351095.3372834.

To Trust or Distrust Trust Measures

- Ueno, T., Sawa, Y., Kim, Y., Urakami, J., Oura, H., Seaborn, K., 2022. Trust in human-AI interaction: Scoping out models, measures, and methods, in: Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY. doi:10.1145/3491101.3519772.
- Vereschak, O., Bailly, G., Caramiaux, B., 2021. How to evaluate trust in AI-assisted decision making? a survey of empirical methodologies. *Proc. ACM Hum.-Comput. Interact.* 5. doi:10.1145/3476068.
- Victor, P., Cornelis, C., De Cock, M., Da Silva, P.P., 2009. Gradual trust and distrust in recommender systems. *Fuzzy Sets and Systems* 160, 1367–1382. doi:10.1016/j.fss.2008.11.014.
- Watson, D., Clark, L.A., Tellegen, A., 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology* 54, 1063–1070. doi:10.1037/0022-3514.54.6.1063.
- Weitz, K., Schiller, D., Schlagowski, R., Huber, T., André, E., 2021. “let me explain!”: exploring the potential of virtual agents in explainable AI interaction design. *Journal on Multimodal User Interfaces* 15, 87–98. doi:10.1007/s12193-020-00332-0.
- Whittaker, T.A., Schumacker, R.E., 2022. A beginner’s guide to structural equation modeling. 5 ed., Routledge, New York, NY. doi:10.4324/9781003044017.
- Wischniewski, M., Krämer, N., Müller, E., 2023. Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY. doi:10.1145/3544548.3581197.
- Yin, M., Wortman Vaughan, J., Wallach, H., 2019. Understanding the effect of accuracy on trust in machine learning models, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY. p. 1–12. doi:10.1145/3290605.3300509.
- Yu, K., Berkovsky, S., Taib, R., Conway, D., Zhou, J., Chen, F., 2017. User trust dynamics: An investigation driven by differences in system performance, in: Proceedings of the 22nd international conference on intelligent user interfaces, Association for Computing Machinery, New York, NY. pp. 307–317. doi:10.1145/3025171.3025219.

Certification Labels for Trustworthy AI: Insights From an Empirical Mixed-Method Study

Nicolas Scharowski
nicolas.scharowski@unibas.ch
University of Basel

Michaela Benk
mbenk@ethz.ch
Mobiliar Lab for Analytics
ETH Zürich

Swen J. Kühne
swen.kuehne@zhaw.ch
School of Applied Psychology
Zürich University of Applied Sciences

Léane Wettstein
leane.wettstein@unibas.ch
University of Basel

Florian Brühlmann
florian.bruehlmann@unibas.ch
University of Basel

ABSTRACT

Auditing plays a pivotal role in the development of trustworthy AI. However, current research primarily focuses on creating auditable AI documentation, which is intended for regulators and experts rather than end-users affected by AI decisions. How to communicate to members of the public that an AI has been audited and considered trustworthy remains an open challenge. This study empirically investigated *certification labels* as a promising solution. Through interviews ($N = 12$) and a census-representative survey ($N = 302$), we investigated end-users' attitudes toward certification labels and their effectiveness in communicating trustworthiness in low- and high-stakes AI scenarios. Based on the survey results, we demonstrate that labels can significantly increase end-users' trust and willingness to use AI in both low- and high-stakes scenarios. However, end-users' preferences for certification labels and their effect on trust and willingness to use AI were more pronounced in high-stake scenarios. Qualitative content analysis of the interviews revealed opportunities and limitations of certification labels, as well as facilitators and inhibitors for the effective use of labels in the context of AI. For example, while certification labels can mitigate data-related concerns expressed by end-users (e.g., privacy and data protection), other concerns (e.g., model performance) are more challenging to address. Our study provides valuable insights and recommendations for designing and implementing certification labels as a promising constituent within the trustworthy AI ecosystem.

CCS CONCEPTS

• Human-centered computing → Empirical studies in HCI.

KEYWORDS

AI, Audit, Documentation, Label, Seal, Certification, Trust, Trustworthy, User study

ACM Reference Format:

Nicolas Scharowski, Michaela Benk, Swen J. Kühne, Léane Wettstein, and Florian Brühlmann. 2023. Certification Labels for Trustworthy AI: Insights From an Empirical Mixed-Method Study. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*, June 12–15, 2023, Chicago, IL, USA. ACM, New York, NY, USA, 13 pages. <https://doi.org/10.1145/3593994>

1 INTRODUCTION

In recent years, the promise of artificial intelligence (AI) in transforming our lives has seen widespread advances in all sectors of society. AI is increasingly guiding our consumer choices [52], reshaping service by automatizing tasks [28], assisting managers in hiring decisions [42], or augmenting clinical decision-making [71]. In light of increasingly ubiquitous AI and its profound impact on human lives, various government institutions, scientific communities, and the general public are engaged in a widespread discourse on how to ensure trustworthy AI [31, 33, 36, 43] for both low-, and high-stake scenarios [11].

To this end, a large body of work has focused on identifying the principles that underlie trustworthy AI [36]. They include mitigating bias and unfairness in AI systems [41], explaining the reasoning of AI decisions [39], setting up mechanisms to hold AI accountable [36], and ensuring user privacy [60]. However, as trust is determined by people's perception [40, 43], efforts to design trustworthy AI are hampered by a lack of understanding of how to communicate trustworthiness to people, for instance, through documentation or other transparency affordances [43]. Particularly for end-users¹, trusting AI can be a challenge, as they lack the necessary expertise and knowledge to evaluate the various trustworthiness principles (e.g., robustness, privacy, fairness) [4, 37].

Motivated by these challenges, this work builds on research highlighting the pivotal role of *auditability* as an enabler of trust in AI [7, 65] and its crucial role in creating an "AI trustworthiness ecosystem" [2] by ensuring that the principles of trustworthy AI are met. Auditing refers to mechanisms that evaluate and ensure compliance with regulations and ethical standards [54]. Various methods have been proposed to increase AI systems' transparency and, thereby auditability, such as through the use of model documentation or information about datasets [14, 21]. While AI documentations are

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

FAccT '23, June 12–15, 2023, Chicago, IL, USA

© 2023 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0192-4/23/06.

<https://doi.org/10.1145/3593013.3593994>

¹In line with prior work [39, 58, 68], we define end-users in this paper as laypeople (i.e., non-experts in data science or machine learning) who may be affected directly or indirectly by the outcomes of AI systems.

valuable artifacts to inform audit decisions, they are tailored to regulators and experts and not intended to certify and communicate to end-users that an AI has met the auditing criteria.

For this reason, our work focuses on communicating the outcomes of auditing processes to end-users, a topic that has received little attention in previous work. Specifically, we investigate the use of *certification labels*, which are commonly used in other domains, such as food and energy [10, 16, 62]. Certification labels are relevant in the context of trustworthy AI for three reasons. First, through the use of simple language, icons, or color-coding, they are usually designed to be accessible to various stakeholder groups, including end-users with limited knowledge and time [24]. Second, if reflecting a genuine and credible auditing process, certification labels can communicate the criteria used in an audit, thereby serving as a "trustworthiness cue" for end-users [44, 57]. Third, labels have shown to promote trustworthiness of a product in other domains [64] facing similar challenges on how to certify that a product meets certain criteria, such as agricultural standards (e.g., organic foods [16]) or low ecological impact (e.g., sustainable hotels [10]). However, end-users' attitudes toward AI certification labels and their effectiveness in communicating the trustworthiness of AI remain to be explored.

We addressed this gap by conducting a mixed-method study with both interviews ($N = 12$) and a census-representative survey ($N = 302$) with end-users. Our results provide evidence that certification labels can effectively communicate AI trustworthiness. Qualitative findings revealed that end-users have positive attitudes toward AI certification labels and that labels can increase perceived transparency and fairness and are regarded as an opportunity to establish standards for AI systems. Particularly, data-related concerns expressed by end-users, such as privacy and data protection, can be mitigated through the use of certification labels. However, labels may not be able to address all raised concerns, such as model performance, suggesting that they should be considered one promising constituent among others for trustworthy AI. Furthermore, our results provide insights into facilitators and inhibitors for the effective design of certification labels in the context of AI. For example, end-users expressed strong preferences for independent audits and highlighted the challenge of communicating subjective criteria such as "fairness," whose meaning can be ambiguous.

Quantitative findings showed that a certification label significantly increases end-users' trust and willingness to use AI in both low- and high-stake AI scenarios. Nevertheless, end-users reported a higher preference for certification labels in high-stake scenarios (e.g., hiring procedure) than in low-stake scenarios (e.g., price comparison), and the positive effect of a label on trust and willingness to use AI was more pronounced in high-stake scenarios. This suggests that compliance with mandatory requirements for AI in high-stake scenarios could be effectively communicated to end-users through certification labels in addition to the proposed voluntary labeling for low-stake AI scenarios [11, 61].

To summarize, our study is the first to demonstrate the potential of certification labels as a promising approach for communicating to end-users that an audit has certified an AI to be trustworthy. We contribute to the trustworthy AI literature by highlighting opportunities and challenges for designing and effectively implementing certification labels.

