The Impact of AI Trustworthiness Labels on the Perception of AI Products

Christina U. Pfeuffer
Department of Psychology - Human-Technology Interaction
Catholic University of Eichstätt-Ingolstadt
Eichstätt, Germany
christina.pfeuffer@ku.de

Abstract

Most often, potential users aren't well informed about the trustworthiness of AI products when selecting them. They may therefore show misplaced (dis-)trust towards AI products. Here, participants were presented with (hypothetical) AI products (smart fridges, voice assistants) of (hypothetical) brands that were paired with a graphical, traffic light-like label conveying low to high AI trustworthiness or they were presented with AI products without such a label (baseline). Higher trustworthiness levels as indicated by the AI trustworthiness labels increased trust, acceptance, and the intention to use AI products, as well as the monetary value participants attributed to AI products, but did not affect the evaluation of an AI product's brand. These findings suggest that labels effectively communicated differences in AI trustworthiness. Interestingly, the evaluation of unlabeled AI products corresponded to AI products labeled with an intermediate trustworthiness level, highlighting biased assessment and the need to communicate AI trustworthiness to potential users.

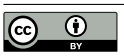
CCS Concepts

• **Human-centered computing** → Human computer interaction (HCI); Empirical studies in HCI; Human computer interaction (HCI); HCI design and evaluation methods; User studies; • **Social and professional topics** → Computing / technology policy; Government technology policy; Governmental regulations Security and privacy; Human and societal aspects of security and privacy; Privacy protections.

Keywords

artificial intelligence, AI trustworthiness, labels, regulation, perception of AI

ACM Reference Format:
Christina U. Pfeuffer. 2026. The Impact of AI Trustworthiness Labels on the Perception of AI Products. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3772318.3790776



CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3790776

1 Introduction

Artificial intelligence (AI) is permeating more and more aspects of our everyday lives promising to benefit both individuals, organizations, and society as a whole. For instance, required text, image, or video material can now easily be generated using generative AI and AI can support us in optimizing products or services, in enhancing productivity and efficiency, and in lowering the costs of tasks [19]. However, AI also bears dangers (e.g., supporting/increasing biases and discrimination, spreading of misinformation, safety risks of mistakes and failures). Its potential can therefore only be realized when human-AI interactions are suitably shaped [2, 15, 16, 18, 19, 22, 29, 32, 33]. Thus, it is important that users as well as those who might become users (potential users) are able to assess the trustworthiness of AI before and when using it (see e.g., criteria for trustworthy AI put forward by the European Commission, [11], see also [52] for theoretical considerations).This helps ensure they do not misplace their (dis-)trust in a corresponding AI product, obstructing its acceptance and use (but see e.g., [19, 27] for corresponding challenges).

The acceptance and adoption of technical systems and AI has most commonly been considered and evaluated using variants of the Technology Acceptance Model (TAM; e.g., [8, 39, 56]), the Unified Theory of Acceptance and Use of Technology (UTAUT; e.g., [56]) and their extensions. Concerning the relationship between trust, acceptance, and the adoption of technology/AI, the TAM was recently extended to intelligent systems (i.e., AI; Intelligent Systems TAM, ISTAM; [58]), integrating several models of trust in technology into the TAM (integrated model of organizational trust: [40]; trust in automation: [23, 34]). According to the ISTAM, trust is one of, if not the essential precursor of technology/AI acceptance and use (see [59] for a model similarly showing the antecedents and impact of trust in the context of social media). This implies that it is essential to ensure that the general public can trust in AI to an appropriate degree to support its further adequate acceptance and adoption in various domains of our lives.

A trustor's (e.g., a user's) trust derives from a corresponding assessment of a trustee's (e.g., an AI brand's) trustworthiness. That is, trustors determine whether a trustee is worthy of receiving trust (perceived trustworthiness) [34, 41, 52]. This assessment relies on trustworthiness cues (e.g., [52]; see e.g., [4], for a taxonomy of such cues) which can be used to determine system characteristics (e.g., information on system uncertainties). Note that a person's trust in an AI or corresponding trust and trustworthiness judgements (perceived trustworthiness) do not necessarily align with the AI's actual, objective trustworthiness considering all relevant pieces of information or with a trustworthiness assessment that would

be justified given the information available to users [11, 52]. This study focuses on perceived trustworthiness irrespective of an AI product's actual, objective trustworthiness as might be determined by experts based on its features. To separate these two aspects and focus on perceived trustworthiness, I deliberately presented only hypothetical AI products.

At present, it is very difficult, effortful, and time-consuming for potential users to gain detailed information on the trustworthiness of AI in order to develop an informed assessment (i.e., reading corresponding fineprint containing legal information and documentation, searching for missing details, etc.). It is important to note that both misplaced distrust, preventing AI acceptance and adoption [7, 58], as well as misplaced trust, leading to expectancy violations and reduced AI acceptance and adoption in the longer term [24, 49], are detrimental to a further successful integration of AI into society and our everyday lives (see also [52]). Moreover, recent studies showed that (potential) users are, at present, hardly able to assess the trustworthiness of AI based on the information available to them [19, 52]. Furthermore, they generally have relative low understanding of AI and its potential applications [19, 27]. Conversely, there is strong, often unjustified public endorsement of at least some AI products [19]. (Potential) users often inaccurately (or at least unjustifiedly based on the available information/cues [52]) presume that safeguards like those proposed in the criteria for trustworthy AI put forward by the European Commission [15] are in place for AI products (e.g., for AI used in healthcare or human resources contexts; [19]). That information on the trustworthiness of an AI is not available in an easy-to-access and easy-to-process format may strongly contribute to this. The lack of reliable, easy-to-access, and easy-to-process information on AI products might also be one of the reasons why most (potential) users judge AI based on heuristics rather than assessing detailed information on corresponding AI products [3, 36, 38]. One example for such a heuristic is that (potential) users are more trusting towards AI developed by universities or research institutions as compared to AI developed by private companies [19].

1.1 Aims of the Current Study

In the present study, I aimed to replicate and substantially extend prior work (see section 2) on the impact of AI trustworthiness labels with multiple levels indicating low to high trustworthiness.

First, whereas prior work focused on trust, I additionally determined whether the impact of AI trustworthiness labels and biases that could previously be shown by comparing evaluations for labeled and unlabeled AI products propagate to AI acceptance and adoption/intention to use. Moreover, I additionally determined whether the evaluation of corresponding (hypothetical) AI brands is also affected by the trustworthiness level indicated by an AI trustworthiness label.

Second, the study my methodology directly builds on (Pfeuffer [47, 48], see section 2) focused on a single dimension contributing to trustworthiness assessment, data security and data privacy. Here, I aimed to determine whether a global AI trustworthiness label that incorporates various dimensions of trustworthiness similarly impacts on trust (and further variables of interest). Determining this is essential to derive clear recommendations for the design of corresponding multi-level AI trustworthiness labels. Whereas a global AI trustworthiness label might be processed and interpreted the fastest, for instance, uncertainty about the contribution of each single dimension could lead to reduced effectiveness or even adverse effects. This calls for a comparative assessment of a global AI trustworthiness label.

Third, central prior work used a setting of very low ecological validity. Here, I established a context more realistically resembling real-life advertisement encounters to replicate and extend prior findings in a more ecologically valid setting.

2 Related Work

2.1 Determining Trustworthiness – Trustworthiness Cues and Labels

To counter the lack of information on an AI's trustworthiness, it has been proposed that trustworthiness cues (e.g., [36, 52], see also [10]; e.g., the transparency of certain information in the documentation) could be used to determine an AI's trustworthiness. Yet, such cues are, unfortunately, themselves not always easily accessible. Recently, studies have begun to investigate AI certification labels as a better-suited alternative to convey AI trustworthiness ([1, 14, 17, 48, 50, 51], see [20] for first label recommendations, see also [6], see also e.g., [12] for the impact of describing AI using verbal labels like "trustworthy" or "reliable" and [35] for the perception of written content labeled as created by AI versus not). Importantly, such label suggestions most often focus on certification labels which simply convey that a certain minimum threshold of trustworthiness has been passed by a corresponding product. In these cases, only the presence versus absence of such a binary certification label indicates a product's trustworthiness. However, in practice, a corresponding certification process is, at present, often optional and not obligatory. Thus, (potential) users might be uncertain about whether the absence of the certification label indicates low trustworthiness of a product or brand. It could, for instance, simply show that the brand or product has not been evaluated for the label yet. In turn, (potential) users might arrive at a too positive trustworthiness assessment, presuming that a corresponding product without a certification label has not been labeled yet instead of perceiving the product as untrustworthy. This constitutes an advantage of multi-level labels (i.e., labels conveying the degree to which respective trustworthiness criteria are met, e.g., [47, 48]) as compared to certification labels that qualify products only by their presence/absence. The existence of a corresponding multi-level label conveys that the product has been certified, whereas its level conveys a more nuanced assessment of the product's trustworthiness. Whereas certification labels are typically orientated at minimum thresholds (e.g., minimum requirements according to corresponding legislation), multi-level labels can convey both the adherence to certain minimal requirements (e.g., lowest label level) as well as the degree to which additional aspects corresponding to an ideal, desirable standard are implemented (e.g., medium to high label levels).

2.2 Trust Calibration with Multi-Level Trustworthiness Labels

A recent study by Pfeuffer [47, 48] took inspiration from the Nutri-Score label (established European food label indicating nutritional value) and assessed to what degree a similar graphical, traffic light-like label (as compared to a text-based label) can convey information on the data security and data privacy of AI products and correspondingly affect (potential) users perception and evaluation of these AI products. That is, Pfeuffer [47, 48] focused on persons who have not yet decided to buy and/or use a (hypothetical) AI product (i.e., potential users). Participants were presented with two types of AI products, smart fridges and voice assistants (each represented only by a corresponding, generic icon) that were first shown without a label (baseline phase) and then presented again with a data security and data privacy label indicating a low, intermediate, or high data security and data privacy level achieved by the hypothetical AI product. Participants indicated their trust in and AI anxiety towards each AI product. Pfeuffer [47, 48] found that trust increased and AI anxiety decreased with higher data security and data privacy levels indicated by the corresponding label. Importantly, baseline ratings corresponded to the rating of an AI product with an intermediate data security and data privacy level. This suggests the existence of a bias towards intermediate judgements when evaluating AI products without further information on them. In a second phase, participants were further presented with two data security and data privacy labels of different levels and asked how much more they would be willing to pay for the AI product of the respective higher data security and data privacy level. Again, the value participants attributed to an AI product (i.e., how much more they were willing to pay) scaled with data security and data privacy levels. Pfeuffer [47, 48] argued that these findings suggest that multi-level labels might be an appropriate means to convey aspects of AI trustworthiness to potential users. This could provide an appropriate assurance mechanism for AI trustworthiness (see also [19]) and help (potential) users calibrate their AI trustworthiness judgements. Note that the study of Pfeuffer [47, 48] showed differences between multi-level graphical and text-based labels that conveyed the same information. Text-based labels were overall associated with higher trust and lower AI anxiety, whereas participants were willing to pay more for AI products labeled with a graphical label, which was described as easier to process. Importantly, effects of data security and data privacy levels did not systematically differ between label types. Thus, it was not possible to determine one label type that would generally be more recommendable for multi-level AI trustworthiness labels.

