Beyond Black Boxes:
AI Certification Labels Increase Adoption through Trust

Filipa de Almeida
Universidade Católica Portuguesa, Católica Lisbon School of Business & Economics, Católica Lisbon Research Unit in Business and Economics, Lisboa,Portugal
[0000-0002-3112-3965] filipadealmeida@ucp.pt

Joana Glaum
1KOMMA5° GmbH, Hamburg, Germany

Kiall Hildred
Universidade Católica Portuguesa, Católica Lisbon School of Business & Economics, Católica Lisbon Research Unit in Business and Economics, Lisboa,Portugal
[0000-0002-7076-0422]

Ian J. Scott
NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal
[0000-0001-9699-4473]

This is the final Author Peer Reviewed version of:

Almeida, F. D., Glaum, J., & Scott, I. (2026). Beyond Black Boxes: AI Certification Labels Increase Adoption through Trust. Psychology and Marketing. https://doi.org/10.1002/mar.70185

Funding/Acknowledgments: We acknowledge the support from FCT – Portuguese Foundation of Science and Technology for the project, under the projects UID/00407/2025 and UID/04152/2025 - Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS - https://doi.org/10.54499/UID/04152/2025 (2025-01-01/2028-12-31) and UID/PRR/04152/2025 https://doi.org/10.54499/UID/PRR/04152/2025 (2025-01-01/ 2026-06-30).

This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions. This article may not be enhanced, enriched or otherwise transformed into a derivative work, without express permission from Wiley or by statutory rights under applicable legislation. Copyright notices must not be removed, obscured or modified. The article must be linked to Wiley's version of record on Wiley Online Library and any embedding, framing or otherwise making available the article or pages thereof by third parties from platforms, services and websites other than Wiley Online Library must be prohibited.

Beyond Black Boxes: AI Certification Labels Increase Adoption through Trust

Author accepted manuscript for: de Almeida, F., Glaum, J., Hildred, K., & Scott, I. J. (2026). ***Beyond black boxes: AI certification labels increase adoption through trust*****.** ***Psychology & Marketing*****.**

Acknowledgements

We acknowledge the support from FCT – Portuguese Foundation of Science and Technology for the project, under the projects UID/00407/2025 and UID/04152/2025 - Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS - https://doi.org/10.54499/UID/04152/2025 (2025-01-01/2028-12-31) and UID/PRR/04152/2025 https://doi.org/10.54499/UID/PRR/04152/2025 (2025-01-01/ 2026-06-30).

Data availability statement

The data that support the findings of these studies are available at https://osf.io/6zbhn/overview?view_only=d0b4d45d953242c1aa9f83ac2ef64bcd

Beyond Black Boxes: AI Certification Labels Increase Adoption through Trust

Abstract

Artificial intelligence (AI) is being increasingly used in daily life. Using these systems depends on end-users' trust in them, and auditing processes have been proposed to determine and communicate whether an AI system is trustworthy. However, many end-users lack the expertise to both assess the trustworthiness of AI systems and to understand the outputs of auditing. Certification labels have been proposed as a non-technical solution to communicating AI trustworthiness to end-users. This research tests the effectiveness of Trustworthy AI certification labels on end-users' trust in and behavioural intention to use (BIU) AI. In three studies, we show that using a simple certification label increases cognitive and affective trust in AI and BIU it. Moreover: cognitive trust mediates the label's effect; labels displaying Trustworthy AI principles further increase perceived trustworthiness and BIU; and exposure to labelled AI systems reduces trust in and BIU non-labelled systems. This work contributes to literature on trust and acceptance of AI and on the effectiveness of certification labels for signalling trustworthiness, and offers insights for the design of effective AI certification labels.

Keywords: artificial intelligence; trustworthy AI; auditing; certification labels; technology acceptance

1 Introduction

Artificial intelligence (AI) is rapidly becoming a central part of our daily lives (Haque & Li, 2024). And yet, how AI systems work is often a "Black-Box" (Mersha et al., 2024), leading to concerns about their trustworthiness (Crockett et al., 2021). While these concerns reflect legitimate risks, they have also increased distrust towards AI (Crockett et al., 2021), hindering its full potential (Lukyanenko et al., 2022). Therefore, building trust in AI has become a top priority on political, commercial, and legal agendas (AI HLEG, 2019).

Achieving widespread trust requires AI systems to be trustworthy (Kaur et al., 2022) and to communicate that to end-users (Laine et al., 2024; Liao & Sundar, 2022). However, as not all end-users have the required understanding to assess trustworthiness in AI systems (Mersha et al., 2024), many researchers (e.g., AI HLEG, 2019; Cihon et al., 2021; Scharowski et al., 2023; Stuurman & Lachaud, 2022) have suggested certification as a non-technical solution.

However, while certification labels are widely used, such as "fairtrade" or "organic" (Cornudet et al., 2025; Mazzù et al., 2023), research on the effectiveness of AI certification labels is limited. Furthermore, as trust and acceptance of AI can be highly context-dependent (Zehnle et al., 2025), the present research addresses the need for a comprehensive, experimental model connecting the design of certification labels to AI acceptance through both cognitive and affective trust, while accounting for both high- and low-stakes contexts and possible spillover effects. This holistic approach extends earlier research focused on simpler models of the relationships between these factors (e.g., Jacovi et al., 2021; Lee & See, 2004; Liao & Sundar, 2022; Scharowski et al., 2023). Specifically, the present research seeks to answer the following questions:

RQ1: Can certification labels increase end-users' trust and acceptance of AI?

RQ2: What is the effect of the level of information in such labels?

RQ3: How may the application of such labels affect unlabelled AI systems?

To address these questions, we followed a quantitative empirical approach, conducting three experiments testing how the presence of certification labels in AI-use scenarios affects participant's trust and intention to use AI. Moreover, we manipulate the information present on the labels and we manipulate the presence of a label on one AI system to assess how it impacts the trust and intention to use unlabelled AI systems.

This research contributes to literature on the importance of trust in AI and AI acceptance by showing that a simple TAI certification label increases cognitive and affective trust and behavioural intention to use AI; that cognitive trust is the primary mediating mechanism; that labels explicitly displaying TAI principles produce stronger effects than minimal labels; that these effects hold across both high- and low-stakes contexts; and that exposure to a labelled system reduces cognitive trust in unlabelled systems, with implications for AI market dynamics and certification policy and label design.

2 Literature Review and Hypotheses Development

2.1 Trust, Trustworthiness, and AI Acceptance

Trust has been defined as "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" (Mayer et al., 1995, p. 712). Accordingly, trust depends on the environment, outcomes, and interactions between trusting parties (Bhattacharya et al., 1998).

An important factor in whether trust occurs is the trustworthiness of the trustee, which refers to the perceived set of characteristics that justify an individual's willingness to trust them (Colquitt et al., 2007; Mayer et al., 1995). Mayer et al. (1995) conceptualised trustworthiness as comprising three dimensions, which each map onto distinct evaluative judgments: ability, the skills and competence to perform a task; benevolence, a caring and

thoughtful manner characterized by goodwill; and integrity, adherence to a set of mutually acceptable principles for behaviour (ABI). This ABI model has demonstrated its versatility across a range of disciplines (Afroogh et al., 2024; Hoff & Bashir, 2015), is often employed to determine the trustworthiness of technologies (Afroogh et al., 2024; McKnight et al., 2011; Söllner et al., 2016), and can be mapped onto the cognitive and affective dimensions of trust (Tomlinson et al., 2020) introduced by Johnson and Grayson (2005).

In the context of AI, an applicable conceptualisation of trust is as a belief formed through the evaluation of certain attributes of an object (Jacovi et al., 2021; Lee & See, 2004; Li et al., 2025). Specifically, cognitive trust, being knowledge-driven, would be based on a user's rational evaluation of an AI system's competence and reliability (ability), whereas affective trust would be based on an emotional bond between the user and the system (benevolence and integrity; Chen & Sundar, 2023). Furthermore, trust links to trustworthiness by comprising an individual's pre-existing views about AI systems with how they interpret aspects designed to signal the trustworthiness of those systems (Lee & See, 2004; Vanneste & Puranam, 2025).

Indeed, trust has been recognised as a key factor in AI acceptance (Li et al., 2025; Mersha et al., 2024; Söllner et al., 2016), leading researchers to include it as a predictor of AI acceptance within extended versions of Davis's (1985) Technology Acceptance Model (TAM; e.g., Alalwan et al., 2018; Wu et al., 2011), Venkatesh et al.'s (2003) Unified Theory of Acceptance and Use of Technology (UTAUT; e.g., Slade et al., 2015; Hooda et al., 2022), and Gursoy et al.'s (2019) Artificially Intelligent Device Use Acceptance model (AIDUA; e.g., Chi et al., 2023). However, acceptance of an AI system is not automatically implied by trust, but is determined by the trustor's subjective probability of a particular behaviour: their behavioural intention to use (BIU; Fishbein & Ajzen, 1975).

With this in mind, one path to establishing trust in AI involves working on factors that determine and signal its trustworthiness, while recognizing the crucial distinction between actual and perceived trustworthiness (Lee & See, 2004). The Black-Box nature of AI systems means that end-users cannot assess the former, but must instead rely on the latter, and hence the challenge for researchers, engineers, and designers is to match the end-user's trust and the trustworthiness they perceive to the actual trustworthiness of the systems (Lee & See, 2004; Glikson & Wooley, 2020).

This challenge has led to research into Trustworthy AI (TAI), focused on ensuring that trustworthiness is embedded into its development, deployment, and utilization (AI HLEG, 2019; Fukukawa & Trivedi, 2025; Jobin et al., 2019; Kaur et al., 2022), and on producing methods to implement these principles in practice (Mittelstadt, 2019; Papagiannidis et al., 2025). To this end, the European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG, 2019) proposed the following framework of requirements, which states that TAI must have three key characteristics (Lawful, Ethical, and Robust), safeguarding four essential human rights: respect for human control, prevention of harm, fairness, and explicability.

2.2 Communicating Trustworthiness of AI Systems to End-users

Auditing has been proposed as one way to translate the above principles into the realisation of TAI or the communication of trustworthiness. Auditing involves running tests to reveal problematic behaviour and uncover issues (Papagiannidis et al., 2025), leveraging expert knowledge to ensure TAI guidelines are met (Avin et al., 2021; Li & Goel, 2025) and to strengthen fairness (Schiff et al., 2024), accountability (Laine et al., 2024), and control (Falco et al., 2021). However, the outputs of auditing are primarily tailored to knowledgeable regulators and experts, making them ineffective in helping end-users make well-informed judgments about trusting or using AI (Kaur et al., 2022).