2 AUDITING FOR TRUSTWORTHY AI

A growing body of work recognizes the critical role of algorithmic or AI auditing in enabling the trustworthiness of AI systems [2, 37, 65]. Prior work suggests that auditing improves fairness [69], accountability [13], and governance [17], among others. These elements are considered to contribute to trust in and acceptance of AI². Moreover, audits have the ability to expose problematic behavior, such as algorithmic discrimination, distortion, exploitation, and misjudgment [3]. In safety-critical industries such as aerospace, medicine, and finance, audits are a long-standing practice [13]. However, only recently have researchers recognized that these areas could inform AI auditing and acknowledged the importance of considering insights from the social sciences, where audits have emerged from efforts toward racial equity and social justice [66].

While the importance of AI auditing has been identified, the development of common audit practices, standards, or regulatory guidance is ongoing [3, 13] and efforts to create auditing frameworks throughout the AI development life-cycle are still in their early stages [54]. Auditing can be defined as "an independent evaluation of conformance of software products and processes to applicable regulations, standards, guidelines, plans, specifications, and procedures." [29, p. 30]. At least three types of AI auditing can be distinguished, including first-party internal auditing, second-party audits conducted by contractors, and independent third-party audits [13]. However, whether auditing should be conducted by independent third-parties or internally within organizations is a topic of ongoing academic discussion [17, 38, 54], with both approaches having their advantages and drawbacks. Raji et al. argue that external auditing may be constrained by a lack of access to organizations' internal processes and information that are often subject to trade secrets. In contrast, Falco et al. point out that the outcomes of internal audits are typically not publicly disclosed and that it often remains unclear whether the auditor's recommendations are effectively implemented or not. The question of whether end-users prefer internal or external audits remains to be investigated.

In addition to defining standards and best practices for AI auditing, it is crucial to consider how the outcomes of audits can be communicated to different stakeholders with varying knowledge and needs [72]. Current research has mainly focused on approaches for documenting machine learning (ML) models and training datasets. These artifacts play an important role in the AI trustworthiness ecosystem by increasing transparency and allowing auditors and regulators to determine whether principles of trustworthy AI (e.g., fairness, robustness, privacy [36]) have been met [37]. For example, "model cards" [14, 49] disclose information about a model's purpose and design process, its underlying assumptions, and the model's performance characteristics. Similarly, Gebru et al. introduced "datasheets," which summarize the motivation, composition, collection process, and recommended uses for datasets, and Floridi et al. recommended the use of "summary datasheets" and "external scorecards." The former is aligned with the goals of "datasheets" and synthesizes key information about the AI, including its purpose, status, and contact information. The latter is conceptually closely

²The definition of trust in AI and its operationalization is an ongoing debate [31, 56, 65, 67]. As an extensive theoretical discussion is out of scope of this work, we focus on trustworthiness, a property of the trustee, rather than on trust as a process that can be affected by numerous contextual and personal factors [8, 9].

related to "model cards" and evaluates the AI system along several dimensions to form an overall risk score [18].

However, these documentations are tailored to AI practitioners, and regulators [37, 58, 72], rather than end-users affected by AI decisions. Often, end-users have neither the access nor the expertise to understand the technical information that AI documentation provides [1]. It is unlikely that end-users can effectively utilize ML model documentation or data documentation to make informed judgments about trusting or using AI [37]. For this reason, end-users depend on auditors and regulators who can use these artifacts to verify and ensure the trustworthiness of AI. Yet, it remains an open research question of how to effectively communicate to end-users that an audit has considered an AI trustworthy. End-users require accessible communication tailored to their specific values and concerns [72]. A potentially effective way to provide such information is through the use of *certification labels*, which we will introduce in the following.

3 CERTIFICATION LABELS FOR AUDITED AI

Labels are widely used for displaying specific product or service attributes to help consumers make more informed decisions. They are well-established in various fields, such as agriculture [23], food [34], energy [59], and e-commerce [63]. Different kinds of labels exist, and various classification systems have been proposed [30, 61, 62]. For example, in the food industry, "nutrition labels" provide consumers with simplified and easily understandable information to identify a product's nutritional content. While this information can also be found in detailed tables on the back of food packing, for many consumers, this information is too complex, revealing similar challenges end-users face with AI documentation. This is where labels can provide information in a clear and accessible manner, utilizing simple language, icons, and color coding, which makes labels accessible to individuals from different backgrounds [22, 24]. Prior work in consumer research has shown that labels can communicate the outcomes of audits and thereby enhance trust in a product [64].

In this study, we focus on *certification labels*, which certify that a product or service meets one or several criteria and are thus suitable for the case of audited AI. Certification labels are exclusively awarded to products that have undergone an auditing process, typically conducted by a third-party organization [62]. By communicating an institutional assurance of trustworthiness, third-party organizations can serve as "trust surrogates" for the consumer, shifting the trust relation from trust in the AI to trust in the institution that provides the certification [64]. In this case, a certification label serves as a trustworthiness cue [57] that signals compliance with governance structures. Our work thus closely aligns with the proposal by Liao and Sundar, highlighting that the trustworthiness of AI is not inherently given but must be communicated and perceived as such by the user, for instance, through transparency affordances. According to the authors, people then use heuristics (i.e., mental rules of thumb) to evaluate these affordance cues to form judgments about the trustworthiness of AI. The authors further suggest that certifications from regulatory bodies that have audited the AI could serve as trustworthiness cues, invoking these heuristics. Therefore, certification labels in the context of AI are a promising

approach to communicate that a regulatory body has audited an AI and considered it trustworthy.

There have been several initiatives at a national and international level to introduce AI labels in both industry (e.g., [20], [25], [19]) and government (e.g., [15], [46]). These initiatives vary in their intended scope but are mostly still in an early stage. Previous studies have also emphasized the potential of labels as a means of AI certification [27, 58, 61]. Holland et al. proposed the concept of a "Data Set Nutrition Label," which would summarize key aspects of a dataset (e.g., metadata and the data source) prior to the development of ML models. Seifert et al. further suggested labels for trained ML models that independent reviewers have evaluated based on properties such as accuracy, fairness, and transparency. A recent study by Stuurman and Lachaud commented on various labels to provide information to end-users affected by AI decisions. Drawing from the EU Act on AI [12], the study distinguished between low-stake and high-stake AI systems and proposed a voluntary labeling system for AI not considered high-stake. This distinction aligns with recommendations from the EU's "white paper on artificial intelligence," [11] which encourages organizations to use labels to demonstrate the trustworthiness of their AI-based products and services. A survey conducted with individuals and organizations directly or indirectly engaged in audits found that while respondents believed that AI audits should be mandatory, 53% supported mandating them only for high-stakes systems [13]. End-users' perceptions of certification labels in low and high-stakes AI scenarios have not yet been investigated.

Despite this extensive theoretical work on labels in the context of AI and their gradual adoption in industry and government, there is currently a lack of empirical research exploring end-users' attitudes toward AI certification labels and their effectiveness in communicating trustworthiness in low- and high-stake AI scenarios. This study aims to address this research gap and inform current industry and government initiatives.

4 RESEARCH QUESTIONS

Based on the aforementioned considerations, we investigated the following research questions:

RQ1: What are end-users' attitudes toward certification labels in the context of AI?

RQ2: How do certification labels affect end-users' trust and willingness to use AI in low- and high-stake scenarios?

5 METHODS

To answer these research questions, we used a mixed-method research approach consisting of semi-structured interviews and a subsequent survey to collect quantitative data as part of a within-subjects design study. For both the interviews and the survey, we used a scenario-based approach to investigate people's attitudes and the effects of a certification label, inspired by past research [5, 32, 35]. In the interviews, we asked participants about their attitudes toward AI and certification labels. As a follow-up within-subjects study, we implemented a survey to investigate the effect of a certification label quantitatively. The semi-structured interviews served as a basis for the survey and a means to enrich the quantitative results. The quantitative survey complemented the qualitative interviews

by extending our results to a larger census-representative sample. In the following, we will introduce the certification label used in our study before describing the procedures of each method in more detail.

5.1 The certification label

To investigate labels in the context of AI, we used a certification label that has already been developed for the broader context of digital trust. Using an existing label had the advantage that it had undergone an extensive design process and thus did not need to be created from scratch.

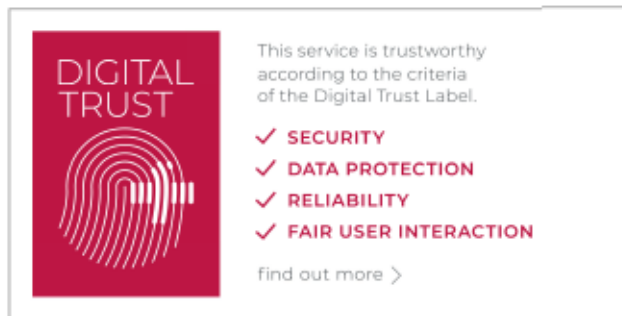


Figure 1: The "Digital Trust Label," which we adopted as a certification label for AI. ©2023 Swiss Digital Initiative

The non-profit foundation Swiss Digital Initiative laid the groundwork for developing this certification label. At the label's core lies a catalog of verifiable and auditable criteria, co-developed by an academic expert group based on a user study on digital trust. A panel of independent experts from academia, data and consumer protection, and digital ethics further developed the label catalog. Involving digital service providers and auditors in the designing process ensured that the criteria were auditable and verifiable. The catalog that forms the basis of the audit currently contains 35 criteria that are summarized into four categories:

- (1) Security (criteria 1 - 12): What is the security standard? The service provider shall, e.g., ensure that the data is encrypted as it transfers so that third-parties cannot access it.
- (2) Data protection (criteria 13 - 20): How is the data protected? The service provider shall, e.g., assume responsibility for the appropriate management of the data.
- (3) Reliability (criteria 21 - 29): How reliable is the service or product? The service provider shall, e.g., take all actions required to safeguard the continuity of the service.
- (4) Fair user interaction (criteria 30 - 35): Is automated decision-making involved? The service provider shall, e.g., ensure that all users receive equal treatment and that there is no data-based service or price discrimination.