Although the study of Pfeuffer [47, 48] provided very relevant findings regarding the impact of multi-level AI trustworthiness labels (there focusing only on the single trustworthiness dimension data security and data privacy), it nonetheless had several shortcomings. For instance, it focused primarily on trust as a precursor of acceptance and intention to use, however, did not test whether effects of AI trustworthiness labels propagated also to AI acceptance and intention to use. Moreover, the setting which presented only two broader AI product categories (smart fridges and voice assistants represented only by an icon), had low ecological validity for actual real-life advertisement contexts.

2.3 The Present Study's Objectives and Contribution to the Literature

First, in the following study building on Pfeuffer [47, 48], I provided a more realistic and ecologically valid setting for AI product evaluation showing both individual product images of (hypothetical) AI products as well as logos of (hypothetical) brands. Like Pfeuffer [47, 48], I intentionally focused on hypothetical AI products (i.e., non-existing smart fridges and voice assistants) and a sample from the general public, that is, potential users who have not yet bought and/or used the respective hypothetical AI products. Potential rather than actual users are of particular interest regarding AI trustworthiness labeling, as they have even more limited access to information on AI products and corresponding trustworthiness cues than actual users.

Second, I aimed to replicate (trust, attributed value) and extend (to acceptance, intention to use, and brand evaluation) their findings for a global AI trustworthiness label that represents a conjunction trustworthiness score (integrating multiple trustworthiness dimensions rather than only focusing on data security and data privacy like Pfeuffer [47, 48]). Specifically, I developed trustworthiness labels based on the seven criteria of trustworthy AI put forward by the European Commission [15]. To achieve optimal comparability between the results of Pfeuffer [47, 48] and the present study, I focused on the same AI product types. Moreover, I used their exact graphical labels, now representing a global AI trustworthiness score in this study. The graphical instead of text-based label type is particularly suited for this endeavor, as it does not contain any text elements which would, inevitably, vastly differ between single-dimensional (Pfeuffer [47, 48]) and global, multi-dimensional (the present study) text-based AI trustworthiness labels.

Third, I further aimed to extend the study of Pfeuffer [47, 48] by additionally assessing the impact of advertisement statements on the perception and evaluation of AI with and without AI trustworthiness labels (see e.g., [26], for evidence that labels such as the Nutri-Score can guard against advertisement claims). Correspondingly, I expected to find that trust, acceptance, intention to use, attributed value, and brand evaluation would increase for AI products with low to high trustworthiness levels as indicated by the corresponding label. In line with the idea that labels can shield (potential) users against advertisement influences, I further expected to see a small or no impact of advertisement statements on labeled AI products. However, I expected the presence versus absence of advertisement statements to affect the perception and evaluation of unlabeled AI products (at baseline).

Fourth, the findings of Pfeuffer [47, 48] pointed towards a bias of (potential) users to attribute intermediate levels of trustworthiness (there only related to data security and data privacy) to AI products when no further information is available. It is essential to replicate and confirm this bias as well as assess to what degree it translates to the acceptance of and the intention to use AI products to ascertain how heavily potential users may unjustifiedly show (dis-)trust, (in-)acceptance, and (non-)intention to use AI products in lack of AI trustworthiness information. This assessment is most relevant to determine the need for corresponding legislation regarding labels conveying the trustworthiness of AI products. Correspondingly, I expected to replicate the finding of Pfeuffer [47, 48] that baseline

trust ratings roughly correspond to the perception and evaluation of AI products of an intermediate AI trustworthiness level and I expected to observe a similar pattern for acceptance and intention to use ratings.

Finally, fifth, I aimed to exploratorily assess the influence potential users' individual characteristics (specifically their general trust in AI irrespective of a specific AI product, i.e., product-unspecific, and their AI literacy) had on the impact AI trustworthiness information conveyed by a corresponding label had on their perception and evaluation of AI products.

This study makes the following empirical contributions to the literature:

- It extends the work of Pfeuffer [47, 48] to a more realistic and ecologically valid setting by providing an assessment context that more accurately reflects real-life AI product advertisement encounters.
- It replicates the findings of Pfeuffer [47, 48] regarding the impact of AI trustworthiness levels indicated by a corresponding label on (potential) users' trust. In particular, it also replicates and thus confirms Pfeuffer's [47, 48] novel finding that (potential) users show a bias to attribute intermediate trustworthiness to AI in the absence of information. The study further extends these findings beyond trust, assessing and demonstrating parallel patterns for AI acceptance and intention to use.
- Moreover, I used the same graphical label type as Pfeuffer [47, 48] to reflect trustworthiness in a global, multi-dimensional context. This allowed me to comparatively assess differences between global, multi-dimensional (the present study) and single-dimensional AI trustworthiness labels (Pfeuffer [47, 48]).
- Additionally, this study is the first to consider the impact of label-indicated AI trustworthiness levels (singularily and in conjunction with advertisement claims) on brand evaluation in this context.
- Finally, this study extends the findings of Pfeuffer [47, 48] by assessing and confirming the moderating effect of a person's individual propensity to trust AI and AI literacy on how effective AI trustworthiness labels are in altering (potential) users' perceptions and evaluations of AI products.

3 Methods

This study was preregistered in the Open Science Framework (OSF; https://doi.org/10.17605/OSF.IO/W6PCM). The materials, data, and analysis scripts are available via OSF (https://doi.org/10.17605/OSF.IO/4ZFUW). This study was approved by the local ethics committee of the Catholic University of Eichstätt-Ingolstadt and adheres to local laws and regulations as well as international standards regarding research involving human participants.

3.1 Participants

A prior study on data security and data privacy labels for AI products assessed 102 participants to find conclusive Bayesian evidence for the impact of an AI product's data security and data privacy level conveyed by a corresponding label on potential users' trust in, anxiety towards, and monetary value attributed to AI products [47, 48]. The impact of multi-level AI trustworthiness labels on the acceptance or intention to use AI products as well as on the evaluation of the corresponding brand has not been assessed previously. Thus, I followed a Bayesian approach for sample size determination [53]. I started with a sample of 100 participants and planned to increase sample size in groups of 20 participants until the Bayesian stopping criterion was reached. My stopping criterion was a Bayes factor BF_{10} (in favour of an effect) or BF_{01} (against an effect/in favour of the null hypothesis) larger than 3 (or smaller than 1/3) for the effect of trustworthiness level and advertisement on the dependent variables trust, acceptance, and intention to use.

The Bayesian stopping criterion was met at sample size 100. All participants provided written informed consent prior to their participation and received monetary compensation for taking part. Participants who aborted the experiment or who failed one of the attention checks were excluded and replaced. Participants (members of the general public who indicated their willingness to participate in online studies; at least 18 years old, German native speakers, residing in Germany or Austria) were recruited via Prolific and the final sample included 100 participants (71 male, 27 female, 2 diverse; age: M = 34.6 years, SD = 10.2 [range: 19 – 68 years]; education: 2 low-level secondary school leaving certificate without and 3 with vocational training, 4 intermediate-level secondary school leaving certificate, 37 A levels/high school diploma, 49 university degree). This European sample was intentionally chosen, as the described criteria for an AI trustworthiness label's level I selected were the seven criteria for trustworthy AI issued by the European Commission [15]. Thus, I selected a sample who would directly be affected by legislation related to these criteria and/or the introduction of corresponding labels. Participants' general, product-unspecific trust in automation/AI, assessed at the beginning of the study via the propensity to trust and trust in automation subscales of the trust in automation questionnaire (TiA; [28]; 5 items; Likert scale from 1 – strongly disagree to 5 – strongly agree; continuous predictor for exploratory analyses), was 15.0 on average (SD = 3.8; [5; 24]). At the end of the study, I additionally assessed participants' AI literacy using the Meta AI literacy scale (MAILS; [5]) which assesses the superordinate factors AI literacy (subscales apply AI, understand AI, detect AI, AI ethics; continuous predictor for exploratory analyses), AI self-efficacy (subscales AI problem solving, learning), and AI self-competency (subscales AI persuasion literacy, AI emotion regulation), and the independent subscale create AI (34 items total; Likert scale from 0 to 10 to indicate individual ability/competence). Participants reported an average of 125.6 (SD = 28.5; [51;179]) on the scales of the superordinate AI literacy factor, a mean score of 7.3 (SD = 11.6; [0;37]) on the create AI subscale, an average of 37.7 (SD = 12.3; [3;60]) on the scales of the superordinate self-efficacy factor, and an average of 44.1 (SD = 9.1; [24;60]) on the scales of the superordinate self-competency factor of the MAILS.

3.2 Stimuli and Apparatus

Participants took part online on their own computers, laptops, or tablets. Participation on mobile phones was not allowed to ensure that stimuli were displayed sufficiently large to inspect all details.

Two types of (hypothetical) AI products (smart fridges, voice assistants) were used. These two AI product types were deliberately

chosen as they were previously used in the study of Pfeuffer [47, 48] and therefore allowed for direct comparisons between the present results and the results of Pfeuffer [47, 48] as well as clear attributions of potential differences in findings. AI product types were introduced and their general functions briefly described at the beginning of the study. Specifically, the central features typically fulfilled by these two types of AI products (e.g., smart fridge: supervising and restocking grocery supplies) as well as the data commonly collected by them (e.g., smart fridge: credit card information for directly ordering groceries) were briefly described in a short information paragraph on the AI products. Per AI product type, eight different, AI-generated images of corresponding AI products (one per combination of trustworthiness level and advertisement condition) were used and randomly assigned to conditions. Furthermore, each AI product was randomly assigned the logo of a (hypothetical) brand. The 16 brand names were pseudowords (UniPseudo, [44], English words, length: 5-6 letters) and the corresponding brand logos containing these brand names were generated using AI. For the advertisement conditions, per AI product type, four different advertisement texts (German) matching the respective product were assigned to the four trustworthiness level conditions (see Figure 1 for exemplary stimuli).

