

In Seal We Trust? Investigating the Effect of Certifications on Perceived Trustworthiness of AI Systems

Magdalena Wischnewski¹ , Nicole Krämer^{1,2} , Christian Janiesch³ ,
Emmanuel Müller^{1,4} , Theodor Schnitzler¹ , and Carina Newen¹ 

1 Research Center for Trustworthy Data Science and Security, Dortmund, Germany

2 Social Psychology: Media and Communication, University of Duisburg-Essen, Duisburg, Germany

3 Enterprise Computing, Technical University Dortmund, Dortmund, Germany


4 Data Science and Data Engineering, Technical University Dortmund, Dortmund, Germany

Abstract

Trust certification through so-called trust seals is a common strategy to help users ascertain the trustworthiness of a system. In this study, we examined trust seals for AI systems from two perspectives: (1) In a pre-registered online study with $N = 453$ participants, we asked whether trust seals can increase user trust in AI systems, and (2) qualitatively, we investigated what participants expect from such AI seals of trust. Our results indicate mixed support for the use of AI seals. While trust seals generally did not affect the participants' trust, their trust in the AI system increased if they trusted the seal-issuing institution. Moreover, although participants understood verification seals the least, they desired verifications of the AI system the most.

Keywords: artificial intelligence, seals of trust, epistemic trust, transparency, formal verification

Notes: We have no conflict of interest to report. This work has been supported by the Research Center Trustworthy Data Science and Security (<https://rc-trust.ai>), one of the Research Alliance centers within the <https://uaruhr.de>

CONTACT Magdalena Wischnewski  • magdalena.wischnewski@tu-dortmund.de • Research Center Trustworthy Data Science and Security • Joseph-von-Fraunhofer-Straße 25 • 44227 Dortmund, Germany

ISSN 2638-602X (print)/ISSN 2638-6038 (online)
www.hmcjournal.com



Copyright 2024 Authors. Published under a Creative Commons Attribution 4.0 International (CC BY-NC-ND 4.0) license.

Introduction

Artificial intelligence (AI) systems are ubiquitous and have become integral to everyday professional and private life. AI systems such as Open AI's ChatGPT or Google's BERT can generate meaningful text (Feuerriegel et al., 2024), other AI systems are components in safety-critical applications such as those that enable autonomous driving (Grigorescu et al., 2020), and even further, AI systems process highly sensitive health information such as echocardiograms (Madani et al., 2018). Simultaneously, these systems and their underlying building blocks, such as deep learning models, have become very complex, aggravating the so-called black box phenomenon. Consequently, knowing when (not) to trust an AI system can be challenging for different stakeholders, from users to decision-makers and even developers. While efforts to develop inherently trustworthy AI systems are much needed, approaches solely focusing on technical aspects are insufficient, as trust results from a system's perceived rather than its actual trustworthiness. Consequently, users sometimes perceive a system inappropriately, placing either too much or too little trust in an AI system.

To help users' trust calibration, different paths can be taken. One popular and well-researched example is explainable AI (XAI), which aims to increase an AI systems' intelligibility by providing explanations for the system's behavior, making internal processes visible, and increasing the overall transparency of the system (Arrieta et al., 2020). Typical methods of XAI are, for example, visual explanations such as heat maps, which highlight areas of input data that were most influential for the system's output, or textual explanations which provide written or oral statements of the explainer. However, XAI is no panacea to cure a lack of trust, and concerns have been raised in terms of users' cognitive biases (Bertrand et al., 2022) and the cognitive burden that explanations pose on users when explanations are not designed with the end-user in mind (Miller, 2019).

In this paper, we aim to counter the shortcomings of XAI and tackle the problem of trust from a different perspective. We empirically explore the effects of AI certifications, so-called *AI seals of trust*. Such seals are credentials which certify that software has been tested and validated to meet specific predefined criteria or standards in various dimensions. Theoretically grounded in works on epistemic trust, trust theory, signaling theory, and persuasion literature, we examined the effects of three different AI seals of trust in a quantitative online experiment. To do so, participants of our study either viewed an AI system with (experimental groups) or without (control group) an AI seal of trust. In addition, in a qualitative part we asked participants in an open-ended format about their preferences for AI certification.

The importance of this work is underlined by initiatives such as the EU AI Act, which suggests certification as a central mechanism to communicate to the public the compliance with industry and legislative requirements. To date, however, empirical studies investigating the effects of such certifications for AI systems are scarce.

Theoretical Background

From Trust in AI to Calibrated Trust in AI

To describe and define *trust in AI*, previous work builds on thoughts from various disciplines, such as philosophy, sociology, and psychology that predominantly examine trust as

an interpersonal judgment between two or more individuals. Moreover, choosing interpersonal trust as a starting point to examine trust in AI seems sensible as humans, at times, react socially to machines (Nass & Moon, 2000). In fact, the most widely adopted definition of trust in automation originates in Mayer et al.'s (1995) dyadic model of organizational trust, in which trust results from a person's (the trustor) perceptions of another person's (the trustee) ability, benevolence, and integrity. While the direct application of an interpersonal trust conceptualization might be appropriate for certain occasions, this is not always the case (Madhavan & Wiegmann, 2007). Hence, emanating from Mayer et al.'s ability-benevolence-integrity framework, Lee and See (2004) postulate that for a person to trust a machine, the person needs to assess the perceived reliability and functionality of an AI (ability = performance), the intentions with which it was built (benevolence = purpose), and the intelligibility of AI (integrity = process). Beyond these three trust antecedents, Lee and See (2004) define trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (p. 54).

Hence, users' trust must be appropriately calibrated to the system's actual trustworthiness (Lee & See, 2004; Madhavan & Wiegmann, 2007; Parasuraman & Riley, 1997). As described above, users' trust depends on various factors, such as the system's overall performance or the perceived integrity of the system. However, cognitive and social psychology insights suggest that users' perceptions can be distorted, possibly leading users to place too little or too much trust in a system. Such a mismatch of the perceived and actual system trustworthiness can result in either the system's disuse (i.e., resistance to use the system) or the system's misuse (over-reliance on the system). Both disuse and misuse pose serious consequences. In the context of semi-automated driving, for example, ignoring and over-relying on autopilot has led to deadly incidents.¹ Hence, reaching calibrated user trust is essential.

To calibrate user trust, different approaches have been taken. Wischniewski et al. (2023) offer a systematic overview of previous approaches. In their work, the authors surveyed different empirical, human-centered interventions to match perceived and actual system trustworthiness for automated systems accurately. Many of the interventions reviewed aim to increase a system's transparency, assisting the users' trust assessments by making the system more intelligible. While some interventions successfully calibrated the users' trust in a system, in some cases, the intervention also increased the users' workload (Kunze et al., 2019) or led to overtrust (Yeh & Wickens, 2001). In addition, adding, for example, explanations for increasing transparency had adversarial effects, eroding the users' trust, which Kizilcec (2016) explained by arguing that the additional information might have been confusing for users, reducing their understanding instead of increasing transparency.