Certification labels have been proposed as a way to bridge this gap, as they signal that a product, procedure, individual, or organization conforms to predefined standards (ISO, 2020). The certification process can also involve auditing (King et al., 2005) conducted by a trusted third-party (Cihon et al., 2021), thereby leveraging experts to confirm the system's quality (Blösser & Weihrauch, 2024). This promotes transparency and motivates companies and designers to meet established standards (Cihon et al., 2021). The important advantage of certification labels over auditing alone is that they are designed to be easily understandable by many stakeholders (Sundar et al., 2021) and have been shown to enhance trust across a variety of goods, such as health ratings on food (Delepedra et al., 2025), social responsibility labelling on clothing (Cuesta-Valiño et al., 2024), and energy efficiency certifications on appliances (Schuitema et al., 2020). This suggests that certification labels are particularly appropriate for communicating TAI, meeting the challenge of matching the end-user's trust in an AI system, and the trustworthiness they perceive, to the actual trustworthiness of that system (Afroogh et al., 2024; Jacovi et al., 2021; Lee & See, 2004). Therefore, we propose the following hypothesis:

H1a. *A certification label for TAI increases trust in AI systems.*

However, research suggests that the effectiveness of certification labels in increasing end-users' cognitive and affective trust in AI systems may be impacted differently. For example, the findings of Schuitema et al. (2020) on energy efficiency labels suggest that they have a greater effect on cognitive trust than affective trust, which is supported by Chen and Sundar (2023), who showed that certification labels enhance users' trust by signalling data credibility. Because certification labels primarily communicate signals of competence, reliability, and procedural integrity, they are theoretically more aligned with cognition-based evaluations of trustworthiness than with the relational and emotional foundations characteristic of affective trust (Johnson & Grayson, 2005). Accordingly, we would expect

labelled AI systems to be evaluated predominantly on factors such as competence and reliability (Li et al., 2025), which Glikson and Woolley (2020) have shown is a key factor in developing cognitive trust. This expectation is further supported by the findings of Shamim et al. (2023), that the perceived reliability of AI has a strong effect on cognitive trust, and those of Xu et al. (2026), that cognitive trust plays a greater role than affective trust in overall trust in AI systems. Other findings support this by demonstrating that affective trust tends to be more important in decision-making in human-centred scenarios (Ferràs-Hernández, 2018), such as in business-to-business relationships, where affective trust is more important in the early phases than cognitive trust (Dowell et al., 2015). Based on these findings, we propose the following hypothesis:

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

As trust has been recognised as a key factor in AI acceptance (for reviews, see Glikson & Wooley, 2020; Li et al., 2025), which is generally determined by BIU (Fishbein & Ajzen, 1975; used in the present research to measure acceptance), we proposed three additional hypotheses:

H2. *A certification label for TAI increases behavioural intention to use AI systems.*

H3a. *Trust in AI (both cognitive and affective) predicts behavioural intention to use AI systems.*

H3b. *Trust in AI (both cognitive and affective) positively mediates the effect of a TAI certification label on behavioural intention to use AI systems.*

Furthermore, the technical nature of AI systems, and the fact that certification labels are intended to signal that a third party has confirmed the system's quality (Blösser & Weihrauch, 2024), suggest that their effectiveness may primarily leverage an individual's beliefs about the competence and reliability of the system, leading us to expect cognitive trust

to play a stronger mediating role than affective trust in the relationship between the certification label and BIU (Chen & Sundar, 2023), for which we propose the following hypothesis:

H3c. *Compared to affective trust, cognitive trust in AI is a stronger positive mediator of the effect of a TAI certification label on behavioural intention to use AI systems.*

As the certification label is only granted to AI systems that meet TAI criteria, its mere presence should signal trustworthiness. However, research has shown that adding more explicit information on similar kinds of certification labels increases trust and purchase intentions, such as on social responsibility labelling for clothing (Cuesta-Valiño et al., 2024), or nutrition labels (Mazzù et al., 2022) and health certifications on food (Sigurdsson et al., 2023). This suggests that explicitly stating TAI principles on a certification label may increase end-users' trust, leading us to propose an additional hypothesis:

H4. *A certification label that includes TAI requirements leads to higher trust (both cognitive and affective) in an AI system than a certification label that does not include TAI requirements.*

2.3 Trust and Trustworthiness Across High- and Low-Stakes Contexts

Many studies on trust, and on trust in AI specifically, focus on high-stakes scenarios by default (Afroogh et al., 2024). This is a reasonable intuition, as there is little value in trusting another entity if the outcomes of its decisions are inconsequential. However, a study by Johnson (2025) showed that a healthy distrust of AI decreased when the stakes were high, even if people spent more time making a decision, while Freel et al. (2025) showed that in low-stakes scenarios, subtle differences in the timing and severity of errors an AI makes can have a greater effect on trust levels. Moreover, Scharowski et al. (2023) found that individuals have a higher preference for AI certification labels in high-stakes scenarios than in low-stakes scenarios. Such findings suggest that how trust-related behaviours change

across the spectrum of high and low stakes is not a simple, one-dimensional function. Accordingly, the present research also aims to test how stakes affect the impact of certification labels, to which we propose the following hypothesis:

H5a. *Stakes moderate the effect of label condition on cognitive and affective trust in AI, such that differences between label conditions are more pronounced in high-stakes than in low-stakes contexts.*

H5b. *Stakes moderate the effect of label condition on behavioural intention to use AI, such that differences between label conditions are more pronounced in high-stakes than in low-stakes contexts.*

2.4 Negative Spillover Effects of Certification Labels on End-users' Trust in Non-Labelled Goods

While the advantages of certification labels have been well-documented (Chen & Sundar, 2023; Cuesta-Valiño et al., 2024; Schuitema et al., 2020), research has also shown that labelling can have both positive and negative spillover effects (Chen et al., 2023). Some early research showed that noticeable cues on products (such as on eco-labels) can create a negative contrast effect for unlabelled similar products (Sherif & Hovland, 1961), and more recent research has shown similar effects with new types of labels, such as sustainability labels (Anagnostou et al., 2015), suggesting that a certification label for one AI system may inadvertently signal the untrustworthiness of unlabelled systems. However, research into these effects is limited. Therefore, the present research aims to test whether participants exposed to a TAI-labelled system (versus not exposed) trust unlabelled systems less, to which we propose the following hypotheses:

H6a. *A labelled AI system leads to lower trust (both cognitive and affective) in an unlabelled AI system.*

Again, the importance of trust in AI acceptance, measured herein as BIU, leads us to propose two additional hypotheses:

H6b. *A labelled AI system leads to lower behavioural intention to use an unlabelled AI system.*

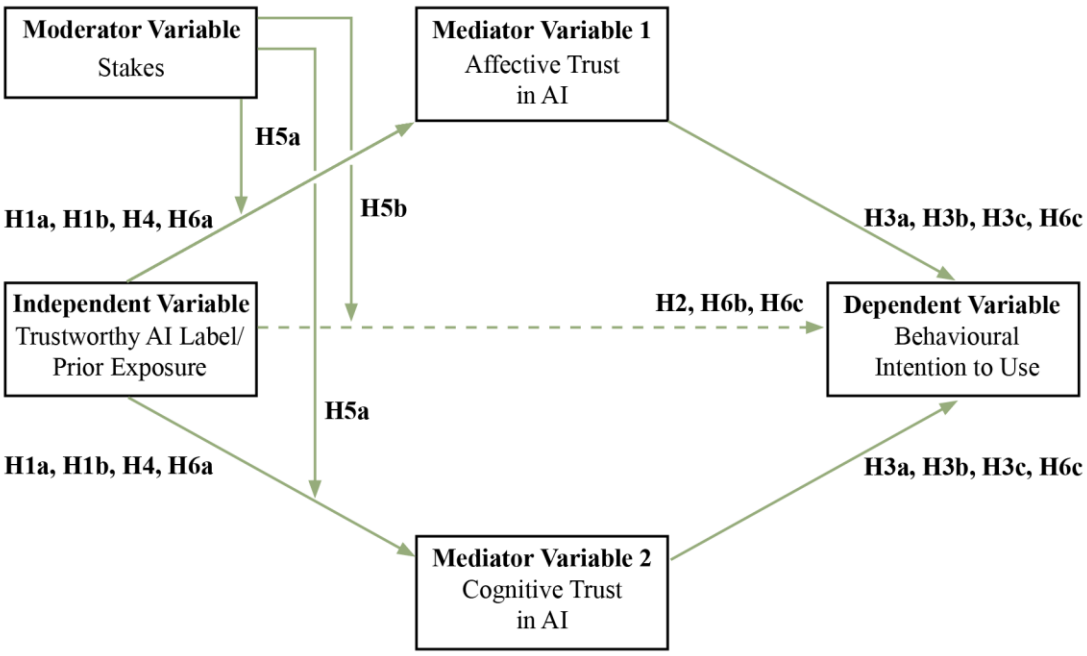
H6c. *A labelled AI system leads to lower behavioural intention to use an unlabelled AI system through both cognitive and affective trust.*

2.5 Overview of Studies and Conceptual Model

These hypotheses are tested across three studies. Study 1 tested the effect of certification labels on trust in and BIU AI, testing Hypotheses H1a–H3c. Study 2 replicated the findings from Study 1 and tested whether a complex label led to stronger effects than a simple label, testing Hypotheses H1a–H4. And Study 3 tested the effect of TAI-labelled AI systems on the perceived trustworthiness of non-labelled systems in both high- and low-stakes scenarios, testing Hypotheses H1a–H3c and H5–H6c. These studies aimed to determine the strength of the relationships in the following conceptual model (Figure 1).

Figure 1

Conceptual Model



3 Study 1

3.1 Participants

A convenience sample was drawn from social media and direct messaging. Participation was voluntary and no rewards were given for successful completion. The required sample size (128) was determined via G*Power (power at 0.8; medium effect size) for the mixed ANOVA and a Monte Carlo simulation for the parallel mediation. After elimination of participants (38) who indicated "1 = *Strongly agree*" or "2 = *Agree*" to the attention check ("*I have never used a smartphone*"), the sample included 128 participants (57.8% female, 39.8% male, 2.3% missing), ranging from 22 to 60 years (M = 27.68, SD = 5.87, N = 128). See Table 1 for more information. Participants reported having a moderately good understanding of AI (UAI; M = 5.30, SD = 1.07).

Sample Statistics, Study 1

Demographic	Value	*N*	%
Education	Secondary education	4	3.1%
	Bachelor's degree	64	50.0%
	Master's degree	50	39.1%
	Doctoral degree	1	0.8%
	Other	1	0.8%
	Missing	8	6.3%
Employment	Employed	59	46.1%
	Freelancer	4	3.1%
	Unemployed	1	0.8%
	Student	41	32.0%
	Worker and Student	20	15.6%
	Missing	3	2.3%
Country of Residence	Austria	2	1.6%
	Belgium	5	3.9%
	Denmark	4	3.1%
	France	1	0.8%
	Germany	87	68.0%
	Netherlands	2	1.6%
	New Zealand	1	0.8%
	Poland	2	1.6%
	Portugal	13	10.2%
	Switzerland	1	0.8%
	Missing	10	7.8%

3.2 Design

The study consisted of a between-subjects experiment with two Label conditions (Present vs. Absent) as the independent variable, BIU as the dependent variable, Affective Trust and Cognitive Trust as the mediators, and UAI as a covariate.