Trustworthiness labels were three-level, graphical labels that resembled a horizontal traffic light (traffic light circles/trustworthiness levels: green = high trustworthiness, yellow = intermediate trustworthiness, red = low trustworthiness; the filled circle indicated the AI product's trustworthiness level; see e.g., [47, 48, 55], for label types used in prior studies). The trustworthiness labels were introduced at the beginning of the study following the introduction of the AI product types. Participants were instructed about the trustworthiness levels and informed that the trustworthiness levels indicated by the trustworthiness label represented the degree to which a corresponding AI product adhered to the seven criteria of trustworthy AI put forward by the European Commission [15] which were each briefly described. Please note that the focus of this study was not to investigate the optimal thresholds for AI trustworthiness levels for corresponding AI trustworthiness labels, but to assess how participants evaluated AI products presented with as compared to without such AI trustworthiness labels. Thus, participants received detailed information on the trustworthiness criteria, but were only informed that the trustworthiness labels indicated whether all of these criteria were jointly overall met to a low (red traffic light label), intermediate (yellow traffic light label), or high degree (green traffic light label). I intentionally did not communicate any specific level allocation criteria, thresholds underlying AI trustworthiness levels, or information on how the seven AI trustworthiness criteria were combined into this global AI trustworthiness score. This was done to ensure that the impact of AI trustworthiness labels could not be attributed to the exact criteria underlying the different AI trustworthiness levels in this study. That is, the label levels were intentionally illustrative without a formal mapping. I chose this approach as, in my opinion, at present, objective criteria for attributing AI trustworthiness levels (globally across criteria or locally for a single criterion) do not yet exist to a degree sufficient for determining/computing the trustworthiness of a particular AI product. This approach is further parallel with to the approach of Pfeuffer [47, 48] and thus supports the comparability of results. In my opinion, setting objective criteria necessitates a transdisciplinary effort (see also e.g., [31]) once the effectiveness of AI trustworthiness labels (i.e., their impact on potential users' perceptions and evaluations of AI) is firmly established. Furthermore, as the present study used only hypothetical AI products, the AI features of which were not detailed, there were no objective criteria for allocating AI trustworthiness levels to these hypothetical AI products in the present context. This, in turn, allowed me to randomly pair AI product pictures, brand logos, and AI trustworthiness labels to thus better partial out the impact of AI product appearance in my analyses.

3.3 Design and Procedure

Participants first provided informed consent, filled in demographic information as well as answered five items of the TiA questionnaire ([28]; subscales propensity to trust and trust in automation; Likert scale from 1 – strongly disagree to 5 – strongly agree; continuous predictor for exploratory analyses). Then, they received information on the two exemplary AI product types (smart fridge, voice assistant) used in this study and their function as well as on the AI trustworthiness labels and the meaning of the three trustworthiness levels indicated by them (low, intermediate, high degree of adherence to the 7 criteria for trustworthy AI set forth by the European Commission, including a brief description of each of the 7 criteria; see [15]). Two attention check questions were integrated in this part to filter out and replace inattentive participants.

The subsequent experiment consisted of three phases (see Figure 1 for study design and time course). In the first phase, the label phase, trustworthiness level (baseline/none vs. low vs. intermediate vs. high) and advertisement (present vs. absent) were systematically manipulated and participants' trust in, acceptance of, an intention to use AI products were assessed (16 trials, 8 per AI product type: 4 trustworthiness levels x 2 advertisement conditions). The following label comparison phase compared two AI trustworthiness label levels per trial (trustworthiness level comparison: low-intermediate vs. intermediate-high vs. low-high) and assessed the monetary value participants attributed to AI products that reached the respective higher trustworthiness level (6 trials, 3 per AI product type: 3 trustworthiness level comparisons). The final brand evaluation phase assessed how the evaluation of the (hypothetical) brands (represented by their logos) was affected by the trustworthiness level and advertisement conditions encountered in the label phase (16 trials: 4 trustworthiness levels x 2 advertisement conditions x 2 AI product types).

In the label phase, per trial, participants were presented with one image of a (hypothetical) AI product together with a (hypothetical) brand logo, an AI trustworthiness label (AI trustworthiness labels were absent in the baseline conditions), and a brief advertisement text related to the function and potential of the respective product (in the advertisement present conditions only; e.g., "Hört zu. Denkt mit."/Eng: "Listens. Thinks things through."; see Figure 1 for an example of the stimuli). A trial's AI product image, brand logo, AI trustworthiness label (if applicable), and advertisement text (if applicable) were presented together for 4 seconds first. Then, the dependent variable questions appeared below and participants were able to answer. Per trial, participants were asked to rate their

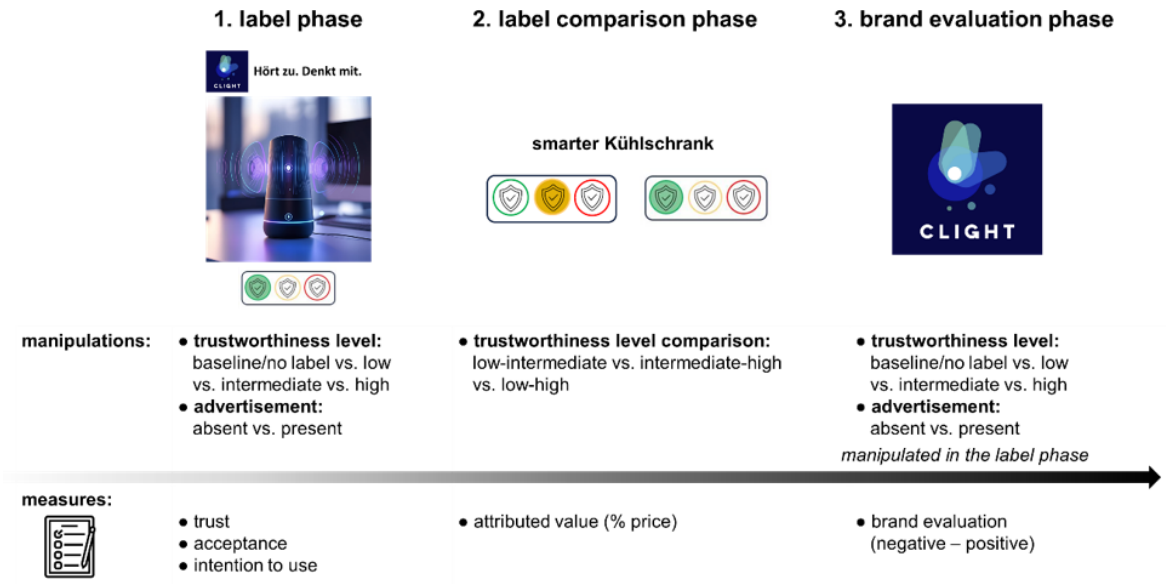


Figure 1: Study Design and Time Course

trust in (trust in automation subscale of the trust in automation questionnaire; [28]; 2 items; Likert scale from 1 – strongly disagree to 5 – strongly agree), acceptance of (3 items; [19]; Likert scale from 1 – strongly disagree to 5 – strongly agree), and intention to use (behavioral intention subscale of the Unified Theory of Acceptance and Use of Technology, UTAUT2, questionnaire; [57]/German translation: [21]; 3 items; Likert scale from 1 – strongly disagree to 7 – strongly agree) the corresponding AI product. Participants either first encountered the eight trials of the smart fridges and then the eight trials of the voice assistants or vice versa (condition order, AI product image, and brand logo independently randomized).

Subsequently, in each trial of the label comparison phase, participants were informed which AI product type (smart fridge/"smarter Kühlschrank" vs. voice assistant/"KI-gestützter Sprachassistent") they should consider (no AI product images were shown) and were presented with two AI trustworthiness labels (trustworthiness level comparison: low-intermediate vs. intermediate-high vs. low-high) on the left (respective lower trustworthiness level) and right (respective higher trustworthiness level). They were asked to indicate how much more they would be willing to pay for an AI product of the respective higher AI trustworthiness level (percent price relative to the AI product with the respective lower AI trustworthiness level). Again, the order of AI product types was randomized. Per AI product type, the order of AI trustworthiness level comparisons was randomized.

In each trial of the final brand evaluation phase, one of the brand logos used in the label phase was presented again and participants were asked to evaluate the brand logos (Likert scale from 1 – very negative to 9 – very positive).

At the end of the study, participants filled out the MAILS questionnaire to assess their AI literacy (Meta AI literacy scale; [5]; 34 items across 9 subscales, representing 4 superordinate factors; Likert scale from 0 to 10 to indicate individual ability/competence in the respective AI-related domain). They were then debriefed and had the chance to provide comments on anything they deemed relevant before ending the study and receiving their monetary compensation.

4 Results

4.1 Data Preparation

The scores for trust, acceptance, and intention to use were computed by averaging the ratings on the items of the respective scales per trial (i.e., per participant, condition, and AI product type). For the analyses of trust, acceptance, and intention to use across trustworthiness levels and advertisement conditions, difference scores, computed by subtracting the respective rating in the baseline - no label, no advertisement condition from each condition with a label, were used. This allowed me to directly display the result pattern as deviations from baseline due to the presented AI trustworthiness labels (method adapted from Pfeuffer [47, 48]). I further used the raw scores for trust, acceptance, and intention to use to compare the baseline conditions in which advertisement was present and absent to quantify the impact of the advertisement on AI product evaluation in the absence of AI trustworthiness labels. For the analyses of attributed value, I excluded outliers (2.8%), that is, answers that deviated by more than 3 SDs from the sample mean. For attributed value and brand evaluation, the raw answers for the respective question were entered into the analyses.

For exploratory analyses, I additionally considered the influence of a person's product-unspecific general trust in automation/AI (initial TiA questionnaire, propensity to trust) and their AI literacy (superordinate factor AI literacy of the MAILS questionnaire). For these analyses, TiA and AI literacy sum scores were z-standardized to be used as continuous predictors.

4.2 Data Analyses

I chose a Bayesian analysis approach to quantify both evidence in favour of as well as against an effect of trustworthiness level/trustworthiness level comparison and advertisement on the dependent variables of interest. Bayes factors BF_{10} show the evidence in favour of an effect (i.e., the alternative hypothesis), whereas Bayes factors BF_{01} quantify the evidence against an effect (i.e., in favour of the null hypothesis). I report all results in terms of Bayes factors BF_{10} in favour of an effect. That is, a $BF_{10} = 4$ indicates that the data are 4 times more likely given the data under the alternative hypothesis than the null hypothesis (i.e., a corresponding effect exists). Conversely, a $BF_{10} = 0.25$ (1/4) indicates that the data are 4 times more likely under the null hypothesis (i.e., there is evidence against an effect). All analyses relied on the default Jeffreys–Zellner–Siow (JZS) priors (Cauchy distributions with $r = \sqrt{2}/2$), which are standard objective priors for Bayes factor computation. These priors are designed to provide stable evidence quantification without imposing strong prior beliefs about effect size magnitude. Furthermore, this prior choice equates the analysis approach of Pfeuffer [47, 48] and thus optimized the comparability of results.