Even though these transparency interventions have shown mixed effects, there are other reasons to question these approaches. First, many interventions are not developed for end-users but for developers themselves to make the inner workings of AI more transparent (Miller, 2019). However, explanations are likely to be less successful without the end-users in mind. Second, implementing additional measures such as explanations to increase users' trust shifts the responsibility of being trustworthy from the AI system and its developers to the users, who must determine whether the AI system is trustworthy. Third, previous

1. See, for example, <https://www.nts.gov/news/press-releases/Pages/NR20200225.aspx> (accessed February 5, 2024).

research has also shown that some users do not want to know how systems, in particular AI systems, work. They would rather stay willfully ignorant because they fear that knowing how a system operates might stop them from using it (Ngo & Krämer, 2022a).

To conclude, while understanding- and transparency-enhancing approaches aiming to increase user trust indeed hold benefits, they also come with many downsides. In the next section, we suggest a different approach to user trust: epistemic trust through AI seals of trust.

Epistemic Trust in AI and Trust in AI-as-an-Institution

One of the main assumptions of understanding- and transparency-enhancing approaches to increase trust in AI, such as explanations or cues, is that users carefully assess the trustworthiness of AI to know whether they can trust it or not. Implicitly, this assumption often entails that users make rational choices about a system, that is, choices based on accurate perception and inference. However, as shown in the previous section, this assumption does not always hold.

We suggest that an alternative to such understanding-based trust is *epistemic trust*. Individuals show epistemic trust (see also, *trust in testimony*, Coady, 1992), whenever they accept communication or communicated knowledge from others as trustworthy, generalizable, and relevant (Sperber et al., 2010). In other words, when individuals trust what others tell them, they show epistemic trust. One could quickly assume that, as such, epistemic trust is equal to blind trust. However, individuals only assume information to be truthful and relevant when contextual or content cues like source credibility or plausibility evaluations do not indicate otherwise (Gilbert et al., 1993).

In the context of AI systems, showing epistemic trust in the communication of especially experts can ease their trust assessments, as it is easier for them to ask “Whom to believe?” instead of attempting to understand the AI system. Examining epistemic trust in science communication, Bromme and Gierth (2021) argue that, while from a classical logical perspective, to judge the trustworthiness of someone (or something) based on their expertise would be called an *argumentum ad verecundiam* (an argument from authority), a fallacious inference, it is indeed more accessible for individuals to assess the expertise of the scientists than to assess the veracity and scrutiny of the scholarship itself. Hence, establishing epistemic trust in AI systems could help overcome the burden of understanding the system.

Arguments similar to epistemic trust in AI systems also come from within the human-AI interaction community. Knowles and Richards (2021) established the concept of *public trust* in AI. In doing so, they differentiate between trust in a specific, discrete, and identifiable AI from trust in AI as an abstraction, which they call trust in *AI-as-an-institution*. Central, here, is the argument that “individuals do not develop trust in [AI] systems through careful and ongoing assessment of their trustworthiness; instead, one trusts that the system itself has appropriate mechanisms for ensuring trustworthiness” (Knowles & Richards, 2021, p. 264). Knowles and Richards also make clear that the ensuring instances are not the developers of the AI systems but the broader ecosystem that determines the trustworthiness rules

developers must follow. In other words, Knowles and Richards suggest that users develop epistemic trust in the ecosystem to ensure the trustworthiness of AI systems.

In their model of public trust, Knowles and Richards (2021) also suggest a four-step process to reach public trust in AI, starting with (1) defining trustworthiness, followed by (2) specifying trustworthiness, (3) enforcing trustworthiness, and (4) reaching trustworthy AI. In their model, the matter of trust calibration is taken over by the ecosystem, ensuring that AI development and outcomes are inherently trustworthy. However, how would an ecosystem communicate the trustworthiness of AI? One answer, included by Knowles and Richards in the fourth step of their model, is by providing certifications which we discuss in the next section.

AI Seals of Trust: Theoretical and Empirical Considerations

Certifications such as AI seals of trust generally “refer to a process in which a company’s processes and services [here: AI] are evaluated against a predefined set of criteria via an audit by a third party, which formally acknowledges that the standard defined by the criteria is met” (Lansing et al., 2019, p. 4). As such, certifications aim to reduce complexity and uncertainties about systems and make it easy for users to identify what is (not) trustworthy. To that end, certifications have been discussed and introduced in various contexts, such as cybersecurity, web assurances in e-commerce, or cloud services. For the context of AI, the EU AI Act suggests certification as a central mechanism to communicate compliance with industry and legislative requirements to the public (see Article 44 in Chapter 5 “Standards, Conformity Assessment, Certificates, Registration”²).

To introduce seals of trust to the field, it is crucial to consider the effectiveness of such measures. Theoretically, arguments supporting seals of trust have previously predominantly been grounded in (1) trust theory, (2) signaling theory, and (3) persuasion literature, in particular, the elaboration likelihood model (ELM).

From the perspective of trust theory, seals of trust communicate to users through trust-assuring arguments that a system can fulfill the specific requirements laid out in the contract between trustor and trustee. In doing so, in trust theory, seals of trust become part of an institutionalized mechanism that ensures trust. In signaling theory, the main focus is on the communication process of one party to the other. Central here is the assumption of an *information asymmetry* wherein one party is less informed (the trustor) than the other (the trustee). Providing information in the form of seals of trust “are signals which are actions that parties take to reveal their true type” (Kirmani & Rao, 2000, p. 66).

In contrast to trust theory and signaling theory, the ELM is more explicit in how seals are perceived. At its core, the ELM describes how individuals process persuasive arguments by following either a peripheral route of processing which requires less cognitive effort, or a central, more effortful route of information processing. Theoretically, seals of trust function as cues that can effortlessly be processed via the peripheral route. However, processing via the central route is also possible when seals of trust induce deeper elaboration (Lowry et al., 2012).

2. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52021PC0206>

While all three theoretical approaches assume positive effects of seals of trust, empirically, previous scholarship has been inconclusive. On the one hand, some authors have found no effects. For example, McKnight et al. (2004) found no effects of, what they called, privacy assurance and industry endorsement seals on trust in web business. The authors explain their results, suggesting that participants either did not notice the seal or did not know what it was supposed to signal. Similar results were obtained by Kim et al. (2008), who found no effect of seals on trust but also pointed to a lack of understanding and familiarity with the seal's meaning. On the other hand, in a more recent study, Kim et al. (2016) found that Web Assurance Seal Services (WASS) were effective instruments to increase users' trust and mitigate their concerns about e-commerce platforms. Moreover, results for the positive effects of seals on trust in the context of e-commerce are supported by findings from Mavlanova et al. (2016). In doing so, the authors differentiated between internal (company's certification) and external (third-party certifications) signals. Their results indicate that, although both signals increased trust, only external signals also increased the perceived quality of the seller. Joining results against and in favor of seals of trust, Adam et al. (2020) introduce the trust tipping point. Examining the effectiveness of seals of trust in the context of online websites, the authors found that below a certain trustworthiness threshold, seals effectively increased users' trust. However, with raising trustworthiness, the seals could not increase users' trust further.