Informed consent from participants was obtained prior to all studies taking place. All procedures were performed in compliance with relevant laws and institutional guidelines and have been approved by the appropriate institutional committees[1].

3.3 Materials and Measures

3.3.1 AI System. The fictional AI system referred to in this study was drawn from the high-stakes scenarios used in Scharowski et al. (2023) and described as "an AI system called *MyJob* for evaluating job applications", which would "determine based on the provided information whether or not [the participant] will be invited for an interview." This scenario was used as the context of hiring and promotion decisions has been classified as high-risk by the European Commission's AI Act (2024). It is also a context in which the perceived trustworthiness of AI has been found to be low (Deriu et al., 2024). For full details of the survey and instructions, see Appendix A.

3.3.2 Certification Label for TAI (IV). The IV consisted of the presence or absence of the label show in Figure 2. The design and accompanying information were based on key success factors in AI regulation (Scharowski et al., 2023; Stuurman & Lachaud, 2022). In alignment with the suggestions of researchers and academic literature on labelling, an explanation of the meaning of the label was added to ensure its clarity and to prevent the potential for misinterpretation (Schebesta, 2019; see Appendix A for details).

[1] The studies were approved under the regulations of the Ethics committee of Nova IMS and MagiC Research Center, with the number reference OTHER2023-10-246486.

Figure 2

Simple Certification Label for TAI

3.3.3 Behavioural Intention to Use (DV). The DV was the participant's acceptance of the AI system as measured by their BIU (α = .89, 3 items; 1 = *Strongly disagree*, 7 = *Strongly agree*; for example, "*I will use AI systems like MyJob in the future.*"), adopted from relevant previous research (Castelo et al., 2019; Johnson & Grayson, 2005), with minor changes to relate it to the AI scenario.

3.3.4 Trust in AI (M). The trust measure was based on the ABI model used for trust in people (Mayer et al., 1995) and trust in technology (McKnight et al., 2011), mapped onto cognitive trust (ability; α = .94; 5 items; 1 = *Strongly disagree*, 7 = *Strongly agree*; for example, "*MyJob is competent.*") and affective trust (benevolence, integrity; α = .86, 6 items; 1 = *Strongly disagree*, 7 = *Strongly agree*; for example, "*MyJob cares about my well-being.*"), drawn from prior research (Castelo et al., 2019; Johnson & Grayson, 2005), with minor changes to relate it to the AI scenario.

3.3.5 Understanding of AI (CV). UAI was measured with a single item, "*How would you rate your understanding of Artificial Intelligence?*" (1 = *Extremely bad*, 7 = *Extremely good*). This was included as a covariate in the model as it has been shown to be a predictor of trust (Orbán & Stefkovics, 2025).

3.4 Procedure

Participants were first asked to rate their UAI, and then were randomly and evenly assigned to one of two label conditions (Present vs. Absent). In each, the participants were

asked to imagine themselves as a job applicant for a company using the *MyJob* AI system to evaluate applications, to which they would need to submit an application form, CV, and personal data. A short definition of AI was presented to ensure that all participants had an equal base-level understanding and thereby reduce variability in trust. Participants in the label present condition were also shown a TAI certification label (Figure 2) and accompanying information. All participants were then surveyed on their trust and BIU toward the AI system and asked an attention check question. Finally, participants were asked demographic questions and debriefed. See Appendix A for full survey details and Appendix B for details on the measures.

3.5 Results

A multiple mediation using Model 4 of PROCESS (v.5, Hayes, 2018; 95% CI, 5000 bootstrap replications) was conducted to test the mediating effects of affective and cognitive trust in the relationship between the certification label and BIU, with UAI included as a covariate. The analysis (Table 2) showed that presence of the label was a significant predictor of both affective trust (a_1) and cognitive trust (a_2), providing support for H1a. A regression analysis showed the difference between these effects to not be significant ($b = 0.10$, $p = .578$), providing no support for H1b. The direct effect of the label's presence on BIU was significant (c_1), providing support for H2. Also, both affective (b_1) and cognitive trust (b_2) were significant predictors of BIU, showing support for H3a.

Table 2

Direct Effects Between Label, Affective and Cognitive Trust, and BIU, controlling for UAI for Study 1

								95% CI	
Predictor	Outcome	*b*	*β*	*SE*	*df*	*t*	*p*	Lower	Upper
Label	Affective Trust (a_1)	0.84	0.36	0.20	124	4.29	<.001	0.45	1.23
Label	Cognitive Trust (a_2)	0.95	0.36	0.22	124	4.37	<.001	0.52	1.38
Affective Trust	BIU (b_1)	0.88	0.69	0.08	123	10.72	<.001	0.71	1.04

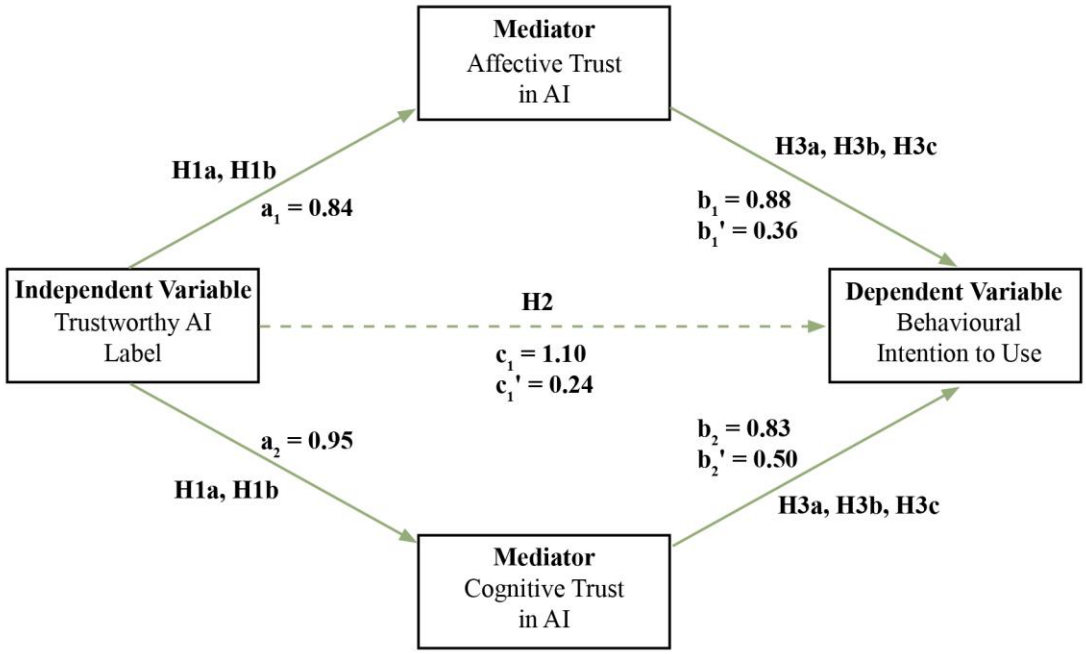
Cognitive Trust	BIU (b_2)	0.83	0.74	0.07	123	11.78	<.001	0.69	0.97
Label	BIU (c_1)	1.10	0.37	0.25	123	4.43	<.001	0.61	1.59

Note. One participant was excluded from the regression analyses due to missing data on the UAI covariate.

However, when affective and cognitive trust were included as mediators, the direct effect of the label's presence (c_1') was no longer significant ($b = 0.24$, $p = .196$), and the indirect effects were significant for both affective (b_1'), $b = 0.36$, *BootSE* = 0.13, 95% CI [0.14, 0.65] and cognitive trust (b_2'), $b = 0.50$, *BootSE* = 0.15, 95% CI [0.22, 0.83], indicating that they fully mediated the effect, supporting H3b, with a total effect of $b = 1.10$, $SE = 0.25$, $t = 4.43$, $p < .001$. The difference between these effects was not significant (95% CI [−0.54, 0.24]), providing no support for H3c. These results are depicted in Figure 3[2]. For the correlation tables, see Appendix C.

Figure 3

Parallel Mediation Model for Effects of Label, Study 1



4 Study 2

[2] UAI did not significantly predict affective trust ($p = .645$), cognitive trust ($p = .180$), or BIU ($p = .787$).

To test the possible strengthened effect from adding more explicit information on certification labels, as found in previous research (Cuesta-Valiño et al., 2024; Mazzù et al., 2022; Sigurdsson et al., 2023), we replicated Study 1 adding a label design that explicitly stated the TAI principles.

4.1 Participants

The sampling method and determination of required sample size were the same as in Study 1. After attention-check elimination (same as in Study 1;7 participants), the sample included 232 participants (64.2% female, 35.8% male), ranging from 19 to 60 years (M = 27.91, SD = 6.82). On average, participants rated themselves as having a moderately good understanding of AI (M = 5.48, SD = 0.91). See Table 3 for more information.

Table 3

Sample Statistics, Study 2

Demographic	Value	*N*	%
Education	Less than Secondary education	2	1.0%
	Secondary education	10	4.0%
	Bachelor's degree	114	49.0%
	Master's degree	92	40.0%
	Doctoral degree	10	4.0%
	Other	1	0.0%
	Not supplied	3	1.0%
Employment	Employed	110	47.0%
	Freelancer	9	4.0%
	Retired	1	0.0%
	Student	84	36.0%
	Unemployed	4	2.0%
	Worker and Student	24	10.0%
Country of Residence	Afghanistan	10	4.0%
	Albania	1	0.0%
	Australia	1	0.0%
	Austria	19	8.0%
	Belgium	7	3.0%
	Croatia	1	0.0%
	Czech Republic	2	1.0%
	Denmark	8	3.0%
	France	10	4.0%
	Germany	142	61.0%
	Hungary	1	0.0%
	Nepal	1	0.0%
	Netherlands	1	0.0%
	Norway	1	0.0%
	Not Provided	15	6.0%
	Portugal	10	4.0%
	United Kingdom	1	0.0%

Zimbabwe	1	0.0%

4.2 Design

The study consisted of a between-subjects experiment with three label conditions (No label vs. Simple vs. Complex) as the IV, BIU as the DV, Affective Trust and Cognitive Trust as the mediators, and UAI as a covariate.

4.3 Materials and Measures

This study measured Affective Trust (α = .94), Cognitive Trust (α = .97), BIU (α = .93), and UAI, as detailed in Study 1, and referred to the same fictional AI system (*MyJob*).