I conducted Bayesian linear mixed model (BLMM) analyses per dependent variable and compared models that did versus did not include a respective effect of interest to calculate Bayes factors BF_{10}. The maximum model per dependent variable included the fixed effects of trustworthiness level (low vs. intermediate vs. high), advertisement (present vs. absent), and their interaction (dependent variables: trust, acceptance, intention to use, brand evaluation) or the fixed effect of trustworthiness level comparison (low-intermediate vs. intermediate-high vs. low-high; dependent variable: attributed value). For the random effects structure, I compared a (fixed effects) maximum model that included intercepts for participants, AI product images, and brand logos (only participants for attributed value) and random slopes for all independent variables and their interaction against a (fixed effects) maximum model that only included the intercepts. When the corresponding BF_{10} indicated that the random slopes contributed to explaining the data, random slopes were included for all model comparisons. When the corresponding BF_{10} indicated that the random slopes did not contribute to explaining the data, models that only contained random intercepts were used for model comparisons. Note that I had indicated that I would separately check the contribution of each random effect in the preregistration of the study but shifted to this approach to reduce computation effort. $BF_{10} > 3$ or $< 1/3$ were considered as conclusive.

For the dependent variables trust, acceptance, and intention to use, I additional conducted Bayesian LMMs to compare ratings at baseline (without an AI trustworthiness label) for AI products displayed with versus without advertisement. These analyses followed the same principles and the maximum model included a fixed effect of advertisement, random intercepts for participants, AI products, and brand logos, and a random slope for advertisement.

Exploratory analyses per dependent variable additionally included the z-standardized continuous predictors product-unspecific general trust in AI (initial TiA score) or AI literacy (superordinate factors AI literacy of the MAILS questionnaire) as well as all possible interactions between each singular continuous predictor and the fixed effects. As the two continuous predictors were correlated, $t(98) = 3.13$, $p = .002$, $r = .30$, separate models each including only one continuous predictor were computed. Due to the computational effort, these exploratory analyses only used models with the corresponding random intercepts, but not random slopes. Moreover, for the exploratory analyses, I started with a minimum fixed effects model that included fixed effects of trustworthiness level, advertisement, and their interaction (dependent variables: trust, acceptance, intention to use, brand evaluation) or the fixed effect of trustworthiness level comparison (dependent variable: attributed value) and tested whether models that additionally included an effect of or interactions with the respective continuous predictor better accounted for the observed data pattern.

All Bayesian model comparisons indicated adequate uncertainties of the estimates. Note that conclusions were stable across different plausible random-effects structures, indicating robustness against model specification.

4.3 Baseline Evaluations of AI products

Trust ratings at baseline (without trustworthiness label or advertisement) were 2.54 (SD = 0.96) for the smart fridges and 2.41 (SD = 0.95) for the voice assistants. Acceptance ratings at baseline (without trustworthiness label or advertisement) were 2.45 (SD = 1.0) for smart fridges and 2.42 (SD = 0.95) for voice assistants. Intention to use ratings at baseline (without trustworthiness label or advertisement) were 2.62 (SD = 1.67) for smart fridges and 2.50 (SD = 1.42) for voice assistants (see Figure 2A-2C for data distributions).

4.4 Trust

As expected, trust increased with higher trustworthiness levels indicated by the AI trustworthiness labels, $BF_{10} = 7.34 \times 10^{231}$ ±2.96%. There was conclusive evidence against an impact of advertisement, $BF_{10} = 1.66 \times 10^{-8}$ ±1.11%, or the interaction of trustworthiness level and advertisement, $BF_{10} = 7.56 \times 10^{-8}$ ±3.17%, on trust. At baseline (without an AI trustworthiness label), there was conclusive evidence against a contribution of advertisement, $BF_{10} = 0.16$ ±0.93% (see Figure 2D).

The exploratory analysis including product-unspecific general trust in AI showed conclusive evidence against an effect of product-unspecific general trust in AI on trust at the level of individual AI products, $BF_{10} = 0.19$ ±2.85%, as well as against the interaction of advertisement and product-unspecific general trust in AI, $BF_{10} = 0.11$ ±2.25%, and against the three-way interaction, $BF_{10} = 0.02$ ±21.46%. There was conclusive evidence in favour of an interaction between trustworthiness level and product-unspecific general trust in AI, $BF_{10} = 1.09 \times 10^{4}$ ±2.13% (see Figure 4A). Descriptively, for AI products of higher trustworthiness levels, trust increased with increasing product-unspecific general trust in AI, whereas for AI products with low trustworthiness levels, trust decreased with increasing product-unspecific general trust in AI. Similarly, the exploratory analysis including AI literacy showed conclusive evidence against an effect of AI literacy, $BF_{10} = 0.19$ ±1.60%, against the interaction of advertisement and AI literacy, $BF_{10} = 0.13$ ±2.74%, and against the three-way interaction, $BF_{10} = 0.04$ ±3.84%. However, there was conclusive evidence for an interaction of trustworthiness level and AI literacy, $BF_{10} = 8.08$ ±1.58% (see Figure 4B). Again, trust

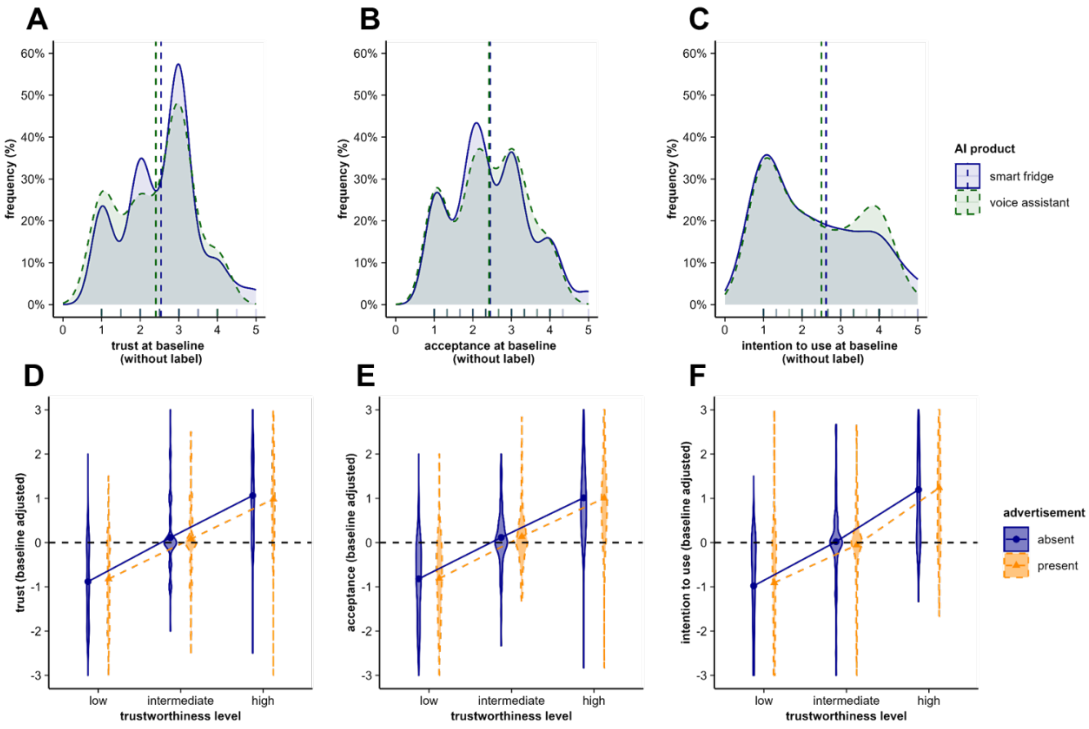


Figure 2: Trust, Acceptance, and Intention to Use Ratings at Baseline as well as per Trustworthiness Level and Advertisement Condition. A) to C) show participants' ratings in the baseline condition (without AI trustworthiness label) in the absence of advertisement text. D) to F) display participants' ratings relative to a participant's respective baseline (without label and without advertisement) rating of the corresponding AI product per trustworthiness level (low vs. intermediate vs. high) and advertisement (absent vs. present) condition. A value of 0 indicates a rating equivalent to a participant's rating of the AI product at baseline without any trustworthiness label or advertisement text. Values below 0 indicate ratings lower than at baseline and values above 0 indicate ratings higher than at baseline. Violins around the respective mean depict the corresponding rating distribution per condition.

ratings descriptively increased with increasing AI literacy for AI products of high trustworthiness, whereas trust ratings decreased with rising AI literacy for AI products of low trustworthiness.

Assessing the impact of the continuous predictors at baseline, expectedly, I found conclusive evidence that trust ratings increased with product-unspecific general trust in AI scores, $BF_{10} = 8.66 \times 10^{7} \pm 3.58\%$. There was conclusive evidence against an interaction of advertisement and product-unspecific general trust in AI, $BF_{10} = 0.07 \pm 3.33\%$. For the continuous predictor AI literacy, I also observed increases in trust at baseline with increasing AI literacy scores, $BF_{10} = 27.01 \pm 1.15\%$, but conclusive evidence against an interaction of advertisement and AI literacy, $BF_{10} = 0.12 \pm 7.02\%$.

4.5 Acceptance

Acceptance increased with higher trustworthiness levels indicated by the AI trustworthiness labels, $BF_{10} = 4.89 \times 10^{233} \pm 1.28\%$. There was conclusive evidence against an impact of advertisement, $BF_{10} = 3.48 \times 10^{-9} \pm 1.54\%$, and against the interaction of trustworthiness level and advertisement, $BF_{10} = 8.06 \times 10^{-9} \pm 1.89\%$, on acceptance. At baseline (without an AI trustworthiness label), there was conclusive evidence against a contribution of advertisement, $BF_{10} = 0.02 \pm 1.12\%$ (see Figure 2E).