Concluding from previous empirical findings, we know that seals of trust can effectively increase trust. However, the effectiveness might be reduced when (a) users do not notice the seals of trust, (b) users do not know the function of the seal of trust, (c) the seal of trust is granted internally, and (d) user trust is already at a high level.

The Present Study

Based on the theoretical and empirical findings elaborated above, for this study, we assume that:

H1: An AI system with an AI seal of trust is perceived as more trustworthy than an AI system without an AI seal of trust.

Moreover, we are also interested in how a seal of trust would affect each trust dimension (performance, process, and purpose). However, empirical differentiations between the three trust dimensions are rare. Hence, we did not formulate a directional hypothesis but instead posed the following research question:

RQ: How does a seal affect the three trust dimensions (performance, process, and purpose)?

Going beyond the mere presence (or absence) of a seal, we are also interested in the specific content of such a seal. What exactly should be certified? As it stands, trustworthy AI can refer to various aspects. While we hypothesize that any seal of trust would help to increase the users' trust perceptions (see H1), we also assume differences between different

seals (H2), relating to how familiar users are with the seals' content (H3a) and how well users understand what the seal certifies (H3b). More formally stated, we hypothesize:

H2: The three trust seals differ in their perceived trustworthiness, with certification of training data receiving the highest trust, followed by certification of transparency and certification through formal verification.

H3a: The seals' perceived trustworthiness partly depends on the perceived familiarity with the seals' content. The more familiar users are with the content of the seal, the higher the perceived trustworthiness of the seal.

H3b: The seals' perceived trustworthiness partly depends on the perceived understanding of users of the seals' content. The more intelligible seals are for users, the higher the perceived trustworthiness of the seal.

In addition, as the literature reviewed above suggests, trust in the certifying body will also affect how a seal is perceived. Hence, we assume:

H4: The seals' perceived trustworthiness partly depends on the perceived trustworthiness of the certifying body. The higher the perceived trustworthiness of the certifying body, the higher the perceived trustworthiness of the seal.

Because the literature on the possible effects of AI seals of trust is scarce, we also included a more explorative approach to better understand users' needs and expectations. Hence, in addition to the directional hypotheses, we included a qualitative part in which we asked participants to elaborate on which aspects of AI systems should be certified through an AI seal of trust.

Method

The study received ethical approval from the ethics committee of the University of Duisburg-Essen. All hypotheses and analyses were pre-registered via [OSF—Open Science Framework](#).

Sample and Study Design

To test our hypotheses and research question, we conducted an online study with a between-group design. To that end, we collected data from $N = 453$ participants who were randomly assigned to one of four conditions. The sample consisted of 220 females, 218 males, 12 nonbinary, and three participants who preferred not to disclose their gender identity. All participants were recruited via the crowd-sourcing platform Prolific. Participants' mean age was 37.94 ($SD = 12.69$) and ranged from 18 to 80 years. The highest degree for two participants was a middle school degree, for 184 a high school degree, for 194 a Bachelor's degree, for 48 a Master's degree, for four a PhD, and 21 indicated to have received another degree.

Manipulated Variable: The AI Seal of Trust

The four experimental conditions reflected the different trust seals, in addition to a control group. To that end, we selected three certifications which correspond to archetypical levels of insight into the inner workings of AI systems: (1) The quality of the training data ($n = 114$)—that is, even if the AI system is a black box, certifications based on the input (i.e., training data) may assist in assessing the system's trustworthiness, (2) the transparency (e.g., explainability) of the AI system ($n = 114$)—as it relates the input and output of a black box approximate system behavior, and (3) the formal verification of a AI system ($n = 113$)—as it guarantees desirable behavior of the system by white-boxing it. In addition to these different certifications, we included one control group ($n = 113$), which did not receive any seal of trust.

In addition to a brief description about the respective trust seal (all detailed descriptions can be found in the online supplementary material C), participants saw an image of a seal (see Figure 1). Because the design of a seal likely affects the end-users' trustworthiness perceptions, we reduced this effect by adding the following statement to the visual representation of the seal: "Please be aware that due to copyright reasons, we cannot represent the actual seal. The representation you see here is just a placeholder for this study."

FIGURE 1 Visualization of the AI Trust Seal That Participants Saw in the Study



Procedure

After agreeing to the informed consent, participants were introduced to a working definition of AI (see the online supplementary material A for details). We included this information to ensure that all participants understood the terminology similarly. Afterward, participants of the experimental groups were introduced to the concept of AI seals of trust with the following text:

“Artificial intelligence (AI) is recognized as a strategically important technology that can contribute to a wide array of societal and economic benefits. However, it is also a technology that may present serious risks, challenges, and unintended consequences. Within this context, trust in AI systems is necessary for the broader use of these technologies in society. It is, therefore vital that AI-enabled products and services are developed and implemented responsibly, safely, and ethically. But how to know whether one can trust AI? One way to make this trust judgment easier for users are so-called AI seals of trust. Such AI seals of trust

are granted by independent and neutral intermediaries who assess whether AI fulfills trustworthiness standards. Similar to food certifications and labels, these AI seals signal to users the state of an AI.”

Next, participants saw the different seals of trust and were introduced to different AI systems certified with AI seals of trust. Participants of the control group were directly introduced to the AI system and did not view information on the seals of trust. After viewing the AI systems, participants were asked to answer several questions about one of these AI systems. Before closing the study with a manipulation check and the debriefing, participants were informed about all three possible seals of trust, after which, in an open question, participants were asked to indicate which of the three seals they found most important (ranking question), and what they expect from an AI seal of trust.

Stimulus Material

Participants read short descriptions of four different AI systems and their functionalities. While modeled after real-world applications to avoid prior exposure effects, all systems were hypothetical and did not exist. The systems were: (1) CheckMySkin, a mobile application to check for skin cancer, (2) Drive Tek, an autonomous driving system, (3) Sound Shuffle, a music recommendation system, and (4) FindYou, a hiring system. The texts participants read can be found in the online supplementary material B.

To increase the generalizability of our results, half of the participants answered questions about the system CheckMySkin, whereas the other half answered questions about the system Drive Tek. Participants in the experimental groups saw both of these systems alongside an AI seal of trust. For the analysis, both conditions were joined.