If an organization would like its digital product or service (e.g., a chatbot) to receive the certification label, it can voluntarily request an audit and thus participate in the certification process. After a scoping call with third-party auditors, an audit is performed along the criteria catalog. The audit leads to an audit report detailing the performance per criterion, which is double-checked by an independent label certification committee composed of auditing experts. If

non-conformities are identified, the organization applying for the label must fix the identified issues, e.g., adjust its privacy policy. After a successful auditing report, the certification label is awarded for a period of three years with two audits during that period.

5.2 Scenario selection

Participants were presented with real-world examples of AI systems, adapted from Kapania et al., namely *medical diagnosis*, *loan approval*, *hiring procedure*, *music preference*, *route planning* and *price comparison* (see materials on OSF: <https://osf.io/gzpz5k/>). One advantage of using hypothetical scenarios instead of real consumer applications is that differences in participants' prior experience with the applications can be controlled for Kapania et al. and Woods et al. proposed that people's behavior in scenario-based experiments corresponds to their real-life behavior. To answer our second research question and following Kapania et al. we explored both low-stake scenarios (music preference, route planning, price comparison) and high-stake scenarios (medical diagnosis, hiring procedure, loan approval). This distinction was crucial since other researchers [18, 61] and the "EU AI Act" [12] have discussed the use of AI labels for "low-stake" and "high-stake" scenarios. This classification was based on the AI's respective impact on affected parties and the involvement of significant risks, in particular with respect to safety, consumer rights, and the use of personal data.

5.3 Interviews

5.3.1 Participants. Initially, we invited 16 participants to an interview on-site at the university. The recruitment was carried out through a university-internal database and an online marketplace where scientific studies can be advertised. To ensure that our sample consisted of end-users (i.e., laypeople who may be affected directly or indirectly by the outcomes of AI systems), we used screening questions following Kapania et al. and asked potential participants about their knowledge of AI and experience working with AI-based systems. We selected participants who indicated that they have heard about AI but did not work with it and provided a comprehensible description or adequate example of what AI is without overly restricting the valid responses (e.g., "robots" was valid while obvious nonsense answers such as "E.T. the alien" was deemed invalid). In addition, we asked participants to indicate their age, gender, profession, and English language proficiency so that we could design the interviews as balanced as possible and present materials in English. However, four interviews did not take place due to no-shows. We, therefore, conducted 12 interviews with end-users of different backgrounds, ages, and genders that lasted 60 - 90 minutes. The interviews were conducted in German and recorded through field notes and audio recordings. Each participant received compensation in the form of a gift card worth CHF 10.00 from a Swiss retail company. The final sample ($M_{age} = 35.42$, $SD_{age} = 12.50$, $Min_{age} = 23$, $Max_{age} = 66$) consisted of students (P2, P3, P4, P8, P11) enrolled in linguistics and literature (P2), fine arts (P3), and psychology (P4, P8, P11), as well as individuals who described their occupation as a bike messenger (P12), waitress (P1), dancer (P9), course manager (P7), management assistant (P6), intern (P10) and retired teacher (P5). The sample was predominantly female, with ten women and two men.

5.3.2 Procedure. Before the interviews, participants had to read and sign a declaration of consent. In the declaration, we informed participants of the purpose and rationale of the study, the researcher affiliations, the voluntary nature of study participation, and how their data will be analyzed and shared. All personally identifiable information was deleted to ensure privacy, and the anonymous data was stored without actual reference to the participants.

During the interviews, we asked attitudinal questions about AI, specifically where participants saw opportunities and challenges in using AI. We then presented the six scenarios to the participants without specifying the low- and high-stake categorization we had made in advance. Based on the respective headings of the scenarios (e.g., music preference), without further information, we asked participants to order the scenarios via drag and drop from "most impactful" (rank 1) to "least impactful" (rank 6). To ensure comparability, we defined "most impactful" for participants as "the scenario that would have the greatest impact on your personal life." This question aimed to validate our categorization in low- and high-stake scenarios. Next, we presented participants with one low-stake and one high-stake scenario and asked how they differed from one another. After this, participants were introduced to the certification label and asked how they perceived it, whether the label criteria were comprehensible or not, and where they saw opportunities and drawbacks of a certification label. The goal of the interviews was not only to gather qualitative data, but also to identify and determine which questions best suited the subsequent survey. We, therefore, made sure the questions were comprehensible and free of ambiguities. Any difficulties encountered during the interviews were discussed within the research team, and, if necessary, the respective questions were revised or removed. We refer to the digital repository for the complete interview manual.

5.4 Survey

5.4.1 Participants. To gain insights into how a general population perceives a label in the context of AI, we hired a market research agency (<https://www.bilendi.ch/>) to provide us with a Swiss census-representative sample regarding age and gender (quota sampling). We used the same screening questions as in the interviews and initially recruited 395 participants that received CHF 3.00 for taking part in the 15-minute online survey. Following a quality assessment using a self-reported single item as an indicator of careless responding [6, 48], 302 participants remained for data analysis. The sample is census-representative regarding age ($M_{age} = 43.88$, $SD_{age} = 16.08$, $Min_{age} = 18$, $Max_{age} = 79$) and the gender distribution (150 women, 151 men, one non-binary person).

5.4.2 Procedure and measures. The survey consisted of three parts. First, after providing informed consent and a brief introduction to the study, participants were free to select one scenario from the low-stake and one from the high-stake categorization. After making their choice, they received full descriptions of the two scenarios (see Appendix A) and were asked to rate their trust ("how much would you trust the AI in the scenario presented?") and willingness to use ("how much would you be willing to use the AI in the scenario presented?") on a scale from 0 (= not at all) to 100 (= absolutely). In addition, participants were asked in which scenario they would more readily accept the AI's decision/recommendation (i.e., "in

which of the two scenarios would you be more willing to accept the decision/recommendation made by AI?").

Participants were introduced to the certification label in the second part of the survey. They were asked for their impression and rated the importance of each criterion (i.e., "how important are the label criteria for you in the context of AI?") on a scale from 0 (= not at all) to 100 (= absolutely). Participants were also asked what effect the certification label had on their acceptance (i.e., "would you be more likely to accept an AI's decision/recommendation if it had received a label?") and preference (i.e., "in which one of the two scenarios would you prefer the use of a label?"). To understand end-users' preferences regarding external and internal auditing, we included an open-ended question (i.e., "who do you think should be responsible for awarding such a label?").

Finally, in the fourth part, we again let participants rate the AI in the same low- and high-stake scenario on trust and willingness to use, this time with the information that the AI had been awarded a certification label. This second assessment allowed us to examine the certification label's effect on trust and willingness to use ratings. Similarly to the first assessment, we asked participants to justify their ratings and why a label led to increased/decreased or unchanged ratings. At the end of the survey, we asked the participants for feedback and the question, "*in your honest opinion, should we use your data in our analyses in this study? Do not worry, this will not affect your payment. You will receive the compensation either way,*" as an additional quality check. The complete survey can be found on the digital repository.

5.5 Analysis and coding procedure

We used the qualitative interview data to answer RQ1 and the quantitative survey data to answer RQ2. The interview data was evaluated using qualitative content analysis [47], more specifically summarizing content analysis. We followed the procedure according to Mayring and Fenzl by determining the coding unit, paraphrasing, generalization to the level of abstraction, first reduction, and second reduction to form a cross-case category system. Coding was carried out by three researchers who independently went through four interviews each. To ensure consistency, one interview was evaluated by all researchers. Any ambiguities and discrepancies were resolved through open discussions, and the final cross-case category system was formed in a group session. The quantitative data analysis was carried out in R (version 4.2.2. [53]). We used the *ggstatsplot* package (version 0.9.1. [51]) to conduct statistical testing and report *t*-values, standard deviations, and the corresponding *p*-values. We set the level of statistical significance to $\alpha = .05$.