The exploratory analysis including product-unspecific general trust in AI showed conclusive evidence against an effect of product-unspecific general trust in AI on acceptance, $BF_{10} = 0.21 \pm 1.94\%$, as well as against the interaction of advertisement and product-unspecific general trust in AI, $BF_{10} = 0.11 \pm 6.05\%$, and against the three-way interaction, $BF_{10} = 0.02 \pm 19.72\%$. There was conclusive evidence in favour of an interaction between trustworthiness level and product-unspecific general trust in AI, $BF_{10} = 9.65 \times 10^{6} \pm 2.57\%$ (see Figure 4C). Descriptively, for AI products of higher trustworthiness levels, trust increased with increasing product-unspecific general trust in AI, whereas for AI products with low trustworthiness levels, trust decreased with increasing product-unspecific general trust in AI. The exploratory analysis including AI literacy revealed inconclusive evidence against an effect of AI literacy on acceptance, $BF_{10} = 0.37 \pm 2.81\%$, as well as conclusive evidence against the interaction of advertisement and AI literacy, $BF_{10} = 0.10 \pm 2.34\%$, and against the three-way interaction, $BF_{10} = 0.04 \pm 2.94\%$. I found

conclusive evidence in favour of an interaction between trustworthiness level and AI literacy, $BF_{10} = 9.65 \times 10^{6}$ ±2.57% (see Figure 4D). Descriptively, AI products of higher trustworthiness levels showed increases in acceptance with increasing AI literacy. Conversely, AI products with low trustworthiness levels, showed decreases in acceptance with increasing AI literacy.

At baseline, I found conclusive evidence that acceptance ratings increased with product-unspecific general trust in AI scores, $BF_{10} = 4.66 \times 10^{4}$ ±1.06%. Furthermore, there was conclusive evidence against an effect of the interaction between advertisement and product-unspecific general trust in AI on acceptance, $BF_{10} = 0.08$ ±1.43%. For the continuous predictor AI literacy at baseline conditions, acceptance increased with increasing AI literacy scores, $BF_{10} = 28.64$ ±2.05%. There was conclusive evidence against an interaction of advertisement and AI literacy, $BF_{10} = 0.09$ ±1.59%.

4.6 Intention to Use

In line with my expectations, intention to use increased with higher trustworthiness levels indicated by the AI trustworthiness labels, $BF_{10} = 3.02 \times 10^{168}$ ±1.87%. There was conclusive evidence against an impact of advertisement, $BF_{10} = 2.02 \times 10^{-9}$ ±1.84%, and against the interaction of trustworthiness level and advertisement, $BF_{10} = 5.17 \times 10^{-9}$ ±2.44%, on the intention to use the AI products. At baseline (without an AI trustworthiness label), there was conclusive evidence against a contribution of advertisement, $BF_{10} = 0.12$ ±0.65% (see Figure 2F).

The exploratory analysis including product-unspecific general trust in AI showed conclusive evidence against an effect of product-unspecific general trust in AI on the intention to use AI products, $BF_{10} = 0.21$ ±2.94%, as well as against the interaction of advertisement and product-unspecific general trust in AI, $BF_{10} = 0.14$ ±2.28%, and against the three-way interaction, $BF_{10} = 0.02$ ±2.33%. Yet, there was conclusive evidence in favour of an interaction between trustworthiness level and product-unspecific general trust in AI, $BF_{10} = 2.72 \times 10^{12}$ ±12.0% (see Figure 4E). Intention to use descriptively increased with increasing product-unspecific general trust in AI for AI products of higher trustworthiness levels, whereas the intention to use AI products decreased with increasing product-unspecific general trust in AI for AI products of low trustworthiness as indicated by the AI trustworthiness label. The exploratory analysis including AI literacy found conclusive evidence against an effect of AI literacy on intention to use, $BF_{10} = 0.27$ ±6.95%, as well as against the interaction of advertisement and AI literacy, $BF_{10} = 0.10$ ±7.72%, and against the three-way interaction, $BF_{10} = 0.03$ ±1.68%. There was conclusive evidence in favour of an interaction between trustworthiness level and AI literacy, $BF_{10} = 1.71 \times 10^{4}$ ±6.95% (see Figure 4F). Descriptively, the intention to use an AI product increased with increasing AI literacy for AI products of higher trustworthiness levels, whereas the intention to use AI products decreased with increasing AI literacy for AI products of low trustworthiness.

At baseline, I found conclusive evidence that intention to use ratings increased with product-unspecific general trust in AI scores, $BF_{10} = 9.36 \times 10^{5}$ ±1.59%. There was conclusive evidence against an effect of the interaction between advertisement and product-unspecific general trust in AI on the intention to use AI products, $BF_{10} = 0.08$ ±1.51%. At baseline, intention to use ratings further increased with increasing AI literacy scores, $BF_{10} = 30.23$ ±1.21%. There was conclusive evidence against an interaction of advertisement and AI literacy, $BF_{10} = 0.08$ ±1.53%.

4.7 Attributed Value

In line with my expectations, attributed value increased from the trustworthiness level comparison low-intermediate to intermediate-high and low-high, $BF_{10} = 2.55 \times 10^{46}$ ±0.56% (see Figure 3A).

The exploratory analysis including the continuous predictor product-unspecific general trust in AI showed inconclusive evidence against an effect of product-unspecific general trust in AI, $BF_{10} = 0.37$ ±2.05%, as well as conclusive evidence against an interaction of trustworthiness level comparison and product-unspecific general trust in AI, $BF_{10} = 0.05$ ±3.62% (see Figure 4I). The exploratory analysis including the continuous predictor AI literacy showed conclusive evidence against an effect of AI literacy, $BF_{10} = 0.329$ ±3.73%, as well as against the interaction of trustworthiness level comparison and AI literacy, $BF_{10} = 0.04$ ±3.51% (see Figure 4J).

4.8 Brand Evaluation

Contrary to my expectations, there was conclusive evidence against an impact of trustworthiness level on brand evaluations, $BF_{10} = 0.04$ ±2.73%. Evidence against an impact of advertisement, $BF_{10} = 0.29$ ±1.17%, and of the interaction between trustworthiness level and advertisement, $BF_{10} = 0.25$ ±1.57%, on brand evaluations was also conclusive (see Figure 3B).

The exploratory analysis including the continuous predictor product-unspecific general trust in AI revealed that brand evaluations improved with higher product-unspecific general trust in AI scores, $BF_{10} = 601.96$ ±8.52%. Furthermore, there was conclusive evidence against interactions of trustworthiness level and product-unspecific general trust in AI, $BF_{10} = 0.03$ ±8.77%, against advertisement and product-unspecific general trust in AI, $BF_{10} = 0.09$ ±9.54%, and against a three-way interaction of trustworthiness level, advertisement, and product-unspecific general trust in AI, $BF_{10} = 0.03$ ±3.73% (see Figure 4F). The exploratory analysis including the continuous predictor AI literacy showed that brand evaluations improved with higher AI literacy scores, $BF_{10} = 42.29$ ±3.06%. There was conclusive evidence against interactions of trustworthiness level and AI literacy, $BF_{10} = 0.01$ ±3.09%, and advertisement and AI literacy, $BF_{10} = 0.09$ ±3.07%, and inconclusive evidence against a three-way interaction of trustworthiness level, advertisement, and AI literacy, $BF_{10} = 0.40$ ±5.43% (see Figure 4G).

5 Discussion

The aim of this study was threefold. First, I wanted to replicate (trust, attributed value) and extend (acceptance, intention to use, brand evaluation) the findings of a prior study by Pfeuffer [47, 48], assessing the impact of data security and data privacy labels on the perception and evaluation of AI products, to global AI trustworthiness labels based on the criteria for trustworthy AI put forward by the European Commission [15]. Second, I aimed to replicate and extend the corresponding findings of Pfeuffer [47, 48] regarding potential users' biased, unfounded, intermediate trust in unlabeled

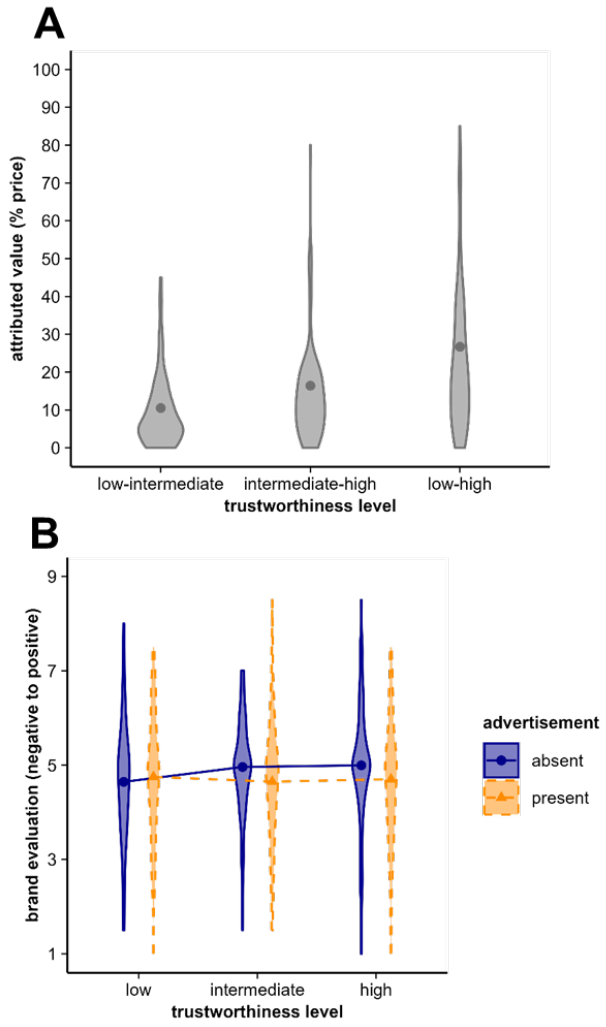


Figure 3: Attributed Value and Brand Evaluation per Condition. A) shows the value attributed to an AI product of a higher trustworthiness level (% price of the AI product of lower trustworthiness participants are willing to pay to receive the respective AI product of higher trustworthiness) for each trustworthiness level comparison (low-intermediate vs. intermediate-high vs. low-high). B) depicts brand evaluation (negative to positive) per trustworthiness level (low vs. intermediate vs. high) and advertisement condition (absent vs. present).

AI products. Third, I aimed to explore the influence of potential users' product-unspecific general trust in AI products and their AI literacy on the effect AI trustworthiness labels had on their perception and evaluation of AI products. To do so, I paired hypothetical AI products with hypothetical brand logos and either presented an AI trustworthiness label (indicating low vs. intermediate vs. high trustworthiness) together with the AI product-brand logo pair or did not present an AI trustworthiness label (baseline condition without label). Furthermore, I manipulated whether an advertisement text was shown with the AI product or not. I assessed the impact of the AI trustworthiness label's trustworthiness level and the advertisement on participants' trust in, acceptance of, and intention to use AI products, as well as their evaluation of corresponding brands. In another phase of the experiment, I further compared AI trustworthiness labels of two different trustworthiness levels (low-intermediate vs. intermediate-high vs. low-high) and assessed how much more participants were willing to pay (attributed value) to receive the AI product of the respective higher trustworthiness level.