Moreover, to increase external validity, we added two additional systems, Sound Shuffle and FindYou, which were always presented without an accompanying seal of trust. Hence, all participants of the experimental groups saw two systems with and two systems without seals of trust, whereas participants of the control group only saw systems without seals of trust.

Measured Variables

All of the following measures were assessed on a 5-point Likert scale, ranging from 1 = “strongly disagree” to 5 = “strongly agree.” For subsequent analyses, items of all measures were summarized to a final mean score.

Trust in a system. Because we wanted to assess trust as thoroughly as possible, we combined items from different scales to measure the three dimensions of trust (performance, process, and purpose) and mistrust. The final measure included 15 items to measure the perceived performance of a system (Cronbach’s $\alpha = .96$), 13 items to measure the perceived process (Cronbach’s $\alpha = .90$), 10 items to measure the purpose of the system (Cronbach’s $\alpha = .87$), and 12 items to measure mistrust (Cronbach’s $\alpha = .94$). All items used to measure the trust dimensions and a supporting exploratory factor analysis can be found in the online supplementary material F.

Perceived familiarity and perceived understanding. We used a three-item measure, adapted from Gefen (2000), to assess the participants' perceived familiarity with a seal's content. The items were "I am familiar with the concept of [. . .]," "I have heard about the possibility to make AI systems better by controlling [. . .]," and "Media often report about controlling [. . .]." Depending on the group participants were allocated to, the blanks were filled by "the training data," "the concept of transparency," or "the concept of formal verification." For the analyses, all items were summarized in one mean score with Cronbach's $\alpha = .91$.

The construct perceived understanding was assessed through the following four items, which were developed following Ngo and Krämer (2022b): "I understand what the seal of trust means," "It is clear to me what the seal certifies," "I could explain in my own words what the certification does," and "I am uncertain about the meaning of the seal." For the analyses, all items were summarized in one mean score with Cronbach's $\alpha = .88$. Both constructs, perceived familiarity and perceived understanding were not assessed by participants of the control group who did not view a seal of trust.

Trust in the certifying body. Trust in the certifying body was assessed through seven items from corporate credibility scale of Newell and Goldsmith (2001). For the analyses, all items were summarized in one mean score with Cronbach's $\alpha = .95$.

Trust in artificial intelligence. Because we did not want the individual's take on AI to interfere with our results, we also included individuals' attitudes toward AI as a covariate, using the ATAI scale of Sindermann et al. (2021), which includes five items on an 11-point Likert scale such as "I fear artificial intelligence" or "Artificial intelligence will benefit humankind." For the analyses, all items were summarized in one mean score with Cronbach's $\alpha = .78$.

Qualitative Content Analysis

To better understand the participants' needs and expectations toward an AI seal of trust, we included a ranking question and an open-ended question at the end of our online experiment. In the ranking question, having been introduced to all three possible seals of trust, we wanted to know which of the seals of trust participants found most important. To conclude, we asked:

"Lastly, having seen now three possible AI seals of trust, we are curious whether you have your own opinion about what an AI seal of trust could certify. Below you have some space to let us know what you think would be important."

We analyzed all answers following Mayring's (2014) recommendations for qualitative content analysis (see results section for details).

Results

All data can be accessed via [OSF—Open Science Framework](#).

Manipulation Check

A chi-squared test with the independent grouping variable trust seal and the dependent variable trust seal recall indicated that significantly more participants remembered correctly the seal they saw than those who did not remember correctly ($\chi^2(12) = 747.05, p < .001$). In the control condition, 63.4% of participants remembered correctly ($n = 71$), in the training data condition, 67.5% ($n = 77$), in the transparency condition, 63.15% ($n = 72$), and in the formal verification, 79.6% ($n = 90$).

Hypotheses Testing

In the central hypothesis of this work (H1), we expected that participants trust an AI system certified with an AI seal of trust more than an AI system without certification. To determine the effect of a seal on the participants' trust, we conducted an ANCOVA with the trust score as the dependent variable and the four leveled factor *AI seal of trust* as the grouping variable. As the covariate, we controlled for participants' general trust in AI. The descriptive results of the variables trust and its subdimensions performance, process, and purpose, as well as mistrust grouped by the factor *AI seal*, can be found in Table 1.

TABLE 1 Descriptive Results of the Dependent Variable Trust and Its Subdimensions by Experimental Group

		No Seal	Training Data	Transparency	Formal Proof
Trust	<i>M</i>	3.48	3.53	3.50	3.41
	<i>SD</i>	0.71	0.68	0.76	0.69
Performance	<i>M</i>	3.29	3.49	3.39	3.36
	<i>SD</i>	0.87	0.78	0.93	0.88
Process	<i>M</i>	3.22	3.24	3.22	3.08
	<i>SD</i>	0.85	0.82	0.94	0.87
Purpose	<i>M</i>	3.93	3.87	3.88	3.79
	<i>SD</i>	0.79	0.77	0.76	0.76
Mistrust	<i>M</i>	3.34	3.24	3.35	3.44
	<i>SD</i>	1.05	0.99	1.05	0.94

Results of the ANCOVA indicate that there was no significant difference in the participants' trust scores between the different groups, $F(3,448) = 0.72, p = .54$. Moreover, we also had to reject H2 for which we expected that the training data seal would receive the most trust, followed by the transparency seal, and the formal verification seal.

While the result for H1 indicates that none of the three different seals of trust affected participants' trust perceptions, it could have been the case that the seal affected only subdimensions of trust. For this possibility, we did not articulate a hypothesis but posed RQ1, asking whether the different seals affected the three subdimensions, performance, process,

and purpose differently. In addition to the three subdimensions, we also included the measure for mistrust in RQ1 (note that mistrust was not included in the RQ in the pre-registration). To assess RQ1, we conducted a MANCOVA with the subdimensions performance, process (integrity & transparency), and purpose, as well as mistrust as outcome variables and the four leveled factor AI seal of trust as the grouping variable. Similar to testing H1, we also controlled for individual levels of trust in AI. Results indicate that the three subdimensions, as well as mistrust, were similarly affected by the trust seals, Pillai's trace = .02, $F(3,448) = 1.09$, $p = .075$.