4.3.1 Certification Label for TAI (IV). The complex label (Figure 4, right) adapted the design from the simple label (Figure 4, left), adding the seven guidelines for TAI (AI HLEG, 2019).

Figure 4

Certification Labels Without (Simple) and with (Complex) Requirements for TAI

Simple Label *Complex Label*

4.4 Procedure

The procedure was identical to Study 1. However, participants were randomly and evenly assigned to one of the three label conditions, with the complex label (Figure 4, right) being presented to those participants in the Complex label condition.

4.5 Results

Model 4 of PROCESS (v.5, Hayes, 2018; 95% CI, 5000 bootstrap replications, N = 232) tested the mediating effects of affective and cognitive trust in the relationship between the different labels and BIU, with UAI included as a covariate. Compared to no label, both simple and complex labels had significant direct effects (Table 4) on affective trust and cognitive trust, providing further support for H1a, with both labels having stronger effects on cognitive trust than affective trust, providing support for H1b. Regression analysis showed that differences were significant for both the simple and complex labels[3].

Table 4

Regression Coefficients (and Differences) of Labels on Affective, Cognitive Trust, and BIU, Study 2

							95% CI	
Label	DV	*b*	*β*	*SE*	*t*	*p*	Lower	Upper
Simple	Affective (a_{3s})	1.53	1.37	0.13	11.57	<.001	1.27	1.79
	Cognitive (a_{4s})	2.06	1.47	0.15	13.30	<.001	1.75	2.36
	Difference	0.53	0.10	0.11	4.88	<.001	0.31	0.74
	BIU (c_{2s})	2.21	1.57	0.14	15.65	<.001	1.93	2.48
Complex	Affective (a_{3c})	1.97	1.50	0.14	14.09	<.001	1.69	2.24
	Cognitive (a_{4c})	2.63	1.63	0.15	17.57	<.001	2.33	2.92
	Difference	0.66	0.13	0.09	7.07	<.001	0.48	0.84
	BIU (c_{2c})	2.46	1.57	0.15	16.30	<.001	2.16	2.75

[3] UAI also had a significant positive effect on affective trust, b = 0.14, β = 0.11, SE = 0.06, t(228) = 2.39, p = .018, cognitive trust, b = 0.23, β = 0.14, SE = 0.07, t(228) = 3.35, p < .001, and behavioural intention, b = 0.27, β = 0.16, SE = 0.07, t(228) = 3.93, p < .001.

Direct effects of the label on BIU (Table 4) were also significant for both labels when compared to no label, providing support for H2. Also, both affective (b_3) and cognitive trust (b_4) were significant predictors of BIU, showing support for H3a (Table 5).

Table 5

Direct Effects of Trust on BIU, Study 2

							95% CI	
Label	Trust	*b*	*β*	*SE*	*t*	*p*	Lower	Upper
Simple	Affective (b_{3s})	1.08	0.86	0.05	20.18	<.001	0.97	1.18
	Cognitive (b_{4s})	0.90	0.90	0.04	25.33	<.001	0.83	0.97
Complex	Affective (b_{3c})	1.05	0.89	0.04	24.58	<.001	0.97	1.14
	Cognitive (b_{4c})	0.89	0.92	0.03	29.60	<.001	0.83	0.94
Total	Affective (b_3)	1.06	0.85	0.04	25.34	<.001	0.98	1.14
	Cognitive (b_4)	0.89	0.90	0.03	31.56	<.001	0.83	0.95

Comparing the mediation models for the two labels showed that, relative to no label, both had significant indirect effects on BIU through affective trust and cognitive trust (Table 6) supporting hypothesis H3b. The total effects were also significant (Table 6). An analysis of bootstrapped confidence intervals showed that the indirect effects of cognitive and affective trust did not significantly differ from each other for both the simple and complex labels, providing no support for H3c.

Table 6

Contrast of Specific Indirect Effects Using Percentile Bootstrapped CIs, Study 2

				95% CI	
Label	Trust	*b*	*BootSE*	Lower	Upper
Simple	Affective (b_{3s}')	0.49	0.18	0.13	0.85
	Cognitive (b_{4s}')	1.03	0.20	0.69	1.48
	Difference	−0.54	0.34	−1.29	0.06
	Total	1.52	0.17	1.22	1.87
Complex	Affective (b_{3c}')	0.69	0.22	0.24	1.11
	Cognitive (b_{4c}')	1.38	0.27	0.88	1.95
	Difference	−0.69	0.46	−1.66	0.11

Total	2.07	0.20	1.68	2.49

Regression analysis showed that the complex label condition was associated with significantly greater affective trust than the simple label condition (a_5), $b = 0.41$, $\beta = 0.30$, $SE = 0.10$, $t(152) = 4.26$, $p < .001$, 95% CI [0.22, 0.60], and greater cognitive trust (a_6), $b = 0.55$, $\beta = 0.31$, $SE = 0.12$, $t(152) = 4.50$, $p < .001$, 95% CI [0.31, 0.79], providing support for H4. The differences in these effects are shown in Figure 5, and all effects are illustrated in the mediation model in Figure 6.

Figure 5

Effect of Different Certification Labels for TAI on Affective and Cognitive Trust.

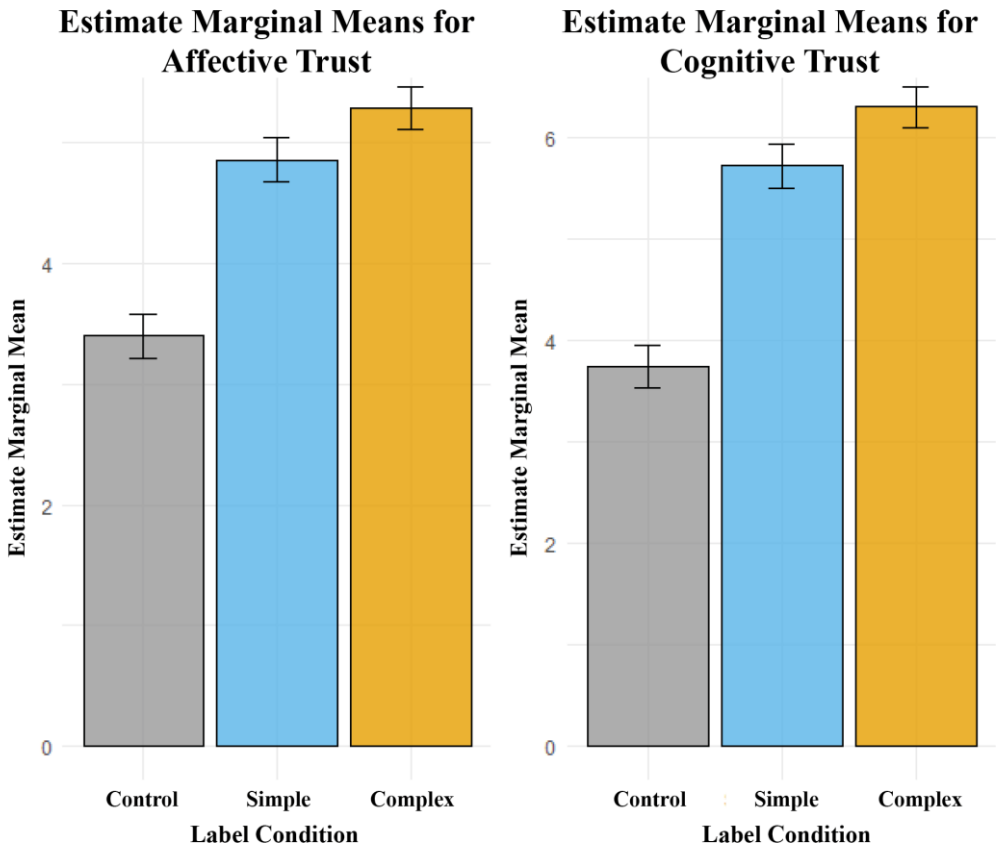
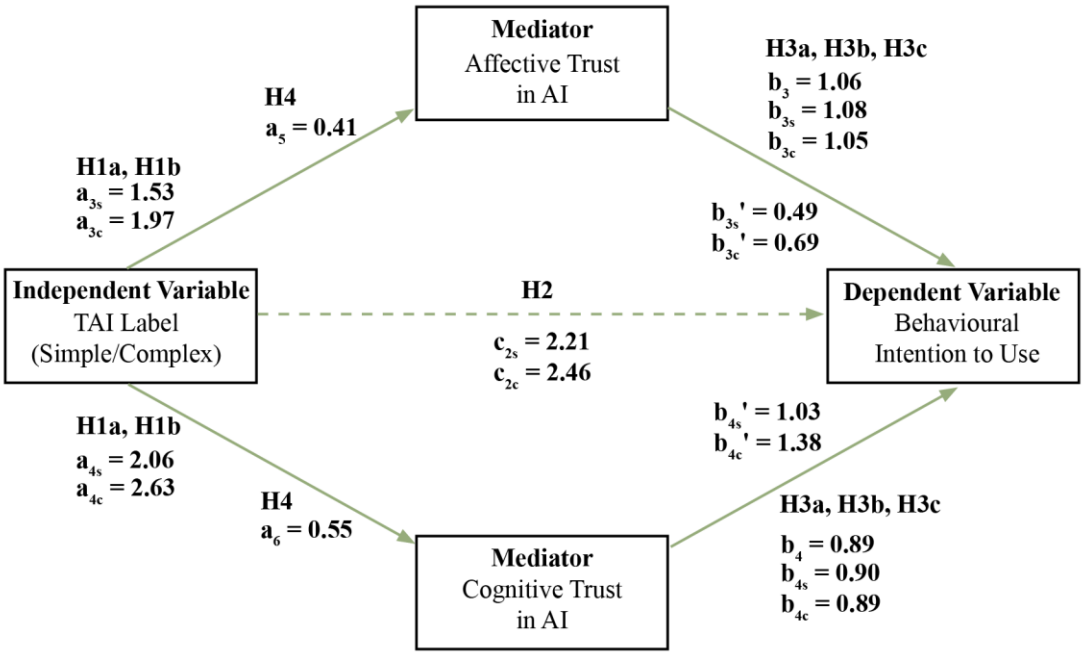


Figure 6

Parallel Mediation Model for Effects of the Simple and Complex Labels, Study 2



5 Study 3

Studies 1 and 2 had participants evaluate an AI within the high-stakes context of recruitment, which was chosen due to the effectiveness of labels within such scenarios (Scharowski et al., 2023). However, we wanted to test differences in how labels might interact with trust and BIU between high- (e.g., recruitment) and low-stakes scenarios. (e.g., music recommendation; Scharowski et al., 2023). Therefore, in Study 3, stakes were manipulated orthogonally to explore their effect. Furthermore, we aimed to test for potential spillover effects of labelling AI systems. Therefore, we created a design (detailed below) to measure differences in the evaluation of AI systems between participants who had been previously exposed to a label and those who had not.