6 RESULTS

6.1 Attitudes toward certification labels

The content analysis of the interview data resulted in 127 case-specific categories, which were further consolidated across participants into 25 categories. These cross-categories were grouped into the following topics: "*AI-related concerns, risks, problems,*" "*AI-related opportunities, advantages,*" "*attitudes toward certification labels,*" and perceived "*differences between low- and high-stakes scenarios*". For the purpose of this study, we focus on the topic "*attitudes toward certification labels,*" as this was the most relevant

Table 1: End-users' attitudes toward certification labels

Category	Subcategory	Example quote
Opportunities for certification labels	Increasing trust	"Because if it is monitored and these various criteria have to be met in order to get the label, then I as a consumer can, of course, trust better and also know that there are perhaps controls and random checks, so I would definitely trust more." (P6)
	Increasing perceived transparency	"I think that if there is such an established label, it will certainly help to increase transparency." (P6)
	Increasing perceived fairness	"With the Fair User Interaction aspect, yes, probably so [fairness is increased]. ... if the AI is now checked for this, and it can be determined that it is not data-based, treated differently." (P12)
	Auditing of AI systems	"Because I'm not an expert in the field and the label ..., gives me proof ... that it's tested by experts." (P4)
	Establishing standards for AI systems	"So I could imagine that if it is a bit more standardized, so to speak, because you have to meet certain standards, that it could introduce a general level of fairness." (P3)
	Covering relevant concerns	"The concern [responsibility] was covered and then just the general concern with all just how our data is also used and hopefully not misused, or yes. That is also covered." (P10)
Facilitators for effective certification labels	Additional label information	"[I would like to] find out what this 'Fair User Interaction' means, what it refers to, how my data is protected ... how is it designed and who monitors this label. Exactly by whom was it created and by whom it is administered, awarded and so on, that's what I would like to know." (P12)
	Independent party awarding the label	"Ideally, it would be an overarching body that is, for example, also external and has the competences and the knowledge ... ideally, an NGO that runs it without any vested interest." (P12)
	Recognition of label	"If many companies get involved in using this label. Then I think it could have an impact." (P9)
	Clarity of label criteria	"[The criteria] are totally comprehensible to me, in any case. It's also something that would be important to me if I were to use such a program." (P9)
	Actuality of label	"You could say that the label guarantees that work on AI is ongoing." (P11)
Limitations of certification labels	Unaddressed concerns	"What you could include is a criterion for the AI. That an AI has been used enough times and has, for example, been 99% correct and always had the right answers, rather than 80%." (P4)
	Lack of persuasiveness	"I think there are still a lot of people, or some people, who will be critical of these systems even though it has a label." (P3)
Inhibitors for effective certification labels	Overabundance of labels	"Because you can see that in the organic sector, there are now 20 labels and as a consumer you can almost no longer categorize them, so I think it's so important now that there is also Bio-Suisse [an organic label] or something like that in Switzerland, they have established themselves well, but I think you always have to stick to that as a label." (P6)
	Vacuousness of label criteria	"I find these four points are so common. And bad news is, maybe we don't really analyze what is written. Or don't even read. I can't speak of everyone, but speaking of myself. I often just don't read that message. Beautiful words, but all blah blah blah." (P2)
	Subjectivity of label criteria	"Yes, so what is complete transparency? That brings us back to fairness ... what is fair? These are all such subjective terms that, in my eyes, you can't use like in natural sciences - where you calculate and then there's a result - it's soft science where you're working in." (P5)
	Overlaps of label criteria	"Overlap; I think it all goes a bit in a similar direction, except maybe the last point [Fair User Interaction], which is a bit different again." (P10)

to our current research objective. Categories may consist of further subcategories. Table 1 contains the subcategories and corresponding example quotes from end-users' attitudes toward certification labels. The complete content analysis with all topics is available on the digital repository.

6.1.1 Opportunities and facilitators. Participants in the interview study indicated that the label covered essential concerns. The content analysis revealed that the topic "concerns, risks, and problems" predominantly consisted of data-related concerns such as data privacy (i.e., protecting data from attack and malicious use), data storage (i.e., how data is handled and stored), and third-party involvement (i.e., unwanted and unknown disclosure of data). Regarding data-related concerns, a certification label for AI systems was perceived as an effective tool to convey compliance with these requirements and hold the certified parties more accountable. In particular, the security and data protection criteria were perceived as minimal standards that must be met for them to consider using AI. Participants emphasized that a certification label provides a certain level of transparency that removes the burden of examining these

criteria from end-users. In addition, they viewed the certification labels and corresponding auditing process as an opportunity for more fairness and to establish standards for AI systems, allowing them to compare products and services critically. The interviewed participants indicated that a certification label could increase their trust for all these reasons.

For a label to be convincing, participants emphasized that additional information regarding the label is needed. This includes information about the label's criteria (i.e., how were they formed?), the auditing process itself (i.e., how were these criteria weighted?), and the auditors (i.e., who was responsible for awarding a label?). Participants also placed a strong emphasis on the independence of the auditing process, noting that the auditors should have no financial ties to or other direct dependencies on the organizations for whose products or services the label is awarded in order not to undermine their credibility. Additionally, participants stressed the importance of widespread participation in the auditing and certification process, as this was deemed necessary for adopting AI standards and the label's credibility. As a crucial factor for the

effectiveness of a certification label, participants identified regular updates that align with industry standards and best practices to ensure that the label remains relevant and useful.

6.1.2 Limitations and inhibitors. While participants acknowledged that a certification label covers essential issues, they also noted that it does not address all their AI-related concerns. These concerns included the lack of model performance (e.g., accuracy measures). Some participants noted that a certification label alone could even lead to "blind trust" in AI systems without accuracy measures. Additionally, participants noted that while a certification label provides some level of transparency, it does not provide complete documentation (e.g., source code) of the AI system and the ethical reasoning behind the auditors' decision to approve the use of AI in a particular application in the first place. As a result of these limitations, participants felt that a certification label might not be sufficiently persuasive to convey trustworthiness for critical individuals.

Furthermore, participants identified several reasons why a certification label may not be effective. One reason was a potential overabundance of labels with different standards, diluting compliance with regulations and leading to confusion among end-users. In line with this, participants emphasized the importance of ensuring that the label's criteria are not just "empty promises" but that they are actually adhered to by organizations. They also pointed out the difficulty of measuring the label's criteria and the degree of subjectivity involved. Concepts such as security and fairness can mean different things to different people. Results showed that some criteria were more easily understood (e.g., security) than others (e.g., fair user interaction). For example, 11/12 participants implied that the definition of the security criteria covered what they had in mind. For data protection, this was the case for 9/12 participants, followed by 8/12 participants for reliability. However, merely 2/12 participants indicated that the criterion "fair user interaction" captured what they thought it would encompass. In addition to these differences in comprehension, participants pointed out conceptual overlaps for some criteria (e.g., security and data protection) that were not readily understood without further clarification. All these factors might diminish the effectiveness of a certification label.

6.2 Effects of certification labels

Participants in the survey study were asked to select one case each from the high-stake (medical diagnosis, hiring procedure, loan approval) and one from the low-stake (music preference, route planning, price comparison) scenarios without explicitly being informed of this distinction. Validation of this distinction between low- and high-stake was provided by participants' "impactfulness" rankings. Calculating a mode revealed that the three high-stake scenarios were perceived as the most impactful ones (i.e., 1 = medical diagnosis, 2 = hiring process, 3 = loan approval, 4 = price comparison, 5 = music preference, 6 = route planning). The majority of participants indicated that they would be more likely to accept the AI's decision/recommendation in low-risk scenarios (74.2%, $n = 224$) than in high-risk scenarios (17.9%, $n = 54$) and 7.9% ($n = 24$) indicating no preference, which we considered an additional confirmation of the distinctiveness of the two scenarios. Participants in the interview study distinguished between low- and high-stakes scenarios primarily on the level of risk associated with the scenario. They

reported that high-stakes scenarios carry higher self-relevance and long-term consequences.

Before being presented with the certification label, participants reported both higher trust ($M = 66.72$, $SD = 24.27$) and willingness to use ($M = 71.54$, $SD = 25.54$) ratings for the low-stake scenarios, compared to ratings in high-stake scenarios for trust ($M = 49.37$, $SD = 30.76$) and willingness to use ($M = 52.89$, $SD = 32.63$). After being presented with the certification label, participants' trust and willingness to use ratings revealed statistically significant increases in both low- and high-stakes scenarios (see Figure 2). A dependent Student's t -test indicated that the presence of a certification label resulted in the highest increase for trust ($M_{\Delta} = 9.12$, $SD = 17.92$, $t(301) = 8.84$, $p < .001$) and willingness to use ($M_{\Delta} = 8.41$, $SD = 17.69$, $t(301) = 8.26$, $p < .001$) ratings in high-stake scenarios, followed by trust ($M_{\Delta} = 6.57$, $SD = 13.26$, $t(301) = 8.61$, $p < .001$) and willingness to use ($M_{\Delta} = 4.60$, $SD = 17.03$, $t(301) = 4.70$, $p < .001$) ratings in low-stake scenarios. Hedges' g for effect sizes ranged between .27 - .51 and can thus be considered small (for low-stake scenarios) to medium (for high-stake scenarios) [55].

The different ratings depending on low- and high-stake scenarios become evident when considering the violin plots and boxplots (see Figure 2). The ratings for high-stake scenarios are relatively symmetrically distributed across the scale. In contrast, the low-stake scenarios' distribution is heavily left-skewed, with approximately 75% of the data above a rating of 50 for trust and willingness to use. Introducing a certification label for both scenarios leads to a further shift of the distribution to the right and, thus, higher ratings. Plotting the non-aggregated scenarios individually reveals the distributional differences more clearly (see Figure 3). The ratings of the individual high-stakes scenarios are more spread out on the scale than in the case of the low-stake scenarios. Differences in the effectiveness of a label also become apparent from this perspective. The median trust and willingness to use ratings in all scenarios increases in the presence of a label and are more pronounced in the high-stake scenarios.