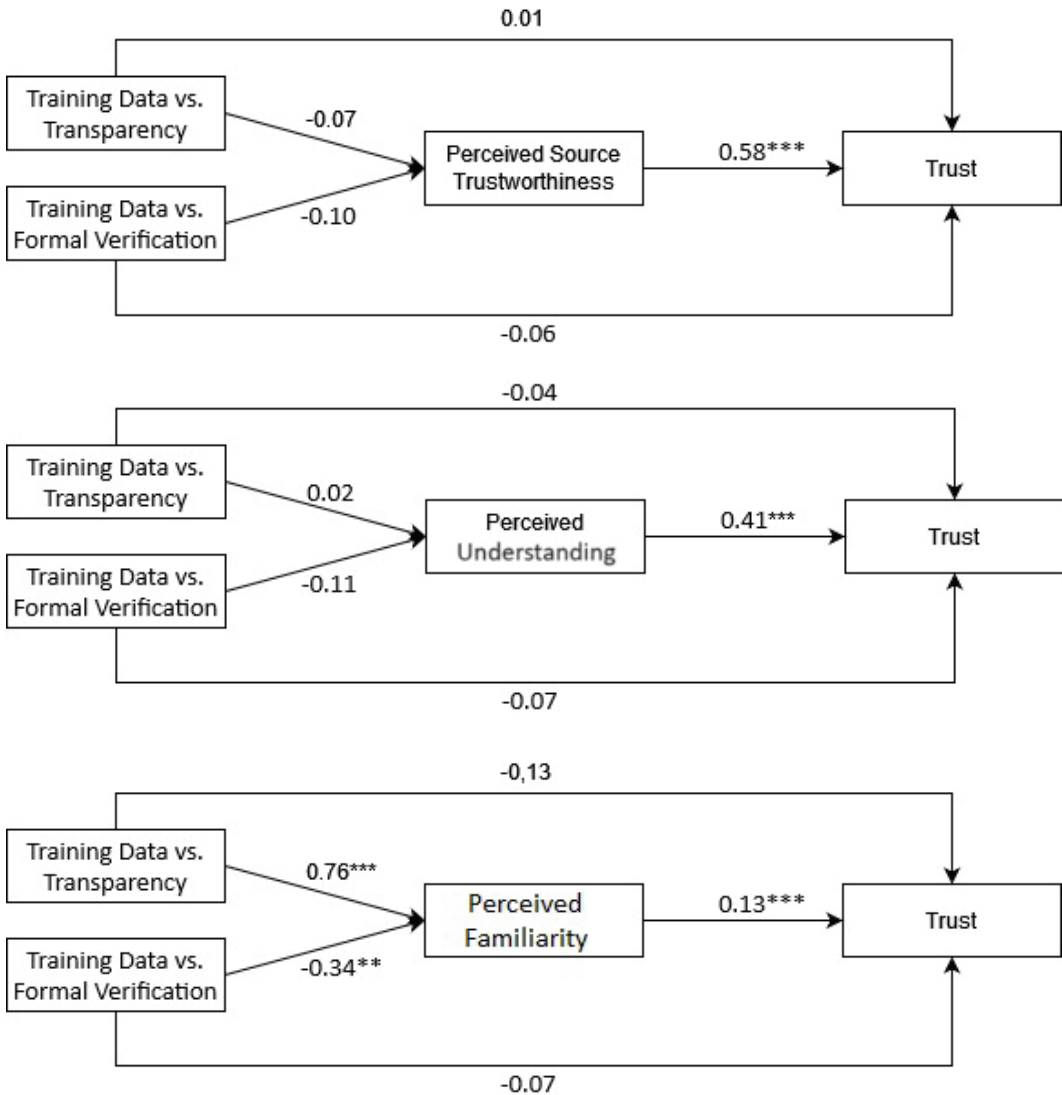
Although we found no differences between the three seals of trust and participants' trust perceptions (see H1), an indirect effect of the seals on trust can still be expected. In H3a and H3b, we suggested that an effect of the seal is at least partly the result of the participants' perceived understanding of the seal's content and the participants' familiarity with the seal's content. In addition, in H4, we anticipate that the effect of the seals might also be the result of the perceived trustworthiness of the institution which issued the seal.

To understand these possible explaining mechanisms, we ran three separate mediation analyses with understanding, perceived familiarity, and perceived source trustworthiness as mediating variables. For this, we used the Process Macro version 4.3.1 for SPSS by Hayes (2017). Furthermore, we used the variable *AI seal of trust* as the independent variable, which was dummy-coded. Participants who viewed the training data seal were entered as a reference category. Participants of the control group were excluded from the analyses as they did not answer questions about their understanding of the seal, their perceived familiarity, and the perceived trustworthiness of the source (see also the elaboration in the methods section). The outcome variable was again trust. We tested the significance of the effects using bootstrapping procedures, computing 5,000 bootstrapped samples with a confidence interval of 95%. All unstandardized path coefficients and significance levels can be found in Figure 2a–c. The full results of the mediation analyses can be found in the online supplementary material D.

The mediation analyses revealed nonsignificant indirect effects for all three variables (understanding, source trustworthiness, and perceived familiarity). For understanding and source trustworthiness, the a-path was insignificant, indicating that the AI seal of trust participants viewed was neither related to the variable understanding nor source trustworthiness. However, the b-path was significant, indicating that both were very strong predictors of trust, with understanding explaining roughly 34% of the trust variance and source trustworthiness explaining roughly 72%. Not surprisingly, these results underline the importance of users understanding what a seal represents and the importance of the issuing source of the seal.

In contrast, we found a significant a-path for perceived familiarity, suggesting that participants were not equally familiar with all AI seals. In particular, we found that participants were more familiar with transparency than verified training data (positive coefficient) but were less familiar with formal verification than training data (negative coefficient). This result partly confirms what we anticipated in H2, suggesting that participants are not equally familiar with the different seal content. Beyond this, the significant b-path indicates that higher familiarity with a seal's content resulted in greater trust.

FIGURES 2a–2c Visual Representation of Mediation Analyses With Unstandardized Path Coefficients

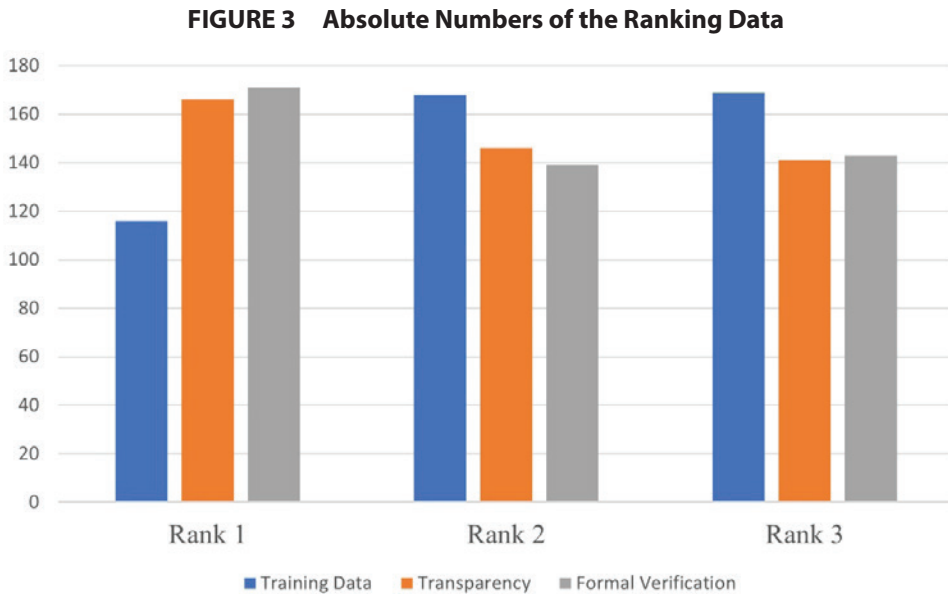


* $p < .05$, ** $p < .01$, *** $p < .001$

Qualitative Results

First, we asked participants to rank the three seals of trust they found most important. With one as the highest rank and three as the lowest, mean results indicate that participants found all three seals of trust similarly important, with formal verification scoring $M = 1.94$, transparency of the AI system $M = 1.94$, and training data $M = 2.12$. While the mean ranks