5.1 Participants

Through Prolific we requested a representative sample of the population of the UK in terms of gender, age, and ethnicity. Because we expected stakes to exert a smaller effect than

the previously examined variables, we powered the study for a small effect and added a 10% buffer for inattentive responses, resulting in a target sample of 866 participants. After attention-check and CAPTCHA elimination[4] (20), 846 participants, were included in the final sample (52.4% female, 46.7% male, 0.6% prefer not to say, 0.4% non-binary/third gender), whose ages ranged from 18 to 84 years (M = 47.00, SD = 15.38), and who, on average, rated themselves as having a moderately good understanding of AI (M = 5.29, SD = 1.10). See Table 7 for more information.

Table 7

Sample Statistics, Study 3

Demographic	Value	*N*	%
Education	Less than Secondary education	4	0.5%
	Secondary education	277	32.7%
	Bachelor's degree	360	42.6%
	Master's degree	141	16.7%
	Doctoral degree	35	4.1%
	Other	29	3.4%
Employment	Employed	499	59.0%
	Freelancer	100	12.0%
	Retired	125	15.0%
	Student	42	5.0%
	Unemployed	46	5.0%
	Worker and Student	9	1.0%
	Other	25	3.0%
Country of Residence	Bulgaria	2	0.2%
	China	4	0.5%
	Germany	3	0.4%
	Ireland	4	0.5%
	Italy	4	0.5%
	Japan	2	0.2%
	Nigeria	3	0.4%
	Other	20	2.4%
	Poland	4	0.5%
	Portugal	2	0.2%
	United Kingdom	790	93.4%
	Missing	8	0.9%

5.2 Design

This study consisted of a mixed 2 (Label: Prior Exposure vs. No Prior Exposure) x 2 (Stakes: Low vs. High) between-subjects design. Label and Stakes conditions were the IVs, BIU the DV, Affective Trust and Cognitive Trust as mediators, and UAI and order were

[4] As the data were collected through an online platform in 2026, we included an additional CAPTCHA item as a quality-control measure to identify potential bot responses.

included as covariates. In both label conditions, participants rated their trust and BIU on two AI systems, sequentially (randomised). In the Prior Exposure condition, the first AI (AI_1) included a label and the second (AI_2) did not; in the No Prior Exposure condition, neither AI included a label (see Figure 7 for a schematic).

5.3 Materials and Measures

The measures of Affective Trust (AI_1: $\alpha = .92$, AI_2: $\alpha = .91$), Cognitive Trust (AI_1: $\alpha = .96$, AI_2: $\alpha = .96$), BIU (AI_1: $\alpha = .84$, AI_2: $\alpha = .72$), and UAI were identical to those used in Studies 1 and 2. The label consisted of the complex TAI label (Figure 4, right) and accompanying information used in Study 2. A manipulation check for Stakes was included (*MyJob*: $\alpha = .95$, *MyUni*: $\alpha = .91$, *MyMusic*: $\alpha = .83$, *MySeries*: $\alpha = .79$). It was composed of 3 items, e.g., "*How personally consequential would the outcome of using MyMusic be for you?*" (1 = *Not much*, 7 = *Extremely*).

5.3.1 AI Systems. This study used four different fictional AI systems. In the High-Stakes condition, AI_1 was the *MyJob* system described in Studies 1 and 2, and AI_2 was described as "an AI system called *MyUni* for evaluating degree applications", which would determine "whether [the participant] will be considered for admission to the course." In the Low-Stakes conditions, AI_1 was described as "an AI system called *MyMusic* for evaluating [the participant's] music preferences", which would "provide [the participant] with song recommendations", and AI_2 was described as "an AI system called *MySeries* for evaluating [the participant's] series preferences", which would "provide [the participant] with series recommendations." These additional scenarios were also adapted from Scharowski et al. (2023), based on the European Commission's AI Act (2024) classification of high- and low-risk applications of AI.

All conditions included a statement indicating that "Some AI systems have received a certification label for trustworthy AI. A certification label for trustworthy AI indicates that

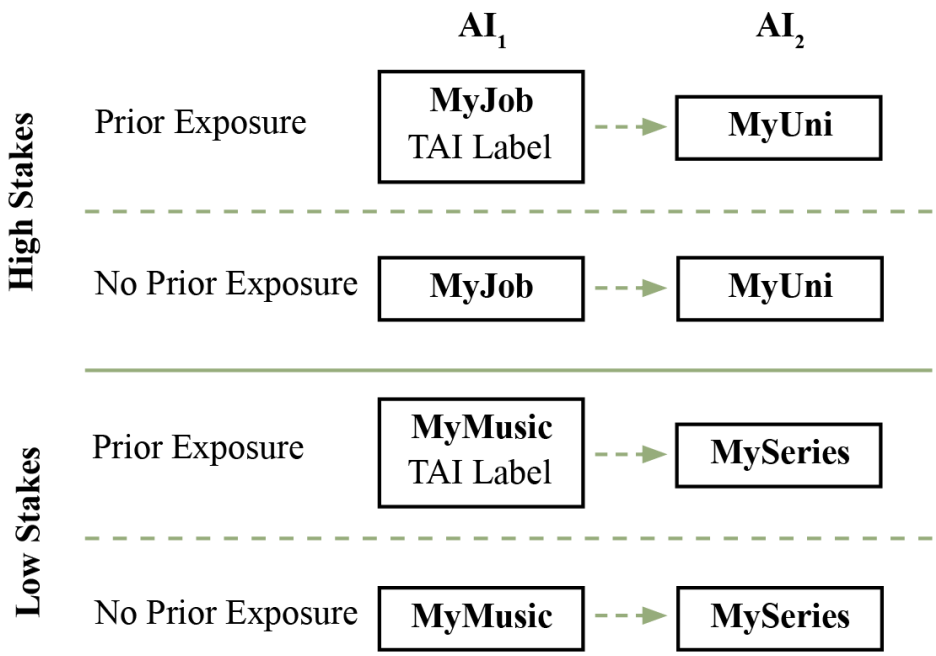
the system meets established criteria to be considered trustworthy." AI systems in Prior Exposure conditions were accompanied by the statement: "*[AI System] has received a certification label for trustworthy AI, as communicated by the logo below. The label was awarded by a foundation specialized in assessing AI trustworthiness. In this process, a panel of experts verified whether a digital service or product offered by a company (e.g., an AI as in the scenarios presented) meets certain criteria regarding trustworthy AI. These criteria were proposed by the European Commission's High-Level Expert Group on AI*." All unlabelled systems were accompanied by the statement, "[AI System] has not received a certification label," ensuring that participants across conditions were equally aware of the existence of AI certification labels. This approach isolates the effect of the focal AI system's certification status, rather than confounding it with mere exposure to information about certification labels, while also making the absence of certification explicit rather than leaving it ambiguous.

5.4 Procedure

Participants were randomly assigned to one of the four conditions (Label: Prior Exposure vs. No Prior Exposure; Stakes: Low vs. High). In each condition, participants repeated the procedure of Study 1 for two AI systems sequentially, with participants in the Prior Exposure condition being shown the complex TAI label (Figure 4, right) and accompanying information on AI_1 but not AI_2, and participants in the No Prior Exposure condition not being shown labels for any of the systems (Figure 7). See Appendix A for full survey details.

Figure 7

Display Sequence for AI Systems (Randomised)



Note. Because order was randomized, half of the participants were assigned to the order depicted in this figure and the other half to the reverse order. Specifically, half of the participants saw *MyUni* or *MySeries* first, followed by *MyJob* or *MyMusic*.

5.5 Results

5.5.1 Stakes. Analysis of the stakes manipulation check showed that the manipulation was successful. Participants in the high-stakes condition reported significantly higher composite manipulation-check scores than those in the low-stakes condition for both AI_1 and AI_2, with means being higher in the high-stakes conditions than the low-stake condition (Table 8).

Table 8

Stakes Manipulation Check Scores for AI_1 and AI_2, Study 3

	Low Stakes		High Stakes					95% CI	
	M	*SD*	*M*	*SD*	*t*	*df*	*p*	Lower	Upper
AI_1	2.96	1.32	4.32	1.70	13.01	790.65	<.001	1.16	1.57
AI_2	3.02	1.43	4.67	1.75	14.96	807.65	<.001	1.43	1.87

5.5.2 Order. Order effects were assessed using separate 2 (order: 1 vs. 2) × 2 (label: present vs. absent) × 2 (stakes: high vs. low) ANCOVAs (with UAI as a covariate) for AI_1 and AI_2 for affective trust, cognitive trust, and BIU. None of these interactions was statistically significant (all *p*s > .05) for any outcome or AI system, suggesting that the results were not meaningfully affected by order. Thus, order was dropped from further analyses.

5.5.3 Moderated Mediation, AI_1. To test the moderating effects of stakes on the relationships in the conceptual model, a moderated mediation model was used (Model 7, PROCESS v.5, Hayes, 2018; 95% CI, 5000 bootstrap replications, $N = 846$) on the measures (affective trust, cognitive trust, and BIU) for AI_1 only, with UAI included as a covariate (Table 9). Consistent with Studies 1 and 2, the presence of the label significantly predicted both affective trust (a_7) and cognitive trust (a_8), supporting H1a. A regression analysis on the interaction of trust type and label was found not to be significant ($b = -0.15$, $p = .20$) indicating that the certification label did not have a significantly stronger effect on cognitive trust than on affective trust, providing no support for H1b, in line with Study 1 but contrary to Study 2. The direct effect of the label's presence on BIU was significant (c_3), providing support for H2, consistent with both Studies 1 and 2. Both affective trust (b_5) and cognitive trust (b_6) were significant predictors of BIU, supporting H3a, consistent with both Studies 1 and 2.[5]

Table 9

Regression Coefficients, Study 3, AI_1

							95% CI	
Predictor	Outcome	*b*	*β*	*SE*	*t*	*p*	Lower	Upper
Label	Affective Trust (a_7)	0.99	0.40	0.11	8.92	<.001	0.77	1.21
High Stake	Affective Trust	−0.15	−0.06	0.11	−1.36	.175	−0.37	0.07
Label * High Stake	Affective Trust (d_1)	−0.12	−0.04	0.16	−0.76	.448	−0.43	0.19

[5] UAI was a marginally significant predictor of affective trust, in line with Study 2, and contrary to Study 1, where it was not significant. It was a significant predictor of BIU, consistent with Study 2 but contrary to Study 1. And it was not a significant predictor of cognitive trust, consistent with Study 1 but contrary to Study 2.