A majority of the survey participants directly indicated that they would prefer the use of a certification label in the selected high-stake scenario (63.2%, $n = 191$), compared to preferring a label in the low-stake scenarios (22.2%, $n = 67$), with 14.6% ($n = 44$) of participants indicating no preference. Regarding the different preferences for certification labels in low- and high-stake scenarios, participants from the interview study expressed a greater demand for a certification label in high-stake scenarios because of the higher scenario complexity, limited individual expertise, and a lack of prior experience with the system. Overall, 81.1% ($n = 245$) of survey participants stated a preference for using an AI with a label, compared to 6% ($n = 18$) that would prefer to use an AI without a label and 12.9% ($n = 39$) that stated no preference. Also, 70.9% ($n = 214$) indicated to be more likely to accept an AI's decision/recommendation if it had received a label, compared to 14.2% ($n = 43$) that indicated "no," and 14.9% ($n = 45$), that made no statement. Survey participants rated the importance of the existing label criteria in the context of AI at a high level with similar ratings for security ($M = 87.72$, $SD = 20.93$), data protection ($M = 85.04$, $SD = 21.81$), reliability ($M = 76.97$, $SD = 23.19$) and fair user interaction ($M = 80.80$, $SD = 23.37$). However, merely 55.3% ($n = 167$) of the participants agreed that the label addresses the concerns/challenges/risks they

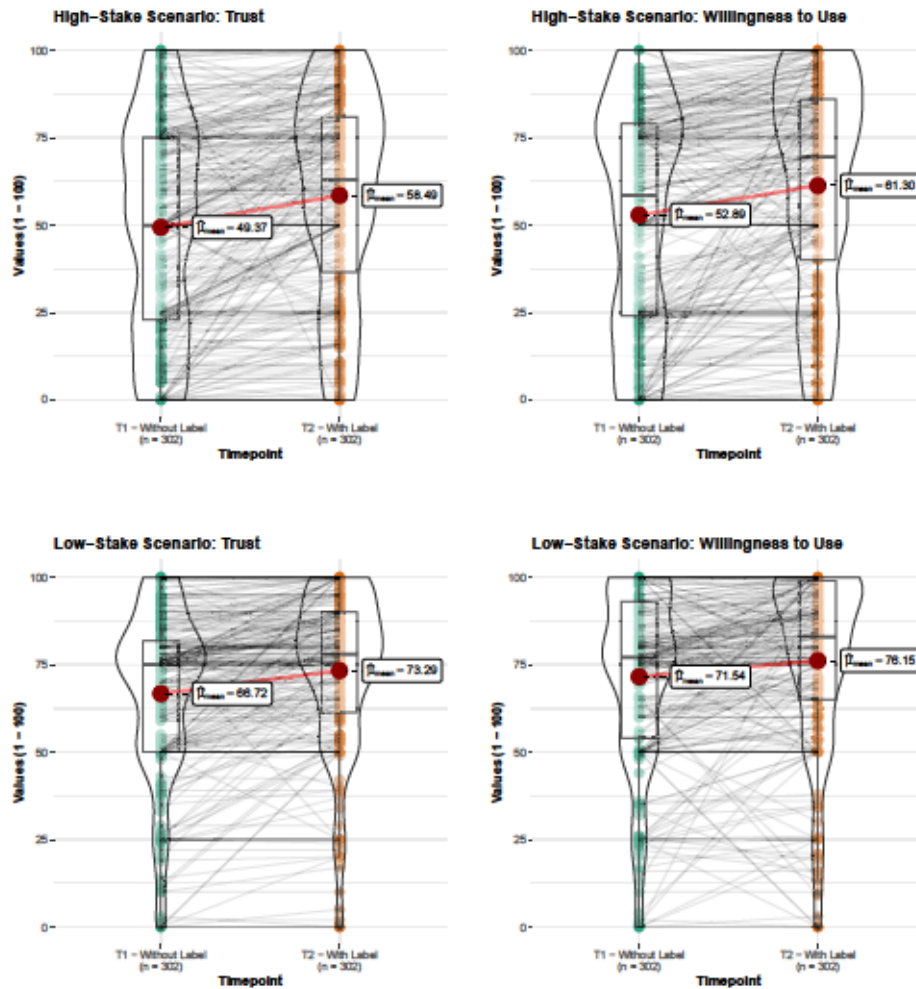


Figure 2: Plots showing the individual scores for trust and willingness to use and their respective changes from T1 (without label) to T2 (with label). The plots also depict the medians, means, and distribution of the aggregated low- and high-stake scenarios. All comparisons revealed statistically significant differences.

see that come with the use of AI, while 20.9% ($n = 63$) stated "no" and 23.8% ($n = 72$) indicated that no statement was possible.

When being asked the question of who should be responsible for awarding a label, the open-ended responses from the survey revealed that a majority of participants expressed a preference for external entities to conduct the auditing, with 48.7% ($n = 147$) of the answers being coded as "government" and 37.4% ($n = 113$) as "NGO." Only 5.3% ($n = 16$) of the answers were coded as "company." Additionally, 8.6% ($n = 26$) of the responses were coded as "other," which included mentions of entities such as "ethic committee," "consumer protection," or "citizen's association."

7 DISCUSSION

The quantitative findings reveal that the presence of a certification label significantly increases participants' trust and willingness to

use AI in *both* low- and high-stake scenarios, thereby answering our second research question. Most participants (81%) of the census-representative survey preferred using AI with a certification label, and a large proportion of participants (71%) responded that they would be more likely to accept an AI's decision or recommendation if it had been awarded a certification label. The results further show that a majority of participants (63%) not only indicated a preference for certification labels in high-stake scenarios, but that certification labels also had a larger effect on trust and willingness to use AI in high-stake scenarios. For example, willingness to use ratings for the "hiring procedure" scenario increased from 36 to 64 points, compared to an increase from 75 to 80 points for the "price comparison" scenario. While Stuurman and Lachaud and the EU's "white paper on artificial intelligence" distinguish between regulating high-stake

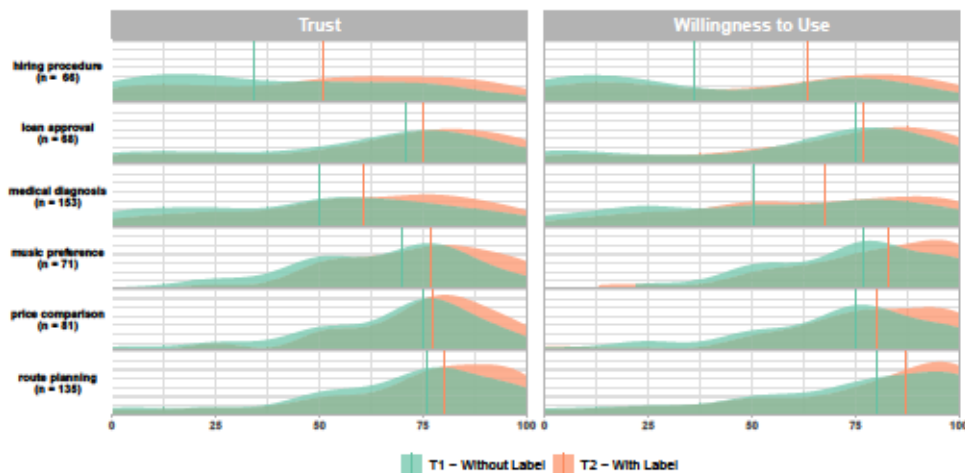


Figure 3: Plots showing the different distributions for trust and willingness to use ratings for the different high-stake (hiring procedure, loan approval, medical diagnosis) and low-stake (music preference, price comparison, route planning) without a label at T1 and with a label at T2.

AI through mandatory requirements and proposed voluntary labeling only for low-stake AI, our results demonstrate the relevance of certification labels for end-users, specifically in high-stake scenarios. Based on these findings, we argue that parallel to voluntary labeling for low-stake AI scenarios, compliance with mandatory requirements for AI in high-stake scenarios could also be communicated through certification labels, potentially increasing end-users' trust in and willingness to use awarded AI systems.

Qualitative findings allowed us to answer our first research question and provide a more nuanced picture of which aspects to consider for effective certification labels in the context of AI. The certification label we investigated in this study was designed for digital trust more generally. However, end-users' attitudes toward the certification label were primarily positive, and the label's criteria of security, data protection, reliability, and fair user interaction were also relevant to end-users in the context of AI. We derive this from survey participants' high "importance" ratings for the existing label criteria. Concerning *opportunities* for AI labels, participants in the interview study indicated that a certification label could increase perceived transparency and fairness and serve as a means to establish standards for AI systems. It became apparent from the interviews that certification labels can especially cover end-users' data-related concerns (e.g., privacy, data protection, and third-party involvement) that map to previous work [65].