do not indicate a great difference between the three seals of trust, which reflects the results of our quantitative analysis, inspecting the absolute number that a seal was ranked first, we can see that participants found the formal verification and transparency of a system most important (see Figure 3).



Because the three trust seals we selected reflect our understanding of importance, we assessed the participants' answers with an open-ended question, asking what participants find most important in an AI seal of trust. We applied descriptive and in-vivo codes in the first coding cycle to capture the participants' answers (Saldaña, 2013). In the second step, all codes were abstracted and summarized into higher-level codes. Throughout both coding cycles, three independent coders worked on the answers. To ensure the quality of the final coding scheme, we calculated Cohen's Kappa on 25% of the answers. In the first round, all three coders arrived at an agreement of $K = .64$. To increase agreement, all three coders discussed and resolved cases of disagreement. Consequently, inter-rater reliability increased to a sufficient $K = .82$ in a second round of coding on different sets of answers.

In the following, we report the most important results of the qualitative content analysis. Overall, the final coding scheme identified seven different categories (see Table 2), which differ in the number of mentions as well as the level of abstraction (number of second-level codes).

TABLE 2 Results of the Qualitative Content Analysis

1st level codes	#	Description	2nd level codes (#)
Trustworthiness	19	Verification of the general trustworthiness of a system without further specification	
Distrust	24	General distrust in the AI or the seals of trust	
Performance	206	Verification of AI's abilities and characteristics	(formal) Verification (60) Safety (55) Accuracy (23) Error-free (20) Re-evaluation (20) Extensive testing (20) Efficiency (8)
Process (transparency)	49	Related to how the AI operates and the intelligibility of its inner workings	
Purpose	66	Verification of the intentions of the AI's developers and the development process	Ethical compliance (13) Privacy (24) Training set quality (25) Copyright compliance (4)
Trustworthy AI seals of trust	36	Verification of the seal-issuing institution	Trustworthy origin of the seal (26) Transparency of the certification process (10)
Destructive AI	20	Verification that AI cannot develop its own agency and intentionally harm humans	

While some participants voiced general support for trust seals, others rejected any certification as well as AI systems as a whole. For example, P84 stated, “nothing would really give me any trust in AI. I am very against the idea of anything AI.” In addition to general mistrust in AI and certifications, participants also voiced concrete concerns about the seal-issuing institution. For example, P163 states, “I don’t necessarily trust these seals of trust because they can always get bought.” This reflects our quantitative results, which underline the importance of source trustworthiness. Moreover, the distrust voiced by our participants reminds of what Dietvorst et al. (2015) call algorithm aversion, a generally negative stance toward anything related to algorithms and AI.

Following trust literature, most participants, however, commented along the lines of the three trust dimensions performance, purpose, and process, with performance-related comments being mentioned by far the most. Among those, most participants wanted a seal of trust to certify that the system does what it was set out to be (formal verification) and its safety.

Related to the issue of safety, a smaller group of participants also voiced the need for, what we call, a nondestructive AI seal of trust. For example, P136 stated that a seal “could certify that the AI can be trusted not to be evil and ruin mankind,” and P389 who noted that a seal “could certify whether the AI’s intentions are true—whether it wants to make humans safe or whether it wants to further its own goals regardless of our safety.”

Discussion

Through quantitative and qualitative data collections, in this work, we investigated the effect of an AI seal of trust on the users’ trust assessments of AI systems as well as the users’ expectations toward such seals respectively.

Quantitative Results: Addressing the Null Effect of the Trust Seal

In a pre-registered online experiment, we tested three different seals of trust (certification of the training data, transparency, and formal verification) and their effects on user trust in an AI system. However, unlike hypothesized, none of the three different seals of trust could significantly increase our participants’ trust in an AI system compared to a control group. A more fine-grained analysis, differentiating trust into its subdimensions performance, process, and purpose, supported this null result. The seals of trust did not affect the trust dimensions differently compared to a control group.

While previous results from different domains would suggest an effect of the certification, this paper’s null results echo previous null results. Examples include McKnight et al. (2004) and Kim et al. (2016), who relate their null findings to users’ not noticing the seal or users’ limited understanding and familiarity of the seal’s content. We can rule out these explanations because we also assessed participants’ understanding of and familiarity with a seal. In addition, the manipulation check indicated that participants remembered the respective trust seals. Instead, we suggest that our results relate to the findings of Adam et al. (2020). The authors suggest that if a system’s trustworthiness is already high, an additional seal of trust cannot increase the trustworthiness any further. We find support for this speculation in the mean trust ratings of our study as we noticed that these fall within 3.41 and 3.53 points, significantly higher than the scale midpoint (2.5 points).

Following theoretical considerations of trust theory and signaling theory, an alternative explanation to the null results is that the trust seals did not signal the intended meaning. Indeed, our seals might not have communicated the trustworthiness of the systems because they are neither well established outside the experimental setting nor granted by a well-known institution (see also next section). Hence, they possibly lacked the epistemic authority to convince our participants.

Moreover, we found that the seals of trust were not perceived differently in terms of understandability but differed in familiarity, with transparency certification being the most well-known, followed by training data and formal verification certifications. Finding differences for familiarity but not understanding indicates that, while knowing of a specific certification method, this knowledge does not necessarily translate into understanding.

Support for Epistemic Trust

We found that independent of which seal participants saw, the higher the participants' trust in the seal-issuing institution was, the higher was the trust in the AI system. In other words, if users trust the institution that grants the seal, trust in the system will increase. Consequently, this shifts the users' trust assessments from the system to the certifying institution. Hence, our result supports the idea of epistemic trust and trust in AI-as-an-institution (Knowles & Richards, 2021). It seems that it is easier for users to ask, "Whom to trust?" instead of attempting to understand AI systems.