UAI	Affective Trust	0.07	0.06	0.04	1.91	.057	0.00	0.14
Label	Cognitive Trust (a_8)	0.77	0.30	0.12	6.65	<.001	0.54	1.00
High Stake	Cognitive Trust	−0.35	−0.14	0.12	−2.98	.003	−0.57	−0.12
Label * High Stake	Cognitive Trust (d_2)	0.04	0.01	0.16	0.23	.821	−0.29	0.36
UAI	Cognitive Trust	0.05	0.04	0.04	1.22	.223	−0.03	0.12
Label	BIU (c_3)	0.55	0.21	0.14	3.93	<.001	0.24	0.82
Label	BIU (mediated; c_3')	−0.13	−0.04	0.08	−1.57	.117	−0.29	0.03
Affective Trust	BIU (b_5)	0.47	0.38	0.05	8.58	<.001	0.36	0.57
Cognitive Trust	BIU (b_6)	0.41	0.35	0.05	7.93	<.001	0.31	0.52
Label * High Stake	BIU (d_3)	0.17	0.03	0.20	0.87	.383	−0.22	0.56
UAI	BIU	0.15	0.11	0.03	4.39	<.001	0.08	0.22

When affective and cognitive trust were included as mediators, the direct effect of the label on BIU (c_3'; Table 9) was no longer significant, while the indirect effects were significant for both affective (b_5'), and cognitive trust (b_6'; Table 10), for both stake levels, indicating that they fully mediated the effect, supporting H3b, and consistent with both Studies 1 and 2. The differences between these effects were not significant (Table 10), providing no support for H3c, consistent with both Studies 1 and 2.

When looking at the indices of moderated mediation, we find no evidence for these indirect effects (d_1, d_2; Table 9), indicating that stakes did not significantly moderate the effect of the label on affective or cognitive trust, providing no support for H5a (Figure 8). Stakes did also not moderate the relationship between labels presence and BIU (d_3), providing no support for H5b.

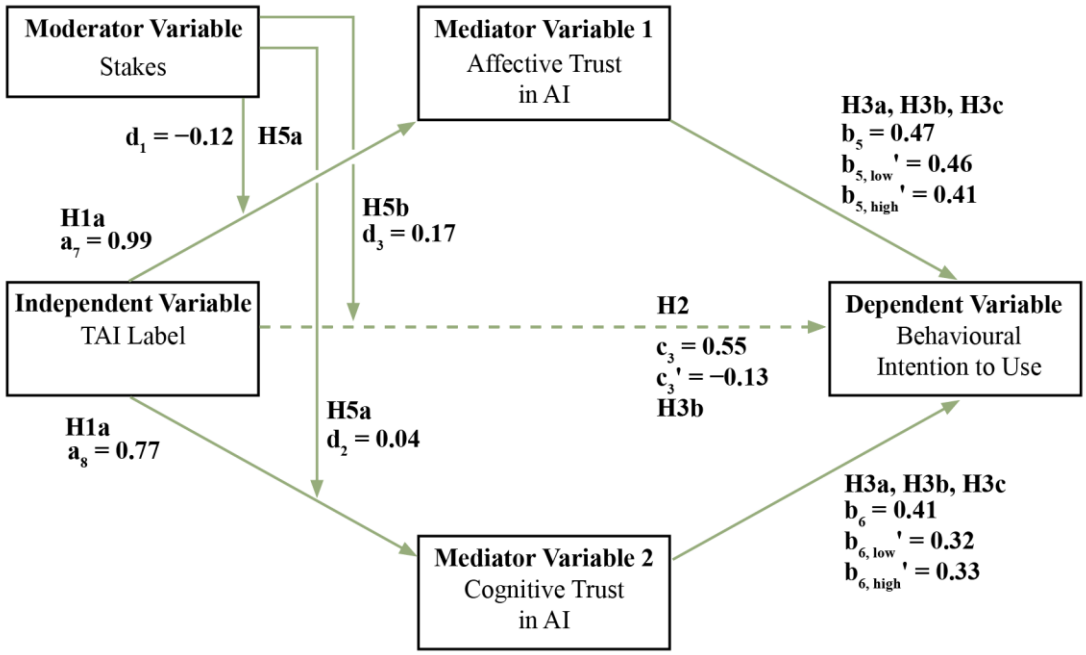
Table 10

Significant Indirect Effects of Label on BIU via Affective and Cognitive Trust, Study 3, AI_1

				95% CI	
Mediator	Stakes	*b*	*BootSE*	Lower	Upper
Affective Trust	Low ($b_{5,\,low}'$)	0.46	0.08	0.32	0.62
	High ($b_{5,\,high}'$)	0.41	0.08	0.27	0.57
Cognitive Trust	Low ($b_{6,\,low}'$)	0.32	0.06	0.21	0.45
	High ($b_{6,\,high}'$)	0.33	0.07	0.20	0.49
Difference	Low	0.24	0.16	−0.07	0.57
	High	0.03	0.13	−0.23	0.27

Figure 8

Parallel Moderated Mediation Model of Effect of Label, Study 3, AI_1



5.5.4 Spillover Effects. To test the effects of prior exposure to the label on the outcomes for the unlabelled system (spillover), a mediation model[6] was used (Model 4, PROCESS v.5, Hayes, 2018; 95% CI, 5000 bootstrap replications, $N = 846$), with affective and cognitive trust as the mediating measures, BIU as the DV, with UAI as a covariate (all for AI_2). First, results (Table 11) show that the responses of those with prior exposure to a label dropped significantly from AI_1 to AI_2 across all three outcomes, with the means for the exposed group above the means for the non-exposed group at AI_1 and below them at AI_2, consistent with a negative spillover effect. Those without prior exposure showed a small but significant decrease in affective trust from AI_1 to AI_2, with no significant changes in cognitive trust or BIU. See Table 11 for descriptive statistics.

Table 11

[6] Given we have no hypotheses concerning stake's role in spillover effects, and stakes did not interact with any of the previous results, we proceeded with the analysis collapsing for this variable, which allows greater statistical power.

Means and Mean Differences Between AI_1 and AI_2, Study 3

		AI_1		AI_2					95% CI_{Diff}	
		M	*SD*	*M*	*SD*	*Diff*	*t*	*p*	Lower	Upper
Affective Trust	Prior Exposure	4.69	1.22	3.53	1.15	−1.15	−21.73	<.001	−1.26	−1.05
	No Prior Exposure	3.75	1.09	3.66	1.12	−0.09	−2.44	.015	−0.16	−0.02
Cognitive Trust	Prior Exposure	5.06	1.19	3.92	1.25	−1.15	−19.44	<.001	−1.26	−1.03
	No Prior Exposure	4.27	1.22	4.21	1.26	−0.06	−1.67	.095	−0.13	0.01
BIU	Prior Exposure	4.04	1.54	3.03	1.38	−1.01	−13.95	<.001	−1.16	−0.87
	No Prior Exposure	3.40	1.43	3.32	1.41	−0.08	−1.50	.134	−0.17	0.02

However, a spillover effect only emerged for cognitive trust, with prior exposure to a label being a significant negative predictor of cognitive trust (a_{10}) in AI_2, but not affective trust (a_9), providing partial support for H6a (Table 12). When looking at the direct effect on BIU (m), we find a spillover effect, providing support for H6b.

When including affective trust and cognitive trust as mediators in the model, the spillover effect on BIU is no longer significant (c_4'; Table 12), suggesting full mediation and supporting H6c. The spillover effect acted on BIU through cognitive trust, with prior exposure having a significant negative indirect effect (b_8'), $b = -0.11$, *BootSE* = 0.04, 95% CI [−0.19, −0.04]. There was no significant indirect effect through affective trust (b_7') $b = -0.05$, *BootSE* = 0.03, 95% CI [−0.11, 0.01]). These results provide partial support for H6c (Figure 9).

Table 12

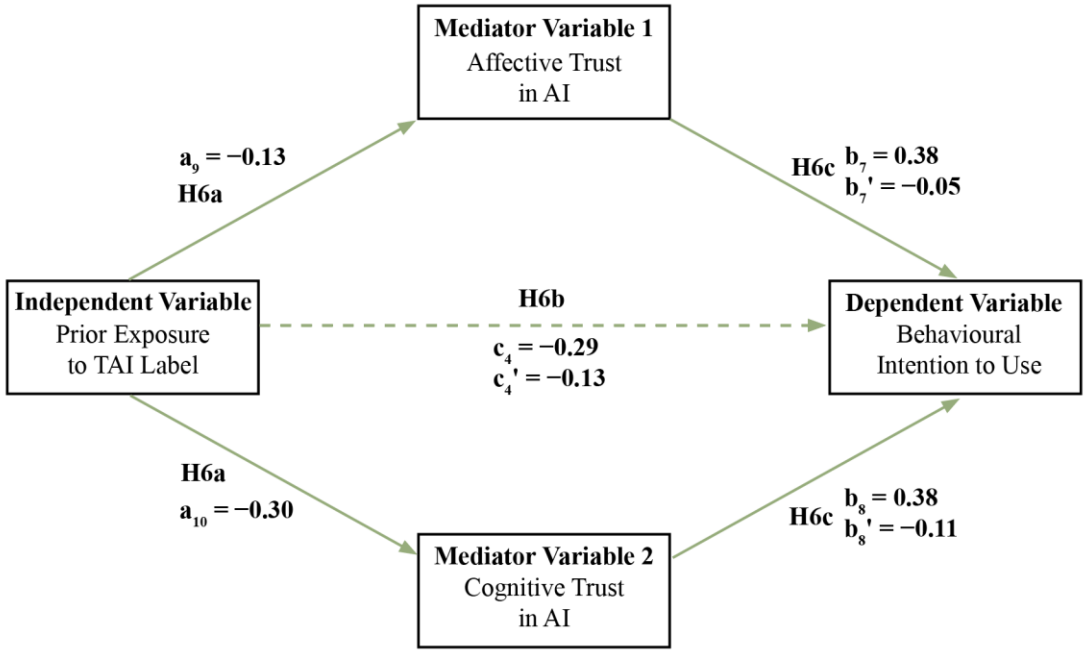
Regression Coefficients, Study 3, AI_2

							95% CI	
Predictor	Outcome	*b*	*β*	*SE*	*t*	*p*	Lower	Upper
Prior Exposure	Affective Trust (a_9)	−0.13	−0.06	0.08	−1.68	.093	−0.28	0.02
UAI	Affective Trust	0.05	0.05	0.04	1.40	.163	−0.02	0.12
Prior Exposure	Cognitive Trust (a_{10})	−0.30	−0.12	0.09	−3.46	.001	−0.47	−0.13
UAI	Cognitive Trust	0.03	0.02	0.04	0.71	.481	−0.05	0.11
Prior Exposure	BIU (c_4)	−0.29	−0.11	0.10	−3.10	<.01	−0.48	−0.10
Prior Exposure	BIU (mediated; c_4')	−0.13	−0.05	0.08	−1.76	.078	−0.28	0.02

Affective Trust	BIU (b_7)	0.38	0.31	0.05	7.19	<.001	0.28	0.48
Cognitive Trust	BIU (b_8)	0.38	0.34	0.05	7.88	<.001	0.28	0.47
UAI	BIU	0.07	0.06	0.03	2.16	.031	0.01	0.14

Figure 9

Parallel Mediation Model of Effects of Prior Exposure, Study 3, AI_2



6 Discussion

6.1 Theoretical Implications

We examined certification labels as a non-technical solution to enhancing trust and acceptance of AI. In line with our hypotheses, we found that certification labels increase trust (H1a), with mixed evidence to suggest that cognitive trust is more affected than affective trust (Study 2, but not Studies 1 or 3; H1b); that certification labels also increase acceptance (H2) of AI systems; that trust predicts acceptance (H3a); that both affective and cognitive trust act as mediators of the effect of the labels on acceptance (H3b), with equal strength (contrary to H3c); that including TAI requirements on the label has a greater effect on both trust and acceptance (H4); that stakes do not moderate the label's effects (contrary to H5a and

H5b); and that labels can have a negative spillover effect—having prior exposure to a labelled AI system reduces cognitive trust in unlabelled systems but not affective trust (partially supporting H6a) or acceptance directly (H6b), although it can act on acceptance through cognitive trust (partially supporting H6c).