However, our results also reveal that certification labels have *limitations* and do not alleviate all issues end-users face regarding the use of AI. Only half of the participants in the survey indicated that a certification label addresses their AI-related concerns/challenges/risks, suggesting that end-users seem to hold differentiated needs. For example, participants in our interviews pointed out that a certification label does not provide indicators about the AI's performance (e.g., accuracy measures). They remarked that performance indicators are essential in deciding in

which cases the AI can be trusted and when it must be questioned. This led participants to remark that a label could inadvertently foster "blind trust" if performance indicators are absent. Thus, we suggest that certification labels should either include performance indicators as part of the label criteria or be supplemented with them. Based on these results, we argue that certification labels can more readily signal trustworthiness than untrustworthiness. This is because it is not possible to distinguish if a digital product or service has not yet been audited or whether it has failed to meet specific audit criteria, particularly if certification labels remain voluntary. We regard certification labels as *one* component of an "AI trustworthiness ecosystem" [2] that meets essential needs for end-users but which ideally should be combined with other transparency approaches to signal untrustworthiness (e.g., accuracy measures) and form a "chain of trust" [65].

As potential *inhibitors* for effective certification labels, participants in our interviews pointed out certain overlaps and the subjective nature of the label's criteria. Ultimately, "fairness" and "security" are subjective judgments that vary from one person to the next, and our results showed that the criterion "fair user interaction," in particular, did not reflect what study participants thought it encompassed. The challenge for auditing of defining and measuring concepts that are inherently difficult to quantify has been discussed by previous research [37, 58, 66]. Our results indicate that this subjectivity is recognized by end-users and can impair the effectiveness of a label. To avoid a discrepancy between, for example, the auditors' definition of fairness and what people commonly associate with this term, auditors should be in dialogue with end-users so that their values are represented in a label. This is in line with Costanza-Chock et al., who had criticized that the involvement of affected communities plays a minor role in AI audits. They argued that real-world harms and sociological phenomena could only be understood by engaging with people to inform auditing.

Our interview results highlight that end-users request not only information on the label's criteria but also information regarding the criteria content (i.e., how they were formed), the auditing process itself (i.e., how the criteria informed the audit), and particularly about the auditors (i.e., who awarded the label). We identified this demand for additional information as a potential *facilitator*, indicating that an effective certification label is more than just a list of evaluation criteria. A large majority (86%) of survey participants responded that either the government (49%) or a non-governmental organization (37%) should ideally be responsible for awarding a label, with only 5.3% of responses indicating that a company should be responsible. Participants in the interview study emphasized the auditors' independence (e.g., financially, with no conflict of interest) as a prerequisite for the effectiveness of a certification label. These findings support the notion that auditing can only foster trust if the auditors themselves are trusted [2] and are in line with results of label studies in other domains [23, 64], which show that third-party certification positively affects trust in eco-labels. We contribute to the ongoing discussion regarding internal vs. external auditing by showing that end-users favor independent auditors. To account for this independence on the one hand and the structural advantages of internal audits on the other, "cooperative audits" [69] could be a way forward, balancing between the advantages and challenges of the two approaches. In addition to these facilitators and inhibitors, auditors and regulators should also be mindful that an overabundance of labels with different standards can inhibit the persuasiveness and trustworthiness of their certification label. Such effects have been reported for eco-labels, where an extensive number of existing labels result in different standards that remain unclear to consumers [26]. These findings speak for a certain harmonization and regulation of certification labels. Moreover, organizational compliance with a label's criteria should be established so end-users do not perceive them as "empty promises" but instead as a means for increased accountability for organizations and more trustworthy AI [37]. A prominent instance of such a challenge is the case of the CE (conformité européenne) marking, in which some products use the mark without actually being manufactured to EU quality standards [45]. This illegitimate use has led, among other things, to the introduction of supplementary certification labels to certify product quality, which unintentionally contribute to consumer confusion [61]. To realize their full potential, certification labels should have a thorough auditing process, be regularly updated to reflect current industry standards, and ideally, be used by a wide range of organizations to increase recognition.

8 LIMITATIONS AND FUTURE WORK

We conducted a within-subjects survey study where participants were presented with the AI scenarios with and without a certification label. While this provided valuable insights into the general effectiveness of certification labels, future work could compare label classes or designs (e.g., nutrition labels vs. certification labels) in a between-subjects experimental design. Certification labels are limited in their ability to communicate untrustworthiness. While other kinds of labels have a more differentiated rating system (e.g., color-codings or grades) that allows comparisons, certification labels only provide dichotomous information by either being present

or not. Thus, it is not possible to differentiate if a product without a certification label is untrustworthy because it failed to meet a label's criteria or has yet to be audited. A between-subjects design could provide evidence about the effectiveness of different kinds of labels and identify the factors that make labels more or less effective in communicating trustworthiness and untrustworthiness.

Moreover, we used single-item questions to measure trust and willingness to use. Trust, in particular, is a complex psychological construct [56] and might not be adequately operationalized using single-items measures. However, a recent study has shown that single-item trust measures are equivalent to validated questionnaires regarding sensitivity to changes in trust and a reliable tool in longer surveys where questionnaires are not feasible [50]. Future work should confirm the effectiveness of certification labels in fostering trust with validated psychometric measures and explore their effect on trusting dynamics that emerge over time in real-world human-AI interactions.

9 CONCLUSION

This study empirically investigated certification labels to communicate trustworthy AI to end-users. For this purpose, we explored end-users' attitudes toward certification labels in the context of AI and how labels affect trust and willingness to use AI in both low- and high-stakes scenarios. We used a mixed-methods approach to collect both qualitative and quantitative data through interviews ($N = 12$) and a census-representative survey ($N = 302$) with end-users. The quantitative results of this study show that certification labels can be a promising way to communicate the outcome of audits to end-users, increasing both trust and willingness to use AI in low- and high-stake AI scenarios. Based on the qualitative findings, we further identified opportunities and limitations of certification labels, as well as inhibitors and facilitators for the effective design and implementation of certification labels. Our work provides the first empirical evidence that labels may be a promising constituent in the more extensive "trustworthiness ecosystem" for AI.

10 FUNDING, DECLARATION OF CONFLICTING INTERESTS AND DATA AVAILABILITY

This research was primarily funded by an independent research group, but additional funding (CHF 2,500.00) was granted by the Swiss Digital Initiative, an independent non-profit foundation, to obtain a representative sample. The entire research process, including the development of the research design, data analysis, interpretation of the results, and the writing of this paper, was conducted exclusively by independent researchers with no other affiliations with the Swiss Digital Initiative Foundation than those mentioned here. All data, corresponding R-scripts, and supplementary materials are available on OSF: <https://osf.io/gzp5k/>.

ACKNOWLEDGMENTS

Special thanks to Ariane Haller and the Swiss Digital Initiative for the permission to use their label for the purpose of our study, especially Nicolas Zahn, who was our contact person at the foundation.