Moreover, in line with predictions of the ELM, knowing a certifying institution might also function as a mental shortcut. Knowing that a certain institution is trustworthy, any communication originating from such an institution should also be trustworthy (see also, *authority heuristic* in Sundar, 2008). For the present work, we could not rely on the authority of a specific institution as our seals might have been less effective because their origin was unknown to the participants. However, adding additional information such as a seal or a seal-issuing institution whose trustworthiness has to be assessed also comes with downsides discussed in the next section.

Qualitative Results

Need for Verifications Without Understanding of Verifications

In the qualitative part of this work, we asked participants to explain what they expect from AI certifications. Through a qualitative content analysis, we found that participant responses mainly fell within the three trust dimensions, performance, process, and purpose, with performance-related certification being mentioned the most. Among the performance category, participants indicated that (formal) verification, the certification that the system does what it was set out to be, was mentioned the most. This is also supported by the ranking data that we collected. Here, formal verification was ranked first most of the time. However, in light of the quantitative results, which indicated that participants knew the least about formal verification compared to transparency and training data, the higher ranking of formal verifications is alarming. Participants found the greatest reassurance in something they understood the least and, in turn, maybe expected it to be most comprehensive and fail-safe. We speculate whether this might be due to participants having given up on other, more well-known methods.

Second-Level Trust Calibrations

Interestingly, some participants mentioned the general need for a trustworthiness certification, whereas others voiced distrust toward any such certification and AI-related system. We relate these contradicting sentiments to what Wischniewski et al. (2023) define as *second-level trust calibrations*, where users have to perform an additional (second level) trust judgment (here: judging the trustworthiness of the seal) on top of the trust judgment concerning the AI system (first level), possibly increasing users' cognitive load. While following persuasion literature which suggests that seals can reduce the users' cognitive load

by offering trust cues, future studies should examine whether cognitive load can also be increased through the additional information that needs to be processed. This is especially true in the context of calibrated trust. Suppose it is the aim that user trust is appropriately calibrated to the AI system's functionality. In that case, users must also find a way to calibrate their trust in the AI seal appropriately.

In addition, the distrust sentiment voiced by our participants also indicates the limits of approaching trust from an epistemic perspective. If the seal-issuing institution is not trusted, users will likely not trust the system. Hence, future studies should assess which cues make an AI seal of trust more trustworthy and which user groups generally distrust AI.

Limitations and Future Studies

The strongest limitation to our study concerns its external validity. First, as currently no established, noncommercial certification body or trust seal exist, all material was hypothetical. Similarly, participants did not directly engage with the AI systems but read different vignettes. Hence, we could not measure how participant trust translated into actual behavioral outcomes. Further, online data collection is limited for decisions in practice as this problem type involves substantial cognitive effort that an online environment may not be able to replicate as well as decision-making often is a high-involvement task and online participants may not meet this criterion.

For future studies, we suggest integrating actual systems into the experimental setting. In addition, with AI systems based on large language models such as GPT-4 being commercialized, it could be interesting, for example, to include such a conversational interface and interactivity in general.

Moreover, as participants likely did not know about AI seals of trust, we had to provide a definition of such. While we tried to be as subtle as possible, describing AI systems as “a technology that may present serious risks, challenges, and unintended consequences” (see Method section), we potentially biased participants to be more critical and vigilant than they initially were, raising participants' overall skepticism toward the presented system. However, as we can see in the overall trust ratings across conditions, participants perceived the systems as relatively trustworthy (mean trust ratings > 3.41 points at a scale midpoint of 2.5 points). In addition, we statistically controlled for participants' general attitudes toward AI by including individuals' attitudes as a covariate in our analyses. Hence, even if a subgroup of users was affected by our definition, it should not have changed our results.

Lastly, as we suggest in the previous section, we speculate that our null results are related to all AI systems being equally trustworthy. To test this interpretation, future studies should experimentally vary the trustworthiness of AI systems by, for example, comparing different levels of system reliability (high vs. low) to investigate whether trust seals can increase the users' trust.

Conclusion

In this work, we investigated the effects of AI certifications, so-called AI seals of trust, on the users' trust in AI systems. We tested three certifications and their effects on global trust

and the trust subdimensions performance, process, and purpose. Unlike hypothesized, we found that the trust seals did not affect users' trust in the AI system. Examining possible underlying mechanisms, we found that a higher understanding of the seal's content as well as familiarity with the seal's content, could increase users' trust. Moreover, we found evidence of epistemic trust. That is, the more participants trusted the seal-issuing institution, the more they trusted the AI system. However, our qualitative results also indicated that some participants reject the idea of an AI seal of trust as they do not trust AI systems or any certifying party. Nevertheless, most participants said they would like to see a system's functionality be certified, specifically, its performance and safety.

Author Biographies

Magdalena Wischnewski (PhD, University of Duisburg-Essen) is a PostDoc at the Research Center for Trustworthy Data Science and Security in Dortmund, Germany. In her work, she investigates human-centric trustworthy AI, such as the calibration of trust, trust assessment, and auditing of AI through AI seals of trust.

 <https://orcid.org/0000-0001-6377-0940>

Nicole Krämer (PhD, University of Cologne) is the Professor for Social Psychology: Media and Communication at the University of Duisburg-Essen.

 <https://orcid.org/0000-0001-7535-870X>

Christian Janiesch (PhD, University of Münster) is the Professor of Enterprise Computing at TU Dortmund University. His research focuses on intelligent systems at the intersection of business process management and artificial intelligence.

 <https://orcid.org/0000-0002-8050-123X>

Emmanuel Müller (PhD, Technical University of Aachen) is the Professor for Computer Science at the Chair of Data Science and Data Engineering at the Technical University Dortmund.

 <https://orcid.org/0000-0002-5409-6875>

Theodor Schnitzler (PhD, Ruhr University Bochum) is an Assistant Professor at the Department of Advanced Computing Science at Maastrich University. His main area of research is user privacy in online environments from both technical and HCI perspectives.

 <https://orcid.org/0000-0001-7575-1229>

Carina Newen is a PhD student in Computer Science at the Research Center Trustworthy Data Science and Security in Dortmund, Germany. She works in interdisciplinary fields such as trustworthy data science from a computer science and psychological perspective

 <https://orcid.org/0000-0001-8721-6856>

Center for Open Science



This article has earned the Center for Open Science badges for Open Data and Open Materials through Open Practices Disclosure. The authors have made their data and materials freely accessible at <https://osf.io/6znvr/>. The article also earned a Preregistered badge for having a preregistered design available at <https://osf.io/c3g6y>.