As hypothesized, perceptions and evaluations of AI products scaled with AI trustworthiness level as indicated by the AI trustworthiness label. Participants showed more trust (replicating [47, 48]), more acceptance, and a more pronounced intention to use AI products labeled with a higher as compared to lower trustworthiness levels. Furthermore, participants were willing to pay more (attributed value) for AI products that achieved a higher trustworthiness level (replicating [47, 48]). These findings replicate the findings of Pfeuffer [47, 48] for trust and attributed value and extend them to more realistic and ecologically valid AI product information material (as compared to generic icons used in [47, 48]) as well as to measures of acceptance and the intention to use AI products. Overall, these findings demonstrate, first and most importantly, that labels conveying AI trustworthiness information in a quick-to-process format effectively communicate AI trustworthiness and correspondingly affect potential users' perception and evaluation of AI products. Second, the present findings show that trustworthiness information conveyed by AI trustworthiness labels also yields an influence on the perception and evaluation of AI products when individual products, brand logos, and advertisement information are displayed (realistic and ecologically valid advertisement setting) rather than only the same generic placeholder product icon (as was the case in [47, 48]). Third, whereas Pfeuffer [47, 48] used a label to convey AI trustworthiness information related to a single dimension (data security and data privacy), the present study incorporated (visually identical) global AI trustworthiness labels that represented the adherence of an AI to a conjunction of trustworthiness criteria (the seven criteria for trustworthy AI put forward by the European Commission; [15]). Nonetheless, I demonstrated that AI perceptions and evaluations scaled based on the trustworthiness level indicated by this label reflecting a conjunction of trustworthiness criteria. This indicates that both overarching, global, multi-dimensional labels reflecting a conjunction of criteria as well as labels reflecting a single evaluation dimension are effective in communicating trustworthiness and in helping (potential) users calibrate their individual AI trustworthiness assessments.

Note, however, that I observed evidence against an impact of AI trustworthiness labels on the evaluation of corresponding AI product brands. Prior research on evaluative conditioning (see e.g., [9, 43], for reviews) in the marketing context showed that manipulations pairing brands with affective stimuli ([43] for a short review) and manipulations which shift product evaluations to the negative or positive can also affect brand evaluation [45]. Yet, studies demonstrating such evaluative conditioning effects most often used multiple pairings of stimuli and a single stimulus pairing (as

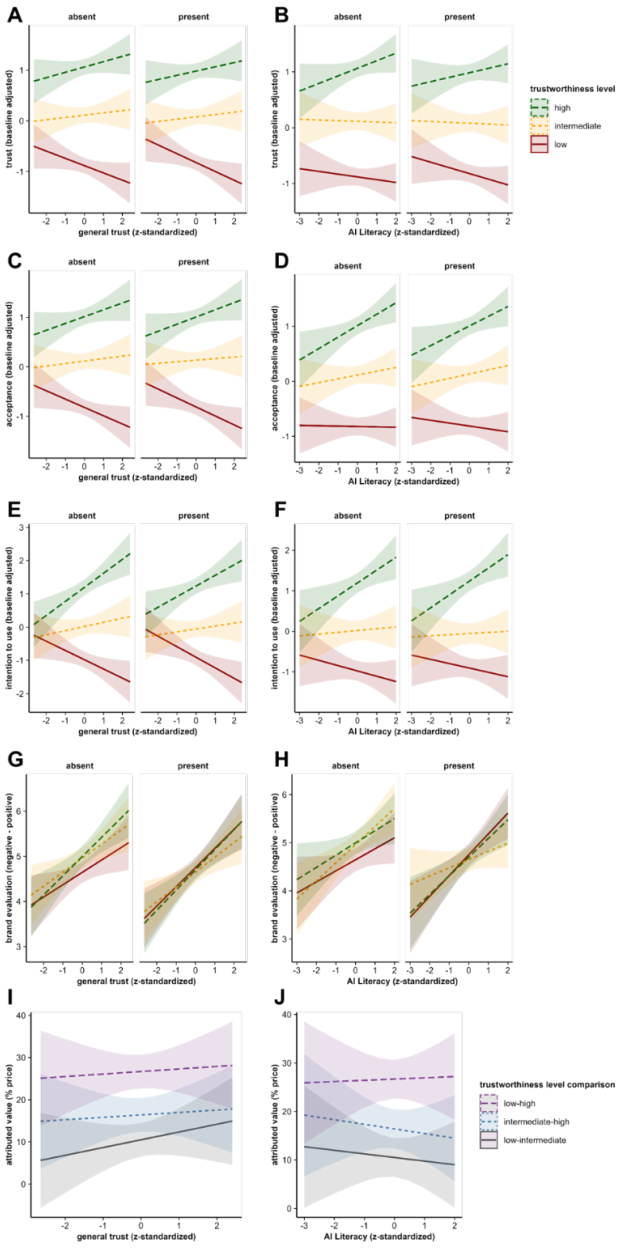


Figure 4: Results of the Exploratory Analyses per Dependent Variable. A) to H) depict the influence of product-unspecific general trust in AI (A, C, E, G) and AI literacy (B, D, F, H; continuous predictors) on the dependent variables trust, acceptance, intention to use, and brand evaluation as well as their interactions with the fixed effects trustworthiness level (low vs. intermediate vs. high; colors corresponding to trustworthiness label's colours per level) and advertisement (absent vs. present). I) and J) show the influence of product-unspecific general trust in AI (I) and AI literacy (J) on attributed value as well as their interactions with the fixed effect trustworthiness level comparison (low-intermediate vs. intermediate-high vs. low-high).

used here) is rarely sufficient for evaluative conditioning effects to emerge, as corresponding associative learning based on a single exposure is too weak (but see [54]). Thus, it remains for a future study to ascertain whether AI trustworthiness labels affect brand evaluations when participants encounter paired AI products, brand logos, and AI trustworthiness labels multiple times before evaluating brands. A first corresponding study, could, for instance, directly built on the present experiment, focus only on pairing AI product-brand logo pairs with an AI trustworthiness label of a low to high level, and show each AI product-brand-label combination multiple times before assessing brand evaluation. Specifically, each time an AI product-brand-label combination is presented, only questions regarding one dependent variable (e.g., trust, acceptance, intention to use, and possibly additional dependent variables of interest, e.g., risk/stakes assessment) could be asked or even only a single question could be asked each time. This would not unduly extend the study duration, but nonetheless lead to multiple exposures and evaluations of the AI product-brand-label combination and would make evaluative conditioning effects on brand evaluations more likely to be observed in the following.

Interestingly, I further observed evidence against an impact of advertisement statements on the perception and evaluation of AI products both at baseline as well as across AI trustworthiness levels. The advertisement statements I used can be considered as promotion messages which have been shown to be effective in positively affecting the evaluation of hedonic products and the willingness to pay for them (e.g., [42]; see e.g., [13, 46], for general reviews on the effect of advertisement). As I ensured that participants viewed AI products for at least 4s before evaluating them, I can also be fairly certain that they processed the advertisement statements. Nevertheless, the advertisement statements I used portrayed benefits of the two types of AI products in a generic way rather than presenting specifics on the potential benefits of an AI product (e.g., statistical claims: advertisement statements using concrete data on how much time can be saved or by how much productivity can be increased; see e.g., [37], for evidence on statistical vs. narrative advertisement claims). Whereas the present study speaks against an impact of advertisement on the evaluation of the AI product types I used in general, further research needs to determine whether other types of advertisement claims (e.g., statistical claims like "increases your productivity by 87%") affect the evaluation of AI products more effectively. Once such an advertisement effect can be shown at baseline, it becomes possible to investigate to what degree AI trustworthiness labels guard against the influence of advertisement (see [26], for corresponding evidence regarding the Nutri-Score) as I intended.

Crucially, assessing trust, acceptance, and intention to use, I relied on baseline-adjusted measures for my main analyses. That is, values of 0 correspond to a person's rating at baseline (without an AI trustworthiness label and without advertisement statements). Values larger than 0 indicate more trust, acceptance, or intention to use than at baseline, whereas values below 0 indicate less trust, acceptance, or intention to use than at baseline. Replicating and extending the findings of Pfeuffer [47, 48], I found that, across dependent measures, intermediate trustworthiness levels corresponded to ratings around 0, that is, to ratings at baseline without any AI trustworthiness label. It is noteworthy that I obtained this

finding even though I intermixed baseline conditions with the three AI trustworthiness levels in comparison to Pfeuffer [47, 48] who first assessed AI product ratings at baseline before manipulating AI trustworthiness (data security and data privacy, specifically). One might argue that participants tended towards mid-scale ratings at baseline in the study of Pfeuffer [47, 48], as they had not received any information to evaluated AI products based on when providing baseline ratings. However, this criticism should not apply in the present study where unlabeled AI products were presented intermixed with AI products with an AI trustworthiness label, directly contrasting baseline and labeled trustworthiness conditions. That participants nonetheless defaulted to baseline ratings that aligned with ratings they provided for intermediate trustworthiness levels (even though there was substantial variance in baseline ratings themselves), points towards common biases in the perception and evaluation of AI products. It appears that, in the absence of relevant information, (potential) users attribute an intermediate level of trustworthiness to any AI product (see also the corresponding discussion in [47, 48]). This puts them at risk to show either inaccurate/unjustified trust or inaccurate/unjustified distrust towards AI products which are both detrimental to the successful implementation of AI in society [7, 24, 49, 58]. In my opinion, especially this finding strongly calls for the implementation of obligatory multi-level AI trustworthiness labels for AI products in corresponding legal regulations concerning AI to support (potential) users of AI products in calibrating their AI trustworthiness assessment. It further illustrates a benefit of multi-level AI trustworthiness labels as compared to certification labels: Comparisons between baseline ratings without further information on AI products and AI product ratings at different AI trustworthiness levels constitute a new and promising method of assessing and tracking biases in the evaluation of AI trustworthiness.

However, please note that the need for AI trustworthiness labels to inform (potential) users' about an AI product's trustworthiness I underscore here based on the results of this study mainly stems from the presently prevalent lack of transparency and the difficulty of obtaining corresponding information regarding AI products. Ideally, (potential) users themselves should be empowered to evaluate the trustworthiness of AI products. Yet, given the present lack of transparency and obstructions to information access on AI products, (potential) users can hardly find trustworthiness cues (see e.g., [36, 52] for corresponding research) to conduct a trustworthiness assessment themselves. This is further complicated by the speed of advancements regarding AI products and a corresponding need to constantly keep up with current developments. Finally, users who already have access to an AI product typically gain access to at least some trustworthiness cues. However, the potential users who have not yet acquired a respective AI product this study focused on are even more limited in terms of the information available to them. Especially regarding these potential users' pre-usage AI trustworthiness assessment, I conclude that my findings highlight existing biases and misconceptions. Thus, AI trustworthiness labels issued by an organization, that is trustworthy itself, can support AI trustworthiness assessment in the absence of more detailed information especially for yet undecided, potential users of AI products. Nonetheless, the existence of AI trustworthiness labels would and should never be a reason to stop advocating for more transparent documentation and communication regarding AI products.