REFERENCES

- [1] M. Arnold, R. K. E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilović, R. Nair, K. Natesan Ramamurthy, A. Olteanu, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, and K. R. Varshney. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4/5 (2019), 6:1–6:13. <https://doi.org/10.1147/JRD.2019.2942288>
- [2] Shahar Avin, Haydn Belfield, Miles Brundage, Gretchen Krueger, Jasmine Wang, Adrian Weller, Markus Anderljug, Igor Krawczuk, David Krueger, Jonathan Lebensold, et al. 2021. Filling gaps in trustworthy development of AI. *Science* 374, 6573 (2021), 1327–1329. <https://doi.org/10.1126/science.abi7176>
- [3] Jack Bandy. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 74 (apr 2021), 34 pages. <https://doi.org/10.1145/3449148>
- [4] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. 2020. Explainable Machine Learning in Deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) (FAT* '20). Association for Computing Machinery, New York, NY, USA, 648–657. <https://doi.org/10.1145/3351095.3375624>
- [5] Reuben Binns, Max van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage'. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, Regan Mandryk, Mark Hancock, Mark Perry, and Anna Cox (Eds.). ACM, New York, NY, USA, 1–14. <https://doi.org/10.1145/3173574.3173951>
- [6] Florian Brühlmann, Serge Petralito, Lena Aeschbach, and Klaus Opwis. 2020. The Quality of Data Collected Online: An Investigation of Careless Responding in a Crowdsourced Sample. *Methods in Psychology* 2 (2020), 100022. <https://doi.org/10.1016/j.metip.2020.100022>
- [7] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian K. Hadfield, Heidy Khlaaf, Jingyong Yang, Helen Toner, Ruth Fong, Tegan Maharaj, Pang Wei Koh, Sara Hooker, Jade Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O'Keefe, Mark Koren, Théo Riffel, J. B. Rubinovitz, Tamay Besiroglu, Federica Carugati, Jack Clark, Peter Eckersley, Sarah de Haas, Maritza Johnson, Ben Laurie, Alex Ingerman, Igor Krawczuk, Amanda Askell, Rosario Cammarota, Andrew Lohn, David Krueger, Charlotte Stix, Peter Henderson, Logan Graham, Carina Prunkl, Bianca Martin, Elizabeth Seger, Noa Zilberman, Seán Ó hÉigeartaigh, Frens Kroeger, Girish Sastry, Rebecca Kagan, Adrian Weller, Brian Tse, Elizabeth Barnes, Allan Dafoe, Paul Scharre, Ariel Herbert-Voss, Martijn Rasser, Shagun Sodhani, Carrick Flynn, Thomas Krendl Gilbert, Lisa Dyer, Saif Khan, Yousha Bengio, and Markus Anderljug. 2020. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. *CoRR* abs/2004.07213 (2020). <https://arxiv.org/abs/2004.07213>
- [8] Christiano Castelfranchi and Rino Falcone. 2010. *Trust theory: A socio-cognitive and computational model*. John Wiley & Sons. <https://doi.org/10.1002/9780470519851>
- [9] Erin K. Chiou and John D. Lee. 2021. Trusting Automation: Designing for Responsivity and Resilience. *Human Factors: The Journal of Human Factors and Ergonomics Society* 65 (2021), 137 – 165. <https://doi.org/10.1177/00187208211009995>
- [10] European Commission. 2022. *EU Ecolabel facts and figures*. Retrieved February 2, 2023 from https://environment.ec.europa.eu/topics/circular-economy/eu-ecolabel-home/business/ecolabel-facts-and-figures_en
- [11] European Commission. 2020. White Paper on Artificial Intelligence: a European approach to excellence and trust. *Official Journal of European Union* L COM(2020) 65 final (19-02-2020). <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0065&qid=1675254609974>
- [12] European Commission. 2021. Proposal for a Regulation of the European Parliament and of the council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. *Official Journal of European Union* L COM (2021) 206 final (21-04-2021). <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206>
- [13] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. 2022. Who Audits the Auditors? Recommendations from a Field Scan of the Algorithmic Auditing Ecosystem. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) (FAcCT '22). Association for Computing Machinery, New York, NY, USA, 1571–1583. <https://doi.org/10.1145/3531146.3533213>
- [14] Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. 2022. Interactive Model Cards: A Human-Centered Approach to Model Documentation. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) (FAcCT '22). Association for Computing Machinery, New York, NY, USA, 427–439. <https://doi.org/10.1145/3531146.3533108>
- [15] Denmark. 2019. National Strategy for Artificial Intelligence. *Ministry of Finance and Ministry of Industry Business Financial Affairs* (March 2019). https://eng.em.dk/media/13081/305755-gb-version_4k.pdf
- [16] Group Ecocert. 2018. *Organic agriculture Europe*. Group Ecocert. Retrieved Jan 14, 2023 from <https://www.ecocert.com/en/certification-detail/organic-farming-europe-eu-n-848-2018>
- [17] Gregory Falco, Ben Shneiderman, Julia Badger, Ryan Carrier, Anton Dahbura, David Danks, Martin Eling, Alwyn Goodloe, Jerry Gupta, Christopher Hart, et al. 2021. Governing AI safety through independent audits. *Nature Machine Intelligence* 3, 7 (2021), 566–571. <https://doi.org/10.1038/s42256-021-00370-7>
- [18] Luciano Floridi, Matthias Holweg, Mariarosaria Taddeo, Javier Amaya Silva, Jakob Mökander, and Yuni Wen. 2022. capAI-A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act. Available at SSRN: <https://ssrn.com/abstract=4064091> (March 23, 2022). <https://dx.doi.org/10.2139/ssrn.4064091>
- [19] Fraunhofer Institute for Telecommunications and HHI Heinrich Hertz Institute. [n.d.]. *Auditing and Certification of AI Systems*. Retrieved Jan 25 2023 from <https://www.hhi.fraunhofer.de/en/departments/ai/technologies-and-solutions/auditing-and-certification-of-ai-systems.html>
- [20] ForHumanity. 2016. *Independent Audit of AI Systems (IAAIS)*. ForHumanity. Retrieved Jan 25 2023 from <https://forhumanity.center/independent-audit-of-ai-systems/>
- [21] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (nov 2021), 86–92. <https://doi.org/10.1145/3458723>
- [22] Samantha Goodman, Lana Vanderlee, Rachel Acton, Syed Mahamad, and David Hammond. 2018. The impact of front-of-package label design on consumer understanding of nutrient amounts. *Nutrients* 10, 11 (2018), 1624. <https://doi.org/10.3390/nu10111624>
- [23] Matthew Gorton, Barbara Tocco, Ching-Hua Yeh, and Monika Hartmann. 2021. What determines consumers' use of eco-labels? Taking a close look at label trust. *Ecological Economics* 189 (2021), 107173. <https://doi.org/10.1016/j.ecolecon.2021.107173>
- [24] Klaus G Grunert, Sophie Hieke, and Josephine Wills. 2014. Sustainability labels on food products: Consumer motivation, understanding and use. *Food policy* 44 (2014), 177–189. <https://doi.org/10.1016/j.foodpol.2013.12.001>
- [25] Sebastian Hallensleben, Carla Hustedt, Lajla Fetic, Torsten Fleischer, Paul Grünke, Thilo Hagendorff, Marc Hauer, Andreas Hauschke, Jessica Heesen, Michael Herrmann, Rafaela Hillerbrand, Christoph Hubig, Andreas Kaminski, Tobias Krafft, Wulf Loh, Philipp Otto, and Michael Puntschuh. 2020. From Principles to Practice – An interdisciplinary framework to operationalise AI ethics. *Artificial Intelligence Ethics Impact Group* (01 April 2020). <https://www.ai-ethics-impact.org/en>
- [26] Rick Harbaugh, John W Maxwell, and Beatrice Roussillon. 2011. Label confusion: The Groucho effect of uncertain standards. *Management science* 57, 9 (2011), 1512–1527. <https://doi.org/10.1287/mnsc.1110.1412>
- [27] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2020. The dataset nutrition label. *Data Protection and Privacy, Volume 12: Data Protection and Democracy* 12 (2020), 1. <https://doi.org/10.48550/arXiv.1805.03677>
- [28] Ming-Hui Huang and Roland T. Rust. 2018. Artificial Intelligence in Service. *Journal of Service Research* 21, 2 (2018), 155 – 172. <https://doi.org/10.1177/1094670517752459>
- [29] IEEE. 2008. IEEE Standard for Software Reviews and Audits. *IEEE Std 1028-2008* vol., no. (15 Aug. 2008), 1–53. <https://doi.org/10.1109/IEEESTD.2008.4601584>
- [30] Iina Ikonen, Francesca Sotgiu, Aylin Aydinli, and Peeter WJ Verleg. 2020. Consumer effects of front-of-package nutrition labeling: An interdisciplinary meta-analysis. *Journal of the Academy of Marketing Science* 48, 3 (2020), 360–383. <https://doi.org/10.1007/s11747-019-00663-9>
- [31] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. 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* (Virtual Event, Canada) (FAcCT '21). ACM, New York, NY, USA, 624–635. <https://doi.org/10.1145/3442188.3445923>
- [32] Maurice Jakesch, Zana Bućinca, Saleema Amershi, and Alexandra Olteanu. 2022. How Different Groups Prioritize Ethical Values for Responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) (FAcCT '22). Association for Computing Machinery, New York, NY, USA, 310–323. <https://doi.org/10.1145/3531146.3533097>
- [33] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1 (2019), 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
- [34] Alexandra Jones, Bruce Neal, Belinda Reeve, Cliona Ni Mhurchu, and Anne Marie Thow. 2019. Front-of-pack nutrition labelling to promote healthier diets: current practice and opportunities to strengthen regulation worldwide. *BMJ global health* 4, 6 (2019), e001882. <https://doi.org/10.1136/bmjgh-2019-001882>
- [35] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. "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* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 158, 18 pages. <https://doi.org/10.1145/3491102.3517533>
- [36] Davinder Kaur, Suleyman Uslu, Kaley J. Rittichier, and Arjan Durrezi. 2022. Trustworthy Artificial Intelligence: A Review. *ACM Comput. Surv.* 55, 2, Article 39 (jan 2022), 38 pages. <https://doi.org/10.1145/3491209>