These results largely support previous research demonstrating the effectiveness of certification labels in aiding trust-formation mechanisms by helping end-users assess a system's trustworthiness (Jacovi et al., 2021; Lee & See, 2004; Liao & Sundar, 2022), which can be strengthened by making the label's connection to TAI requirements more explicit, allowing end-users to better assess the AI's reliability (Jacovi et al., 2021; Liao & Sundar, 2022; Scharowski et al., 2023; Stuurman & Lachaud, 2022).

The findings of the present research regarding the relationships between cognitive and affective trust and AI also align with the mixed findings of previous research (e.g., Chen & Sundar, 2023; Cihon et al., 2021; Glikson & Woolley, 2020; Scharowski et al., 2023; Schuitema et al., 2020). Various characteristics of a given AI system (e.g., transparency, anthropomorphism) seem to have widely varying effects on cognitive and affective trust depending on a broad range of factors (e.g., appearance, domain; Glikson & Woolley, 2020). However, in the context of certification labels, the present findings strongly point to the conclusion that certification labels have their primary effects through cognitive trust. This aligns with the fact that TAI certification labels are designed to signal competence and reliability rather than benevolence (Liao & Sundar, 2022; Mayer et al., 1995).

This design factor may also help explain why negative spillover effects were only found for cognitive trust. That is, because the components of cognitive trust are easier to directly signal through certification labels than the slow building of an emotional bond required for affective trust (see Chen & Sundar, 2023; Glikson & Wooley, 2020; Li et al., 2025; Schuitema et al., 2020), the comparison of AI systems based on certification labels is

more likely to activate the immediate perceptions of trustworthiness characteristic of cognitive deliberations of trust rather than feelings of affective trust, which require sustained experience with the system over time (Dowell et al., 2015).

The results of Study 3 demonstrated that the effects were consistent across different levels of stakes, even though the manipulation check showed the different scenarios to be perceived as high and low stakes. This consistency stands in contrast to research demonstrating a preference for AI certification labels in high-stakes contexts (Scharowski et al., 2023) and differences in trust across high- and low-stakes contexts (Freel et al., 2025; Johnson, 2025). This may suggest that certification labels work as heuristics of symbolic credibility, where a user assumes the high-risk factors to have been taken into account in the certification process signalled by the label. Indeed, this ready interpretability is one of the proposed values of certification labels (Jacovi et al., 2021; Liao & Sundar, 2022). Research testing the effectiveness of different label design factors and their interaction with various contexts would be needed to determine the limit conditions for these effects.

Our findings provide strong support for the inclusion of trust in technology acceptance models in past research (e.g., Alalwan et al., 2018; Chi et al., 2023; Hooda et al., 2022; Li et al., 2025; Mersha et al., 2024; Slade et al., 2015; Söllner et al., 2016; Wu et al., 2011). However, the spillover effects demonstrated in Study 3 also suggest a potential extension to these models by demonstrating that interactions with the certification label of one AI system can affect a user's trust-formation mechanisms with other systems.

6.2 Stakeholder Implications, Contributions, and Limitations

While the OECD have begun developing the*ir AI Trust Standard and Label* (2022), there is no official internationally-accepted certification label for TAI. However, AI's growing use in products and services makes its trustworthiness a prominent challenge in overcoming consumer hesitance. The present findings suggest that certification labels are not

only important technical benchmarks to meet, but may offer a competitive advantage in the AI market. This is especially true if such labels have a negative spillover effect on unlabelled systems, and worsened if certification processes are slow or biased in their distribution, making it difficult for end-users to determine the trustworthiness of unlabelled systems (Scharowski et al., 2023). However, while the design of the label and its accompanying information was based on key success factors in AI regulation (Scharowski et al., 2023; Stuurman & Lachaud, 2022) and literature on labelling (Schebesta, 2019), a manipulation check was not included within the studies, and the effects of certification labels can depend on user familiarity and long-term exposure (Dowell et al., 2015). Therefore, we acknowledge that the novelty of the label and the lack of a manipulation check may limit the ecological validity of the findings.

Furthermore, in developing H1b, we argued that TAI certification labels would have a stronger effect on cognitive than affective trust, given that such labels are designed to signal competence and reliability and that trust in AI systems seems to run more through cognitive than affective trust. However, the scenarios used (e.g., *MyJob*) involved the expectation of divulging sensitive personal data, which may resemble human-centred interactions involving information disclosure that elicit stronger reliance on affective trust (Ferràs-Hernández, 2018; Gill et al., 2024). Indeed, as generative AI systems rapidly advance, their capacity to simulate features of human interaction and handle personal data also increases, which may enhance the salience of affective trust and potentially attenuate the relative dominance of cognitive trust in shaping overall trust responses. Thus, H1b may be bounded by contexts in which the interactions with the AI systems are perceived as highly human-like. The mixed support for H1b across the present studies, together with the non-significant findings testing H3c, suggests that the effects of certification labels on cognitive and affective trust may be more context-dependent than originally theorised. Future research should therefore examine

whether the impact of certification labels on cognitive versus affective trust is moderated by the perceived human-like features of the interaction.

By linking TAI and certification labels, this research contributes significantly to the discussion of trust in AI and AI acceptance. It contributes to supporting the development of certification labels as a positive influence on trust and acceptance of AI, suggesting additional factors to be considered in certification frameworks, such as those of the AI HLEG (2019). Specifically, this work supports previous findings on the effectiveness of certification labels generally (Cuesta-Valiño et al., 2024; Mazzù et al., 2022) and offers specific insights within the context of AI, challenging previous design recommendations that warn against high information content (Stuurman & Lachaud, 2022) and suggesting that comprehensive information on AI certification labels is an important factor in building trust. The findings bolster past studies suggesting that certification labels have a greater effect on cognitive than affective trust (Chen & Sundar, 2023; Cihon et al., 2021; Scharowski et al., 2023; Schuitema et al., 2020).

Research suggests that trust in AI systems is shaped, in part, by users' digital literacy and familiarity with AI (Kozak & Fel, 2024), making sample composition an important consideration for the present findings. In Studies 1 and 2, relatively high proportions of students (44.3% and 48.1%, respectively) reported moderately strong self-assessed understanding of AI (5.30 and 5.48, respectively), which may indicate a level of baseline familiarity that is not fully representative of the broader population. Accordingly, one limitation is that the observed effectiveness of certification labels may be somewhat inflated in samples that are comparatively comfortable engaging with AI. However, this concern is partly alleviated by Study 3, in which participants reported a similar level of AI understanding (5.29) but were recruited via Prolific using quotas to approximate UK population representativeness in terms of gender, age, and ethnicity.

Even with this caveat, the replication of the effects across Studies 1 to 3 enhances confidence in their robustness and strengthens their practical relevance. Specifically, the findings suggest that certification labels may be effective across a wider range of contexts than prior literature has proposed, challenging claims that such labels are only necessary or effective in high-stakes settings (e.g., Afroogh et al., 2024; Freel et al., 2025; Johnson, 2025; Scharowski et al., 2023). For marketers, firms, and policy-makers, this implies that a single certification label may be deployed across diverse applications rather than restricted to high-stakes domains alone. Nevertheless, because trust in AI may vary as a function of familiarity and digital literacy, future research should examine whether these effects hold among populations with lower AI exposure, thereby clarifying the scope and limits of these practical recommendations.

The present research extends previous research on spillover effects on consumer products (Anagnostou et al., 2015; Sherif & Hovland, 1961) into the realm of AI certification, providing valuable insights into the externalities of certification labelling in this emerging industry. It also makes a unique contribution to research on the reliable communication of trustworthiness in AI by being, to our knowledge, the first to show that certification labels specifically for AI can lead to BIU through cognitive and affective trust, and that negative spillover effects on trust and BIU can occur in the context of such labels.

7 Conclusion

As AI seeps into more areas of life, concerns about its trustworthiness have followed. Ensuring these systems are trustworthy and communicating that trustworthiness has become a significant challenge for researchers, developers, and policy-makers. The present research has made a significant contribution to addressing this challenge by investigating and demonstrating that TAI certification labels are an effective non-technical method for communicating trustworthiness and increasing both trust in and the acceptance of AI tools.