Furthermore, it is conceivable that the correct interpretation of trustworthiness cues in AI products also depends on (potential) users' AI literacy. Importantly, I further found evidence that interindividual differences in a person's general, product-unspecific trust in AI and differences in their AI literacy modulated the impact of AI trustworthiness levels conveyed via corresponding labels on participants' perception and evaluation of AI products. The more trusting and literate regarding AI participants were, the more positively they rated their trust, acceptance, and intention to use corresponding AI products with a high trustworthiness level and the more negatively they rated their trust, acceptance, and intention to use corresponding AI products with a low trustworthiness level. That is, AI trustworthiness labels yielded stronger effects on a potential user's perception and evaluation of an AI product when they were generally more trusting and when they experienced themselves as relatively highly literate regarding AI. These results extend prior findings suggesting that higher AI literacy is associated with higher trust in AI and more frequent use of AI ([25, 30]) and present a more nuanced picture. That is, AI literacy only seems to be associated with more positive perceptions of AI like increased trust when AI trustworthiness is evaluated as relatively high. Conversely, however, perceptions and evaluations of AI products labeled as untrustworthy become even more negative for persons with higher AI literacy. At present, these findings have to be interpreted cautiously as a contribution of floor effects to the pattern of results cannot be ruled out. Nevertheless, especially the findings regarding the impact of AI literacy are of interest as they might suggest that potential users might need at least a certain minimum knowledge/understanding of AI to adequately benefit from trustworthiness information conveyed by AI trustworthiness labels (or other trustworthiness cues; see e.g., [36, 52]). This hypothesis should be further investigated in future studies. If it were confirmed, it would advocate for both a need of introducing AI trustworthiness labels as well as measures to increase AI literacy to ensure that the trustworthiness information conveyed by such labels (as well as potentially other trustworthiness cues) yields a strong impact on (potential) users' perception and evaluation of AI products.

Note that the present study which used trustworthiness criteria established within the scope of the EU legislation on AI ([15, 16]) focused on German-speaking EU citizens living in Germany and Austria that is, persons who are also affected by corresponding EU AI legislation. At present, I cannot determine whether my findings regarding the impact of AI trustworthiness labels focusing on criteria particularly discussed in the EU can be extended to, for instance, non-EU citizens who might be used to diverging local AI legislation. The present study communicated trustworthiness criteria that vastly differed from those used in the study of Pfeuffer [47, 48] (which focused on data security and data privacy inspired by a Swiss labelling suggestion [55]). Nonetheless, it replicated the impact of AI trustworthiness label levels on potential users' trust. This tentatively suggests that AI trustworthiness labels yield an impact irrespective of their exact criteria. This might tentatively indicate that corresponding impacts of AI trustworthiness labels

also replicate in other populations. Yet, this first needs to be confirmed empirically. One especially noteworthy population future studies should focus on are actual active users of non-hypothetical AI products and changes in their perceptions and evaluations of AI they are interacting with. These studies should investigate active users' perceptions and evaluations of AI depending on which AI trustworthiness information (e.g., low-high trustworthiness as indicated by an AI trustworthiness label) they initially received on the corresponding AI. Furthermore, my present sample, for instance, mainly consisted of participants who reached high educational levels and therefore allows only tentative conclusions for participants of lower educational levels. Future studies need to reassess participant groups underrepresented in the present sample before clear conclusions can be drawn for them.

Moreover, the present study did not yet consider a multi-layered/multi-agent perspective of the trustee (i.e., the AI product). That is, users' trustworthiness assessments and corresponding trust may be derived based on the AI's brand and or the platform via which it is provided (e.g., Meta, OpenAI, etc.) as well as the content or services delivered by the AI (e.g., recommendations, decisions, outputs), or the contributions of other system users (e.g., training data or reviews; see e.g., [52], see [59] for a corresponding multi-layered trust/trustworthiness perspective in the context of social media usage). Future studies should additionally implement a multi-agent perspective of the trustee. This is of particular interest when comparing actual active users of an AI system and its potential users who do not yet have accessed the AI system and, thus, for instance, have not yet actively experienced the AI system's content and services.

6 Conclusion

Taken together, I show that AI trustworthiness conveyed via corresponding, graphical labels successfully communicates AI trustworthiness information and scales potential users' perception and evaluation of AI extending from trust to acceptance and intention to use. Potential users are at risk for misplacing trust as well as distrust in AI products, as they appear to default to attributing intermediate trustworthiness to AI products they have no further information about. This highlights the need to communicate AI trustworthiness to potential users in an easy-to-process format, for instance, by using graphical AI trustworthiness labels. Finally, my findings suggest that individual trust in AI irrespective of specific AI products and AI literacy may influence the impact of such AI trustworthiness labels and call for further corresponding research.

Acknowledgments

I thank Deborah Werner for her help with preparing the study materials.

References

[1] Martin Adam, University of Goettingen, Sebastian Lins, Karlsruhe Institute of Technology, Ali Sunyaev, Karlsruhe Institute of Technology, Alexander Benlian, and Technical University of Darmstadt. 2024. The Contingent Effects of IS Certifications on the Trustworthiness of Websites. JAIS 25, 3 (2024), 594–617. https://doi.org/10.17705/1jais.00836

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, May 02, 2019. ACM, Glasgow Scotland Uk, 1–13. https://doi.org/10.1145/3290605.3300233

[3] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. Proc. ACM Hum.-Comput. Interact. 5, CSCW1 (April 2021), 1–21. https://doi.org/10.1145/3449287

[4] George J. Cancro, Shimei Pan, and James Foulds. 2022. Tell Me Something That Will Help Me Trust You: A Survey of Trust Calibration in Human-Agent Interaction. https://doi.org/10.48550/arXiv.2205.02987

[5] Astrid Carolus, Martin J. Koch, Samantha Straka, Marc Erich Latoschik, and Carolin Wienrich. 2023. MAILS - Meta AI literacy scale: Development and testing of an AI literacy questionnaire based on well-founded competency models and psychological change- and meta-competencies. Computers in Human Behavior: Artificial Humans 1, 2 (August 2023), 100014. https://doi.org/10.1016/j.chbah.2023.100014

[6] Peter Caven, Zitao Zhang, Jacob Abbott, Xinyao Ma, and Ljean Camp. 2024. Comparing the Use and Usefulness of Four IoT Security Labels. In Proceedings of the CHI Conference on Human Factors in Computing Systems, May 11, 2024. ACM, Honolulu HI USA, 1–31. https://doi.org/10.1145/3613904.3642951

[7] Hyesun Choung, Prabu David, and Arun Ross. 2023. Trust in AI and Its Role in the Acceptance of AI Technologies. International Journal of Human–Computer Interaction 39, 9 (May 2023), 1727–1739. https://doi.org/10.1080/10447318.2022.2050543

[8] Fred D. Davis. 1989. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. MIS Quarterly 13, 3 (September 1989), 319. https://doi.org/10.2307/249008

[9] Jan De Houwer, Sarah Thomas, and Frank Baeyens. 2001. Association learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. Psychological Bulletin 127, 6 (2001), 853–869. https://doi.org/10.1037/0033-2909.127.6.853

[10] Ewart J. De Visser, Marvin Cohen, Amos Freedy, and Raja Parasuraman. 2014. A Design Methodology for Trust Cue Calibration in Cognitive Agents. In Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments, Randall Shumaker and Stephanie Lackey (eds.). Springer International Publishing, Cham, 251–262. https://doi.org/10.1007/978-3-319-07458-0_24

[11] Ewart J. De Visser, Marieke M. M. Peeters, Malte F. Jung, Spencer Kohn, Tyler H. Shaw, Richard Pak, and Mark A. Neerincx. 2020. Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. Int J of Soc Robotics 12, 2 (May 2020), 459–478. https://doi.org/10.1007/s12369-019-00596-x

[12] John Dorsch and Ophelia Deroy. 2025. The impact of labeling automotive AI as trustworthy or reliable on user evaluation and technology acceptance. Sci Rep 15, 1 (January 2025), 1481. https://doi.org/10.1038/s41598-025-85558-2

[13] Martin Eisend and Farid Tarrahi. 2016. The Effectiveness of Advertising: A Meta-Meta-Analysis of Advertising Inputs and Outcomes. Journal of Advertising 45, 4 (October 2016), 519–531. https://doi.org/10.1080/00913367.2016.1185981

[14] Pardis Emami-Naeini, Henry Dixon, Yuvraj Agarwal, and Lorrie Faith Cranor. 2019. Exploring How Privacy and Security Factor into IoT Device Purchase Behavior. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, May 02, 2019. ACM, Glasgow Scotland Uk, 1–12. https://doi.org/10.1145/3290605.3300764

[15] European Commission. Directorate General for Communications Networks, Content and Technology. and High Level Expert Group on Artificial Intelligence. 2019. Ethics guidelines for trustworthy AI. Publications Office, LU. Retrieved February 28, 2025 from https://data.europa.eu/doi/10.2759/346720

[16] European Parliament and Council of the European Union. Regulation (EU) 2022/206 of 21 April 2022 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri$=$CELEX%3A32022R0206

[17] Raphael Fischer, Magdalena Wischnewski, Alexander van der Staay, Katharina Poitz, Christian Janiesch, and Thomas Liebig. 2025. Bridging the Communication Gap: Evaluating AI Labeling Practices for Trustworthy AI Development. https://doi.org/10.48550/ARXIV.2501.11909

[18] George Fragiadakis, Christos Diou, George Kousiouris, and Mara Nikolaidou. 2024. Evaluating Human-AI Collaboration: A Review and Methodological Framework. https://doi.org/10.48550/arXiv.2407.19098

[19] Nicole Gillespie, Steven Lockey, Caitlin Curtis, Javad Pool, and Ali Akbari. 2023. Trust in Artificial Intelligence: A global study. The University of Queensland; KPMG Australia, Brisbane, Australia. https://doi.org/10.14264/00d3c94

[20] Danny S. Guamán, Manel Medina, Pablo López-Aguilar, Hristina Veljanova, José M. Del Álamo, Valentin Gibello, Martin Griesbacher, and Ali Anjomshoaa. 2022. TRUESSEC Trustworthiness Label Recommendations. In Challenges in Cybersecurity and Privacy - the European Research Landscape (1st ed.). River Publishers, New York, 207–230. https://doi.org/10.1201/9781003337492-10