- [37] Bran Knowles and John T. Richards. 2021. The Sanction of Authority: Promoting Public Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) (FAcCT '21). Association for Computing Machinery, New York, NY, USA, 262–271. <https://doi.org/10.1145/3442188.3445890>
- [38] P. M. Krafft, Meg Young, Michael Katell, Jennifer E. Lee, Shankar Narayan, Micah Epstein, Dharma Dailey, Bernese Herman, Aaron Tam, Vivian Guetler, Corinne Bintz, Daniella Raz, Pa Ousman Jobe, Franziska Putz, Brian Robick, and Bissan Barghouti. 2021. An Action-Oriented AI Policy Toolkit for Technology Audits by Community Advocates and Activists. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) (FAcCT '21). Association for Computing Machinery, New York, NY, USA, 772–781. <https://doi.org/10.1145/3442188.3445938>
- [39] Markus Langer, Daniel Oster, Timo Speith, Lena Kästner, Holger Hermanns, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What Do We Want From Explainable Artificial Intelligence (XAI)? A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research. *Artificial Intelligence* 296 (Feb. 2021), 103473. <https://doi.org/10.1016/j.artint.2021.103473>
- [40] John D. Lee and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (March 2004), 50–80. <https://doi.org/10.1518/hfes.46.1.50.30392>
- [41] Bruno Lepri, Nuria Oliver, Emmanuel Letouze, Alex 'Sandy' Pentland, and Patrick Vinck. 2018. Fair, Transparent, and Accountable Algorithmic Decision-making Processes. *Philosophy & Technology* 31 (2018), 611–627. <https://doi.org/10.1007/s13347-017-0279-x>
- [42] Lan Li, Tina Lassiter, Joohee Oh, and Min Kyung Lee. 2021. Algorithmic Hiring in Practice: Recruiter and HR Professional's Perspectives on AI Use in Hiring. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) (AI/ES '21). Association for Computing Machinery, New York, NY, USA, 166–176. <https://doi.org/10.1145/3461702.3462531>
- [43] QVera Liao and S. Shyam Sundar. 2022. Designing for Responsible Trust in AI Systems: A Communication Perspective. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) (FAcCT '22). Association for Computing Machinery, New York, NY, USA, 1257–1268. <https://doi.org/10.1145/3531146.3533182>
- [44] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). ACM, New York, NY, USA, 1–15. <https://doi.org/10.1145/3313831.3376590>
- [45] Consumer Research Associates Ltd. 2007. *EFTA study on certification and marks in Europe*. Retrieved May 4, 2023 from <https://www.efta.int/sites/default/files/publications/study-certification-marks/executeive-summary.pdf>
- [46] Malta. 2019. Malta the ultimate AI Launchpad: a strategy and vision for Artificial Intelligence in Malta 2030. *Parliamentary Secretariat for Financial Services Digital Economy Innovation* (October 2019). https://malta.ai/wp-content/uploads/2019/11/Malta_The_Ultimate_AI_Launchpad_vFinal.pdf
- [47] Philipp Mayring and Thomas Fenzl. 2019. Qualitative Inhaltsanalyse. In *Handbuch Methoden der empirischen Sozialforschung*, N. Baur and J. Blasius (Eds.). Springer VS, Wiesbaden, (pp. 633–648). https://doi.org/10.1007/978-3-658-21308-4_42
- [48] Adam W Meade and S Bartholomew Craig. 2012. Identifying careless responses in survey data. *Psychological methods* 17, 3 (2012), 437. <https://doi.org/10.1037/a0028085>
- [49] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) (FAT* '19). Association for Computing Machinery, New York, NY, USA, 220–229. <https://doi.org/10.1145/3287560.3287596>
- [50] Birthe Nessel, Gnanathusharan Rajendran, José David Aguas Lopes, and Helen Hastie. 2022. Sensitivity of Trust Scales in the Face of Errors. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction* (Sapporo, Hokkaido, Japan) (HRI '22). IEEE Press, 950–954. <https://doi.org/10.1109/HRI53351.2022.9889427>
- [51] Indrajeet Patil. 2021. statsExpressions: R package for tidy dataframes and expressions with statistical details. *Journal of Open Source Software* 6, 61 (2021), 3236. <https://doi.org/10.21105/joss.03236>
- [52] Stefano Puntoni, Rebecca Walker Rezek, Markus Giesler, and Simona Botti. 2020. Consumers and Artificial Intelligence: An Experiential Perspective. *Journal of Marketing* 85, 1 (2020), 131–151. <https://doi.org/10.1177/0022242920953847>
- [53] R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
- [54] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) (FAT* '20). Association for Computing Machinery, New York, NY, USA, 33–44. <https://doi.org/10.1145/3351095.3372873>
- [55] Shlomo S Sawilowsky. 2009. New effect size rules of thumb. *Journal of modern applied statistical methods* 8, 2 (2009), 467–474. <https://doi.org/10.22237/jmasm/1257035100>
- [56] Nicolas Scharowski, Sebastian AC Perrig, Nick von Felten, and Florian Brühlmann. 2022. Trust and Reliance in XAI—Distinguishing Between Attitudinal and Behavioral Measures. *CHI TRAIT Workshop* (2022), 6 pages. <https://doi.org/10.48550/arXiv.2203.12318>
- [57] Nadine Schlicker, Alarith Uhde, Kevin Baum, Martin C Hirsch, and Markus Langer. 2022. Calibrated Trust as a Result of Accurate Trustworthiness Assessment—Introducing the Trustworthiness Assessment Model. (2022). <https://doi.org/10.31234/osf.io/qhwvx>
- [58] Christin Seifert, Stefanie Scherzinger, and Lena Wiese. 2019. Towards Generating Consumer Labels for Machine Learning Models. In *2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI)* (Los Angeles, CA, USA). IEEE, 173–179. <https://doi.org/10.1109/CogMI48466.2019.00033>
- [59] Marcel Stadelmann and Renate Schubert. 2018. How do different designs of energy labels influence purchases of household appliances? A field study in Switzerland. *Ecological economics* 144 (2018), 112–123. <https://doi.org/10.1016/j.ecolecon.2017.07.031>
- [60] Bernd Carsten Stahl and David Wright. 2018. Ethics and privacy in AI and big data: Implementing responsible research and innovation. *IEEE Security & Privacy* 16, 3 (2018), 26–33. <https://doi.org/10.1109/MSP.2018.270116>
- [61] Kees Stuurman and Eric Lachaud. 2022. Regulating AI: A label to complete the proposed Act on Artificial Intelligence. *Computer Law & Security Review* 44 (2022), 105657. <https://doi.org/10.1016/j.clsr.2022.105657>
- [62] Khan MR Taufique, Kristian S Nielsen, Thomas Dietz, Rachael Shwom, Paul C Stern, and Michael P Vandenberg. 2022. Revisiting the promise of carbon labelling. *Nature Climate Change* 12, 2 (2022), 132–140. <https://doi.org/10.1038/s41558-021-01271-8>
- [63] Frauke Mattison Thompson, Sven Tuzovic, and Corina Braun. 2019. Trustmarks: Strategies for exploiting their full potential in e-commerce. *Business Horizons* 62, 2 (2019), 237–247. <https://doi.org/10.1016/j.bushor.2018.09.004>
- [64] Emma Tonkin, Annabelle M Wilson, John Coveney, Trevor Webb, and Samantha B Meyer. 2015. Trust in and through labelling—a systematic review and critique. *British Food Journal* 117, 1 (2015), 318–338. <https://doi.org/10.1108/BJF-07-2014-0244>
- [65] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad van Moorsel. 2020. The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 272–283. <https://doi.org/10.1145/3351095.3372834>
- [66] Briana Vecchione, Karen Levy, and Solon Barocas. 2021. Algorithmic Auditing and Social Justice: Lessons from the History of Audit Studies. In *Equity and Access in Algorithms, Mechanisms, and Optimization* (–, NY, USA) (EAAMO '21). Association for Computing Machinery, New York, NY, USA, Article 19, 9 pages. <https://doi.org/10.1145/3465416.3483294>
- [67] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5 (2021), 1–39. <https://doi.org/10.1145/3476068>
- [68] Ruotong Wang, F. Maxwell Harper, and Haiyi Zhu. 2020. Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–14. <https://doi.org/10.1145/3313831.3376813>
- [69] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. 2021. Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) (FAcCT '21). Association for Computing Machinery, New York, NY, USA, 666–677. <https://doi.org/10.1145/3442188.3445928>
- [70] Sarah Woods, Michael Walters, Kheng Lee Koay, and Kerstin Dautenhahn. 2006. 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. (Istanbul, Türkiye)*. IEEE, 750–755. <https://doi.org/10.1109/AMC.2006.1631754>
- [71] Kun-Hsing Yu, Andrew Beam, and Isaac S. Kohane. 2018. Artificial intelligence in healthcare. *Nature Biomedical Engineering* 2 (2018), 719–731. <https://doi.org/10.1038/s41551-018-0305-z>
- [72] Mireia Yurrita, Dave Murray-Rust, Agathe Balayn, and Alessandro Bozzon. 2022. Towards a Multi-Stakeholder Value-Based Assessment Framework for Algorithmic Systems. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) (FAcCT '22). Association for Computing Machinery, New York, NY, USA, 535–563. <https://doi.org/10.1145/3531146.3533118>

A APPENDIX

A.1 High-stake Scenarios

A.1.1 Medical Diagnosis. Consider the situation where you are searching for potential medical diagnoses. Your insurance is using an AI system called MyHealth for evaluating medical symptoms. You will be required to fill out a form, uploading your medical history, and submit them along with personal information like age, gender, marital status and employment status to MyHealth. Once assessed, MyHealth will determine based on the provided information what your medical diagnosis is.

A.1.2 Hiring Procedure. 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.

A.1.3 Loan Approval. Consider the situation where you are applying for a loan at a bank. The bank is using an AI system called MyLoans for evaluating loan applications. You will be required to fill out a form, specifying the loan amount, and submit them along with personal information like marital status, employment

status, annual income and financial history to MyLoans. Once assessed, MyLoans will determine based on the provided information whether your loan application is successful or not.

A.2 Low-stake Scenarios

A.2.1 Music Preference. Consider the situation where you want to explore new music. You are using an AI system called MyMusik for evaluating your music preference. You will be required to accept terms and conditions of MyMusik which among other things include analyzing your search behavior and already liked songs. Once assessed, MyMusik will provide you with song recommendations.

A.2.2 Route Planning. Consider the situation where you want to get from one place to another place. You are using an AI system called MyMap for evaluating your travelling route. You will be required to accept terms and conditions of MyMap which among other things include analyzing your motion data and already visited places. Once assessed, MyMap will provide you with a route recommendation.

A.2.3 Price Comparison. Consider the situation where you want to sell your car. You are using an AI system called MyCar for evaluating a selling price. You will be required to accept terms and conditions of MyCar which among other things include analyzing your search history on the platform and already sold cars. Once assessed, MyCar will provide you with a selling price recommendation.