References

Afroogh, S., Akbari, A., Malone, E., Kargar, M., & Alambeigi, H. (2024). Trust in AI: Progress, challenges, and future directions. *Humanities and Social Sciences Communications*, *11*, Article 1568. https://doi.org/10.1057/s41599-024-04044-8

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

Anagnostou, A., Ingenbleek, P. T. M., & van Trijp, H. C. M. (2015). Sustainability labelling as a challenge to legitimacy: Spillover effects of organic Fairtrade coffee on consumer perceptions of mainstream products and retailers. *Journal of Consumer Marketing*, *32*(6), 422–431. https://doi.org/10.1108/JCM-11-2014-1213

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

Bhattacharya, R., Devinney, T. M., & Pillutla, M. M. (1998). A formal model of trust based on outcomes. *Academy of Management Review*, *23*, 459–472. https://doi.org/10.5465/amr.1998.926621

Blösser, M., & Weihrauch, A. (2024). A consumer perspective of AI certification – the current certification landscape, consumer approval and directions for future research. *European Journal of Marketing*, *58*(2), 441–470. https://doi.org/10.1108/EJM-01-2023-0009

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

Chen, C., & Sundar, S. S. (2023). Is this AI trained on credible data? The effects of labeling quality and performance bias on user trust. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, (pp. 1–11), Article 816. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3544548.3580805

Chen, X., Wang, X., Spitzley, L., & Nunamaker, J. (2023). Trust and deception with high stakes: Evidence from the friend or foe dataset. *Decision Support Systems*, *173*, Article 113997. https://doi.org/10.1016/j.dss.2023.113997

Chi, O. H., Chi, C. G., Gursoy, D., & Nunkoo, R. (2023). Customers' acceptance of artificially intelligent service robots: The influence of trust and culture. *International Journal of Information Management*, *70*, Article 102623. https://doi.org/10.1016/j.ijinfomgt.2023.102623

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

Cornudet, C., Laporte, M.-E., Berger-Remy, F., Parguel, B., & Richet, J.-L. (2025). When food scanner apps outperform front-of-pack nutrition labels: A conditional process model to foster healthier food choices in times of growing distrust. *Psychology & Marketing*, *42*, 1963–1978. https://doi.org/10.1002/mar.22214

Crockett, K. A., Gerber, L., Latham, A., & Colyer, E. (2021). Building trustworthy AI solutions: A case for practical solutions for small businesses. *IEEE Transactions on Artificial Intelligence*, *4*(4), 778–791. https://doi.org/10.1109/TAI.2021.3137091

Cuesta-Valiño, P., Gutiérrez-Rodríguez, P., García-Henche, B., & Núñez-Barriopedro, E. (2024). The impact of corporate social responsibility on consumer brand engagement and purchase intention at fashion retailers. *Psychology & Marketing*, *41*, 649–664. https://doi.org/10.1002/mar.21940

Davis, F. D. (1985). *A technology acceptance model for empirically testing new end-user information systems: Theory and results* [Doctoral dissertation, Massachusetts Institute of Technology]. DSpace@MIT. http://hdl.handle.net/1721.1/15192

Delapedra, A. T., Costa, M. L. A., Paredes, D. Z., & Silva, A. H. G. R. d. (2025). Decoding food labels: How and why labels influence consumers' responses. *Psychology & Marketing*, *0*, 1–15. https://doi.org/10.1002/mar.70092.

Deriu V., Pozharliev, R., & De Angelis, M. (2024). How trust and attachment styles jointly shape job candidates' AI receptivity. *Journal of Business Research*, *179*, Article 114717. https://doi.org/10.1016/j.jbusres.2024.114717

Dowell, D., Morrison, M., & Heffernan, T. (2015). The changing importance of affective trust and cognitive trust across the relationship lifecycle: A study of business-to-business relationships. *Industrial Marketing Management*, *44*, 119–130. https://doi.org/10.1016/j.indmarman.2014.10.016

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

Ferràs-Hernández, X. (2017). The future of management in a world of electronic brains. *Journal of Management Inquiry*, *27*(2), 260–263. https://doi.org/10.1177/1056492617724973

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behaviour.* Reading. Mass.: Addison-Wesley.

Freel, A., Pias, S. B. H., Šabanović, S., & Kapadia, A. (2025). How misclassification severity and timing influence user trust in AI image classification: User perceptions of high- and low-stakes contexts. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*, (pp. 2906–2923). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3715275.3732187

Fukukawa, K., & Trivedi, R. H. (2025). Empathy, ethics and efficacy: The 3Es of implementing artificial intelligence for consumer encounters. *Psychology & Marketing*, *42*, 2352–2368. https://doi.org/10.1002/mar.22235

Gill, H., Vreeker-Williamson, E., Hing, L. S., Cassidy, S. A., & Boies, K. (2024). Effects of cognition-based and affect-based trust attitudes on trust intentions. *Journal of Business and Psychology*, *39*(6), 1355–1374. https://doi.org/10.1007/s10869-024-09986-z

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

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

Haque, M. A., & Li, S. (2025). Exploring ChatGPT and its impact on society. *AI and Ethics*, *5*, 791–803. https://doi.org/10.1007/s43681-024-00435-4

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach (Methodology in the social sciences)* (2nd ed.). New York, NY: The Guilford Press.

High-Level Expert Group on Artificial Intelligence (AI HLEG) (2019). *Ethics guidelines for trustworthy artificial intelligence*. https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1

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

Hooda, A., Gupta, P., Jeyaraj, A., Giannakis, M., & Dwivedi, Y. K. (2022). The effects of trust on behavioral intention and use behavior within e-government contexts. *International Journal of Information Management*, *67*, Article 102553. https://doi.org/10.1016/j.ijinfomgt.2022.102553

International Organization for Standardization. (2020). *Conformity assessment—Vocabulary and general principles* (ISO/IEC Standard No. 17000:2020). https://www.iso.org/standard/73029.html

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (FaccT '21)*, (pp. 624-635). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3442188.3445923

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

Johnson, D. (2025). *Higher stakes, healthier trust? An application-grounded approach to assessing healthy trust in high-stakes human-AI collaboration*. arXiv. https://doi.org/10.48550/arXiv.2503.03529

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

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

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

Kozak, J., & Fel, S. (2024). How sociodemographic factors relate to trust in artificial intelligence among students in Poland and the United Kingdom. *Scientific Reports*, *14*, Article 28776. https://doi.org/10.1038/s41598-024-80305-5

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

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.

Li, Y., & Goel, S. (2025). Making it possible for the auditing of AI: A systematic review of AI audits and AI auditability. *Information Systems Frontiers*, *27*, 1121–1151 https://doi.org/10.1007/s10796-024-10508-8

Li, B., Lai, E. Y., & Wang, X. (Shane). (2025). From tools to agents: Meta-analytic insights into human acceptance of AI. *Journal of Marketing*, *0*(0), 1–21. https://doi.org/10.1177/00222429251355266

Liao, Q. V., & Sundar, S. S. (2022). Designing for responsible trust in AI systems: A communication perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FaccT '22),* (pp. 1257–1268). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3531146.3533182

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

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

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

Mazzù, M. F., Marozzo, V., Baccelloni, A., & Giambarresi, A. (2023). The effects of combining front-of-pack nutritional labels on consumers' subjective understanding, trust, and preferences. *Psychology & Marketing*, *40*, 1484–1500. https://doi.org/10.1002/mar.21838

McKnight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on*

Management Information Systems, *2*(2), 1–25. https://doi.org/10.1145/1985347.1985353

Mersha, M., Lam, K., Wood, J., AlShami, A. K., & Kalita, J. (2024). Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. *Neurocomputing*, *599*, Article 128111. https://doi.org/10.1016/j.neucom.2024.128111

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

Orbán, F., & Stefkovics, Á. (2025). Trust in artificial intelligence: A survey experiment to assess trust in algorithmic decision-making. *AI & Society*, *40*, 4955–4969. https://doi.org/10.1007/s00146-025-02237-6

Organisation for Economic Co-operation and Development (2022). *AI trust standard & label*. https://oecd.ai/en/catalogue/tools/ai-trust-standard-and-label

Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems*, *34*(2), Article 101885. https://doi.org/10.1016/j.jsis.2024.101885

Scharowski, N., Benk, M., Kühne, S. J., Wettstein, L., & Brühlmann, F. (2023). Certification labels for trustworthy AI: Insights from an empirical mixed-method study. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FaccT '23)*, (pp. 248–260). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3593013.3593994

Schebesta, H. (2019). Control in the label: Self-declared, certified, accredited? On-pack consumer communication about compliance control in voluntary food schemes from a legal perspective. In P. Rott (Ed.), *Studies in European Economic Law and Regulation Vol. 16: Certification – Trust, Accountability, Liability* (pp. 143–161). Springer. https://doi.org/10.1007/978-3-030-02499-4_7

Schiff, D. S., Kelley, S., & Camacho Ibáñez, J. (2024). The emergence of artificial intelligence ethics auditing. *Big Data & Society*, *11*(4). https://doi.org/10.1177/20539517241299732

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

Shamim, S., Yang, Y., Zia, N. U., Khan, Z., & Shariq, S. M. (2023). Mechanisms of cognitive trust development in artificial intelligence among front line employees: An empirical examination from a developing economy. *Journal of Business Research*, *167*, Article 114168. https://doi.org/10.1016/j.jbusres.2023.114168

Sherif, M., & Hovland, C. I. (1961). *Social judgment: Assimilation and contrast effects in communication and attitude change*. Yale University Press: New Haven.

Sigurdsson, V., Larsen, N. M., Folwarczny, M., Fagerstrøm, A., Menon, R. V., & Sigurdardottir, F. T. (2023). The importance of relative customer-based label equity when signaling sustainability and health with certifications and tags. *Journal of Business Research*, *154*, 113338. https://doi.org/10.1016/j.jbusres.2022.113338

Slade, E. L., Dwivedi, Y. K., Piercy, N. C., & Williams, M. D. (2015), Modeling consumers' adoption intentions of remote mobile payments in the United Kingdom: Extending UTAUT with innovativeness, risk, and trust. *Psychology & Marketing*, *32*, 860–873. https://doi.org/10.1002/mar.20823

Söllner, M., Hoffmann, A., & Leimeister, J. M. (2016). Why different trust relationships matter for information systems users. *European Journal of Information Systems*, *25*(3), 274–287. https://doi.org/10.1057/ejis.2015.17

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

Sundar, A., Cao, E., Wu, R., & Kardes, F. R. (2021). Is unnatural unhealthy? Think about it: Overcoming negative halo effects from food labels. *Psychology & Marketing*, *38*(8), 1280–1292. https://doi.org/10.1002/mar.21485

Tomlinson, E. C., Schnackenberg, A. K., Dawley, D., & Ash, S. R. (2020). Revisiting the trustworthiness–trust relationship: Exploring the differential predictors of cognition- and affect-based trust. *Journal of Organizational Behavior*, *41*, 535–550. https://doi.org/10.1002/job.2448

Vanneste, B. S., & Puranam, P. (2025). Artificial intelligence, trust, and perceptions of agency. *Academy of Management Review*, *50*(4), 726–744. https://doi.org/10.5465/amr.2022.0041

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3), 425–478. https://doi.org/10.2307/30036540

Wu, K., Zhao, Y., Zhu, Q., Tan, X., & Zheng, H. (2011). A meta-analysis of the impact of trust on technology acceptance model: Investigation of moderating influence of subject and context type. *International Journal of Information Management*, *31*(6), 572–581. https://doi.org/10.1016/j.ijinfomgt.2011.03.004

Xu, S., Gong, Z., Li, Y., Zhao, Y., & Huang, S. (2026). The trust cost of AI errors: Examining cognitive and affective trust dynamics in human-AI collaboration. *International Journal of Human–Computer Interaction*, *42*(10), 7666–7681. https://doi.org/10.1080/10447318.2025.2560522

Zehnle, M., Hildebrand, C., & Valenzuela, A. (2025). Not all AI is created equal: A meta-analysis revealing drivers of AI resistance across markets, methods, and time. *International Journal of Research in Marketing*, *42*(3, Part B), 729–751. https://doi.org/10.1016/j.ijresmar.2025.02.005