[21] David Harborth and Sebastian Pape. 2018. German Translation of the Unified Theory of Acceptance and Use of Technology 2 (UTAUT2) Questionnaire. SSRN Journal (2018). https://doi.org/10.2139/ssrn.3147708

[22] David Hartmann, José Renato Laranjeira De Pereira, Chiara Streitbörger, and Bettina Berendt. 2024. Addressing the regulatory gap: moving towards an EU AI audit ecosystem beyond the AI Act by including civil society. AI Ethics (November 2024). https://doi.org/10.1007/s43681-024-00595-3

[23] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. Hum Factors 57, 3 (May 2015), 407–434. https://doi.org/10.1177/0018720814547570

[24] Joo-Wha Hong. 2021. Artificial intelligence (AI), don't surprise me and stay in your lane: An experimental testing of perceiving humanlike performances of AI. Human Behav and Emerg Tech 3, 5 (December 2021), 1023–1032. https://doi.org/10.1002/hbe2.292

[25] Kuo-Ting Huang and Christopher Ball. 2024. The Influence of AI Literacy on User's Trust in AI in Practical Scenarios: A Digital Divide Pilot Study. Proceedings of the Association for Information Science and Technology 61, 1 (October 2024), 937–939. https://doi.org/10.1002/pra2.1146

[26] Kristin Jürkenbeck, Clara Mehlhose, and Anke Zühlsdorf. 2022. The influence of the Nutri-Score on the perceived healthiness of foods labelled with a nutrition claim of sugar. PLoS ONE 17, 8 (August 2022), e0272220. https://doi.org/10.1371/journal.pone.0272220

[27] Maria Kasinidou. 2023. Promoting AI Literacy for the Public. In Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 2, March 2023. ACM, Toronto ON Canada, 1237–1237. https://doi.org/10.1145/3545947.3573292

[28] Moritz Körber. 2019. Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation. In Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018), Sebastiano Bagnara, Riccardo Tartaglia, Sara Albolino, Thomas Alexander and Yushi Fujita (eds.). Springer International Publishing, Cham, 13–30. https://doi.org/10.1007/978-3-319-96074-6_2

[29] Adriano Koshiyama, Emre Kazim, Philip Treleaven, Pete Rai, Lukasz Szpruch, Giles Pavey, Ghazi Ahamat, Franziska Leutner, Randy Goebel, Andrew Knight, Janet Adams, Christina Hitrova, Jeremy Barnett, Parashkev Nachev, David Barber, Tomas Chamorro-Premuzic, Konstantin Klemmer, Miro Gregorovic, Shakeel Khan, Elizabeth Lomas, Airlie Hilliard, and Siddhant Chatterjee. 2024. Towards algorithm auditing: managing legal, ethical and technological risks of AI, ML and associated algorithms. R. Soc. Open Sci. 11, 5 (May 2024), 230859. https://doi.org/10.1098/rsos.230859

[30] Esther S. Kox and Beatrice Beretta. 2024. Evaluating Generative AI Incidents: An Exploratory Vignette Study on the Role of Trust, Attitude and AI Literacy. In Frontiers in Artificial Intelligence and Applications, Fabian Lorig, Jason Tucker, Adam Dahlgren Lindström, Frank Dignum, Pradeep Murukannaiah, Andreas Theodorou and Pınar Yolum (eds.). IOS Press. https://doi.org/10.3233/FAIA240194

[31] Frank Krueger, René Riedl, Jennifer A. Bartz, Karen S. Cook, David Gefen, Peter A. Hancock, Sirkka L. Jarvenpaa, Lydia Krabbendam, Mary R. Lee, Roger C. Mayer, Alexandra Mislin, Gernot R. Müller-Putz, Thomas Simpson, Haruto Takagishi, and Paul A. M. Van Lange. 2025. A call for transdisciplinary trust research in the artificial intelligence era. Humanit Soc Sci Commun 12, 1 (July 2025), 1124. https://doi.org/10.1057/s41599-025-05481-9

[32] Sabina Lacmanovic and Marinko Skare. 2025. Artificial intelligence bias auditing – current approaches, challenges and lessons from practice. RAF 24, 3 (May 2025), 375–400. https://doi.org/10.1108/RAF-01-2025-0006

[33] Joakim Laine, Matti Minkkinen, and Matti Mäntymäki. 2024. Ethics-based AI auditing: A systematic literature review on conceptualizations of ethical principles and knowledge contributions to stakeholders. Information & Management 61, 5 (July 2024), 103969. https://doi.org/10.1016/j.im.2024.103969

[34] J. D. Lee and K. A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. Human Factors: The Journal of the Human Factors and Ergonomics Society 46, 1 (January 2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[35] Fan Li and Ya Yang. 2024. Impact of Artificial Intelligence–Generated Content Labels On Perceived Accuracy, Message Credibility, and Sharing Intentions for Misinformation: Web-Based, Randomized, Controlled Experiment. JMIR Form Res 8, (December 2024), e60024. https://doi.org/10.2196/60024

[36] Q. Vera 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, June 21, 2022. 1257–1268. https://doi.org/10.1145/3531146.3533182

[37] Hung-Chou Lin and Sheng-Hsien Lee. 2021. Effects of Statistical and Narrative Health Claims on Consumer Food Product Evaluation. Front. Psychol. 11, (January 2021), 541716. https://doi.org/10.3389/fpsyg.2020.541716

[38] Zhuoran Lu and Ming Yin. 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, May 06, 2021. ACM, Yokohama Japan, 1–16. https://doi.org/10.1145/3411764.3445562

[39] Nikola Marangunić and Andrina Granić. 2015. Technology acceptance model: a literature review from 1986 to 2013. Univ Access Inf Soc 14, 1 (March 2015), 81–95. https://doi.org/10.1007/s10209-014-0348-1

[40] Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An Integrative Model of Organizational Trust. The Academy of Management Review 20, 3 (July 1995), 709. https://doi.org/10.2307/258792

[41] Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An Integrative Model of Organizational Trust. The Academy of Management Review 20, 3 (July 1995), 709. https://doi.org/10.2307/258792

[42] Camelia C. Micu and Tilottama G. Chowdhury. 2010. The Effect of Message's Regulatory Focus and Product Type on Persuasion. Journal of Marketing Theory and Practice 18, 2 (April 2010), 181–190. https://doi.org/10.2753/MTP1069-6679180206

[43] Tal Moran, Yahel Nudler, and Yoav Bar-Anan. 2023. Evaluative Conditioning: Past, Present, and Future. Annu. Rev. Psychol. 74, 1 (January 2023), 245–269. https://doi.org/10.1146/annurev-psych-032420-031815

[44] Boris New, Jessica Bourgin, Julien Barra, and Christophe Pallier. 2024. UniPseudo: A universal pseudoword generator. Quarterly Journal of Experimental Psychology 77, 2 (February 2024), 278–286. https://doi.org/10.1177/17470218231164373

[45] Mauricio Palmeira, Jing Lei, and Ana Valenzuela. 2019. Impact of vertical line extensions on brand attitudes and new extensions: The roles of judgment focus, comparative set and positioning. European Journal of Marketing 53, 2 (February 2019), 299–319. https://doi.org/10.1108/EJM-07-2017-0431

[46] Richard E. Petty and John T. Cacioppo. 1986. The Elaboration Likelihood Model of Persuasion. In Advances in Experimental Social Psychology. Elsevier, 123–205. https://doi.org/10.1016/S0065-2601(08)60214-2

[47] Christina U Pfeuffer. 2025. Instilling (Dis-)Trust in AI Products: Recommendations for the Design of Data Security and Data Privacy Labels. In Proceedings of the 2025 HAI Conference on Human-Agent Interaction, 2025. ACM, 1–3. https://doi.org/10.1145/3765766.3765868

[48] Christina U. Pfeuffer. 2025. Instilling (Dis-)Trust in AI Products: Recommendations for the Design of Data Security and Data Privacy Labels. https://doi.org/10.31234/osf.io/q25nr_v1

[49] Minjin (Mj) Rheu, Yue (Nancy) Dai, Jingbo Meng, and Wei Peng. 2024. When a Chatbot Disappoints You: Expectancy Violation in Human-Chatbot Interaction in a Social Support Context. Communication Research 51, 7 (October 2024), 782–814. https://doi.org/10.1177/00936502231221669

[50] Lisa Schadelbauer, Stephan Schlögl, and Aleksander Groth. 2023. Linking Personality and Trust in Intelligent Virtual Assistants. MTI 7, 6 (May 2023), 54. https://doi.org/10.3390/mti7060054

[51] 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, June 12, 2023. ACM, Chicago IL USA, 248–260. https://doi.org/10.1145/3593013.3593994

[52] Nadine Schlicker, Kevin Baum, Alarith Uhde, Sarah Sterz, Martin C. Hirsch, and Markus Langer. 2025. How do we assess the trustworthiness of AI? Introducing the trustworthiness assessment model (TrAM). Computers in Human Behavior 170, (September 2025), 108671. https://doi.org/10.1016/j.chb.2025.108671

[53] Felix D. Schönbrodt, Eric-Jan Wagenmakers, Michael Zehetleitner, and Marco Perugini. 2017. Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. Psychological Methods 22, 2 (June 2017), 322–339. https://doi.org/10.1037/met0000061

[54] Elnora W. Stuart, Terence A. Shimp, and Randall W. Engle. 1987. Classical Conditioning of Consumer Attitudes: Four Experiments in an Advertising Context. J CONSUM RES 14, 3 (December 1987), 334. https://doi.org/10.1086/209117

[55] Swiss Digital Initiative. 2022. Retrieved February 12, 2023 from https://digitaltrust-label.swiss/criteria/

[56] Venkatesh, Morris, Davis, and Davis. 2003. User Acceptance of Information Technology: Toward a Unified View. MIS Quarterly 27, 3 (2003), 425. https://doi.org/10.2307/30036540

[57] Venkatesh, Thong, and Xu. 2012. Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology. MIS Quarterly 36, 1 (2012), 157. https://doi.org/10.2307/41410412

[58] E. S. Vorm and David J. Y. Combs. 2022. Integrating Transparency, Trust, and Acceptance: The Intelligent Systems Technology Acceptance Model (ISTAM). International Journal of Human–Computer Interaction 38, 18–20 (December 2022), 1828–1845. https://doi.org/10.1080/10447318.2022.2070107

[59] Yixuan Zhang, Joseph D Gaggiano, Nutchanon Yongsatianchot, Nurul M Suhaimi, Miso Kim, Yifan Sun, Jacqueline Griffin, and Andrea G Parker. 2023. What Do We Mean When We Talk about Trust in Social Media? A Systematic Review. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, April 19, 2023. ACM, Hamburg Germany, 1–22. https://doi.org/10.1145/3544548.3581019