References

- Adam, M., Niehage, L., Lins, S., Benlian, A., & Sunyaev, A. (2020). Stumbling over the trust tipping point—The effectiveness of web seals at different levels of website trustworthiness. In *Proceedings of the 28th European Conference on Information Systems (ECIS). Online Conference, June 15–17, 2020*. https://aisel.aisnet.org/ecis2020_rp/3
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., 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>
- Bertrand, A., Belloum, R., Eagan, J. R., & Maxwell, W. (2022). How cognitive biases affect XAI-assisted decision-making: A systematic review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 78–91). <https://doi.org/10.1145/3514094.3534164>
- Bromme, R., & Gierth, L. (2021). Rationality and the public understanding of science. In M. Knauff & W. Spohn (Eds.), *The Handbook of Rationality* (pp. 767–776). MIT Press. <https://doi.org/10.7551/mitpress/11252.003.0084>
- Coady, C. A. J. (1992). *Testimony: A philosophical study*. Clarendon Press.
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114. <https://doi.org/10.1037/xge0000033>
- Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering*, 66, 111–126. <https://doi.org/10.1007/s12599-023-00834-7>
- Gefen, D. (2000). E-commerce: The role of familiarity and trust. *Omega*, 28(6), 725–737. [https://doi.org/10.1016/S0305-0483\(00\)00021-9](https://doi.org/10.1016/S0305-0483(00)00021-9)
- Gilbert, D. T., Tafarodi, R. W., & Malone, P. S. (1993). You can't not believe everything you read. *Journal of Personality and Social Psychology*, 65(2), 221. <https://doi.org/10.1037/0022-3514.65.2.221>
- Grigorescu, S., Trasnea, B., Cocias, T., & Macesanu, G. (2020). A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3), 362–386. <https://doi.org/10.1002/rob.21918>

- Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Press.
- Kim, D. J., Ferrin, D. L., & Rao, H. R. (2008). A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents. *Decision Support Systems*, 44(2), 544–564. <https://doi.org/10.1016/j.dss.2007.07.001>
- Kim, D. J., Yim, M.-S., Sugumaran, V., & Rao, H. R. (2016). Web assurance seal services, trust, and consumers' concerns: An investigation of e-commerce transaction intentions across two nations. *European Journal of Information Systems*, 25, 252–273. <https://doi.org/10.1057/ejis.2015.16>
- Kirmani, A., & Rao, A. R. (2000). No pain, no gain: A critical review of the literature on signaling unobservable product quality. *Journal of Marketing*, 64(2), 66–79. <https://doi.org/10.1509/jmkg.64.2.66.1800>
- 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). <https://doi.org/10.1145/2858036.2858402>
- Knowles, B., & Richards, J. T. (2021). The sanction of authority: Promoting public trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 262–271). <https://doi.org/10.1145/3442188.3445890>
- Kunze, A., Summerskill, S. J., Marshall, R., & Filtness, A. J. (2019). Automation transparency: Implications of uncertainty communication for human-automation interaction and interfaces. *Ergonomics*, 62(3), 345–360. <https://doi.org/10.1080/00140139.2018.1547842>
- Lansing, J., Siegfried, N., Sunyaev, A., & Benlian, A. (2019). Strategic signaling through cloud service certifications: Comparing the relative importance of certifications' assurances to companies and consumers. *The Journal of Strategic Information Systems*, 28(4), 101579. <https://doi.org/10.1016/j.jsis.2019.101579>
- 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
- Lowry, P. B., Moody, G., Vance, A., Jensen, M., Jenkins, J., & Wells, T. (2012). Using an elaboration likelihood approach to better understand the persuasiveness of website privacy assurance cues for online consumers. *Journal of the American Society for Information Science and Technology*, 63(4), 755–776. <https://doi.org/10.1002/asi.21705>
- Madani, A., Arnaout, R., Mofrad, M., & Arnaout, R. (2018). Fast and accurate view classification of echocardiograms using deep learning. *npj Digital Medicine*, 1(1). <https://doi.org/10.1038/s41746-017-0013-1>
- Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human-human and human-automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, 8(4), 277–301. <https://doi.org/10.1080/14639220500337708>
- Mavlanova, T., Benbunan-Fich, R., & Lang, G. (2016). The role of external and internal signals in e-commerce. *Decision Support Systems*, 87, 59–68. <https://doi.org/10.1016/j.dss.2016.04.009>
- Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, 20(3), 709–734. <https://doi.org/10.5465/amr.1995.9508080335>

- Mayring, P. (2014). Qualitative content analysis: Theoretical foundation, basic procedures and software solution. Klagenfurt. <https://nbn-resolving.org/urn:nbn:de:0168-ssoar-395173>
- McKnight, D. H., Kacmar, C. J., & Choudhury, V. (2004). Shifting factors and the ineffectiveness of third party assurance seals: A two-stage model of initial trust in a web business. *Electronic markets*, 14(3), 252–266. <https://doi.org/10.1080/1019678042000245263>
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267, 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1), 81–103. <https://doi.org/10.1111/0022-4537.00153>
- Newell, S. J., & Goldsmith, R. E. (2001). The development of a scale to measure perceived corporate credibility. *Journal of Business Research*, 52(3), 235–247. [https://doi.org/10.1016/S0148-2963\(99\)00104-6](https://doi.org/10.1016/S0148-2963(99)00104-6)
- Ngo, T., & Krämer, N. (2022a). Exploring folk theories of algorithmic news curation for explainable design. *Behaviour & Information Technology*, 41(15), 3346–3359. <https://doi.org/10.1080/0144929X.2021.1987522>
- Ngo, T., & Krämer, N. (2022b). I humanize, therefore I understand? Effects of explanations and humanization of intelligent systems on perceived and objective user understanding. *psyarXiv preprint*. <https://doi.org/10.31234/osf.io/6az2h>
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230–253. <https://doi.org/10.1518/001872097778543886>
- Saldaña, J. (2013). *The coding manual for qualitative researchers* (2nd ed.). SAGE Publications Inc. <https://doi.org/10.1017/CBO9781107415324.004>
- Sindermann, C., Sha, P., Zhou, M., Wernicke, J., Schmitt, H. S., Li, M., Sariyska, R., Stavrou, M., Becker, B., & Montag, C. (2021). Assessing the attitude towards artificial intelligence: Introduction of a short measure in German, Chinese, and English language. *KI-Künstliche Intelligenz*, 35, 109–118. <https://doi.org/10.1007/s13218-020-00689-0>
- Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origg, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, 25(4), 359–393. <https://doi.org/10.1111/j.1468-0017.2010.01394.x>
- Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. In M. J. Metzger & A. J. Flanagin (Ed.), *Digital Media, Youth, and Credibility* (pp. 73–100). The MIT Press. <https://doi.org/10.1162/dmal.9780262562324.073>
- 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* (pp. 49–54). <https://doi.org/10.1145/3544548.3581197>
- Yeh, M., & Wickens, C. D. (2001). Display signaling in augmented reality: Effects of cue reliability and image realism on attention allocation and trust calibration. *Human Factors*, 43(3), 355–365. <https://doi.org/10.1518/001872001775898269>
-