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

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

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

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

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

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

ABSTRACT

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

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**.

KEYWORDS

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

1 INTRODUCTION

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

To this end, a large body of work has focused on identifying the principles that underlie trustworthy AI [36]. They include mitigating bias and unfairness in AI systems [41], explaining the reasoning of AI decisions [39], setting up mechanisms to hold AI accountable [36], and ensuring user privacy [60]. However, as trust is determined by people's perception [40, 43], efforts to design trustworthy AI are hampered by a lack of understanding of how to communicate trustworthiness to people, for instance, through documentation or other transparency affordances [43]. Particularly for end-users[1], trusting AI can be a challenge, as they lack the necessary expertise and knowledge to evaluate the various trustworthiness principles (e.g., robustness, privacy, fairness) [4, 37].

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

---

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

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

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

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

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

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

2 AUDITING FOR TRUSTWORTHY AI

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

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

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

---

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

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

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

3 CERTIFICATION LABELS FOR AUDITED AI

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

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

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

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

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

4 RESEARCH QUESTIONS

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

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

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

5 METHODS

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

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

5.1 The certification label

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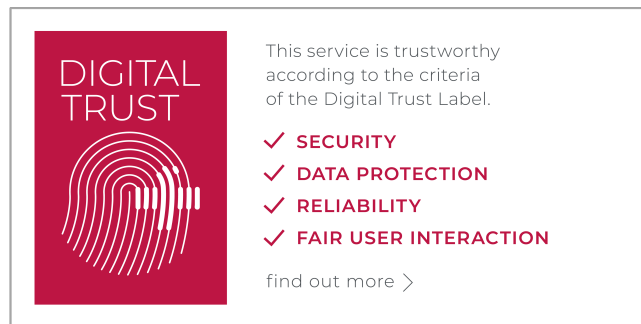

Figure 1: The "Digital Trust Label," which we adopted as a certification label for AI. ©2023 Swiss Digital Initiative

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

(1) Security (criteria 1 - 12): What is the security standard? The service provider shall, e.g., ensure that the data is encrypted as it transfers so that third-parties cannot access it.

(2) Data protection (criteria 13 - 20): How is the data protected? The service provider shall, e.g., assume responsibility for the appropriate management of the data.

(3) Reliability (criteria 21 - 29): How reliable is the service or product? The service provider shall, e.g., take all actions required to safeguard the continuity of the service.

(4) Fair user interaction (criteria 30 - 35): Is automated decision-making involved? The service provider shall, e.g., ensure that all users receive equal treatment and that there is no data-based service or price discrimination.

If an organization would like its digital product or service (e.g., a chatbot) to receive the certification label, it can voluntarily request an audit and thus participate in the certification process. After a scoping call with third-party auditors, an audit is performed along the criteria catalog. The audit leads to an audit report detailing the performance per criterion, which is double-checked by an independent label certification committee composed of auditing experts. If non-conformities are identified, the organization applying for the label must fix the identified issues, e.g., adjust its privacy policy. After a successful auditing report, the certification label is awarded for a period of three years with two audits during that period.

5.2 Scenario selection

Participants were presented with real-world examples of AI systems, adapted from Kapania et al., namely *medical diagnosis, loan approval, hiring procedure, music preference, route planning* and *price comparison* (see materials on OSF: https://osf.io/gzp5k/). One advantage of using hypothetical scenarios instead of real consumer applications is that differences in participants' prior experience with the applications can be controlled for Kapania et al. and Woods et al. proposed that people's behavior in scenario-based experiments corresponds to their real-life behavior. To answer our second research question and following Kapania et al. we explored both low-stake scenarios (music preference, route planning, price comparison) and high-stake scenarios (medical diagnosis, hiring procedure, loan approval). This distinction was crucial since other researchers [18, 61] and the "EU AI Act" [12] have discussed the use of AI labels for "low-stake" and "high-stake" scenarios. This classification was based on the AI's respective impact on affected parties and the involvement of significant risks, in particular with respect to safety, consumer rights, and the use of personal data.

5.3 Interviews

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

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

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

5.4 Survey

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

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

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

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

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

5.5 Analysis and coding procedure

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

6 RESULTS

6.1 Attitudes toward certification labels

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

Table 1: End-users' attitudes toward certification labels

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

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

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

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

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

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

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

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

6.2 Effects of certification labels

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

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

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

A majority of the survey participants directly indicated that they would prefer the use of a certification label in the selected high-stake scenario (63.2%, $n = 191$), compared to preferring a label in the low-stake scenarios (22.2%, $n = 67$), with 14.6% ($n = 44$) of participants indicating no preference. Regarding the different preferences for certification labels in low- and high-stake scenarios, participants from the interview study expressed a greater demand for a certification label in high-stake scenarios because of the higher scenario complexity, limited individual expertise, and a lack of prior experience with the system. Overall, 81.1% ($n = 245$) of survey participants stated a preference for using an AI with a label, compared to 6% ($n = 18$) that would prefer to use an AI without a label and 12.9% ($n = 39$) that stated no preference. Also, 70.9% ($n = 214$) indicated to be more likely to accept an AI's decision/recommendation if it had received a label, compared to 14.2% ($n = 43$) that indicated "no," and 14.9% ($n = 45$), that made no statement. Survey participants rated the importance of the existing label criteria in the context of AI at a high level with similar ratings for security ($M = 87.72$, $SD = 20.93$), data protection ($M = 85.04$, $SD = 21.81$), reliability ($M = 76.97$, $SD = 23.19$) and fair user interaction ($M = 80.80$, $SD = 23.37$). However, merely 55.3% ($n = 167$) of the participants agreed that the label addresses the concerns/challenges/risks they
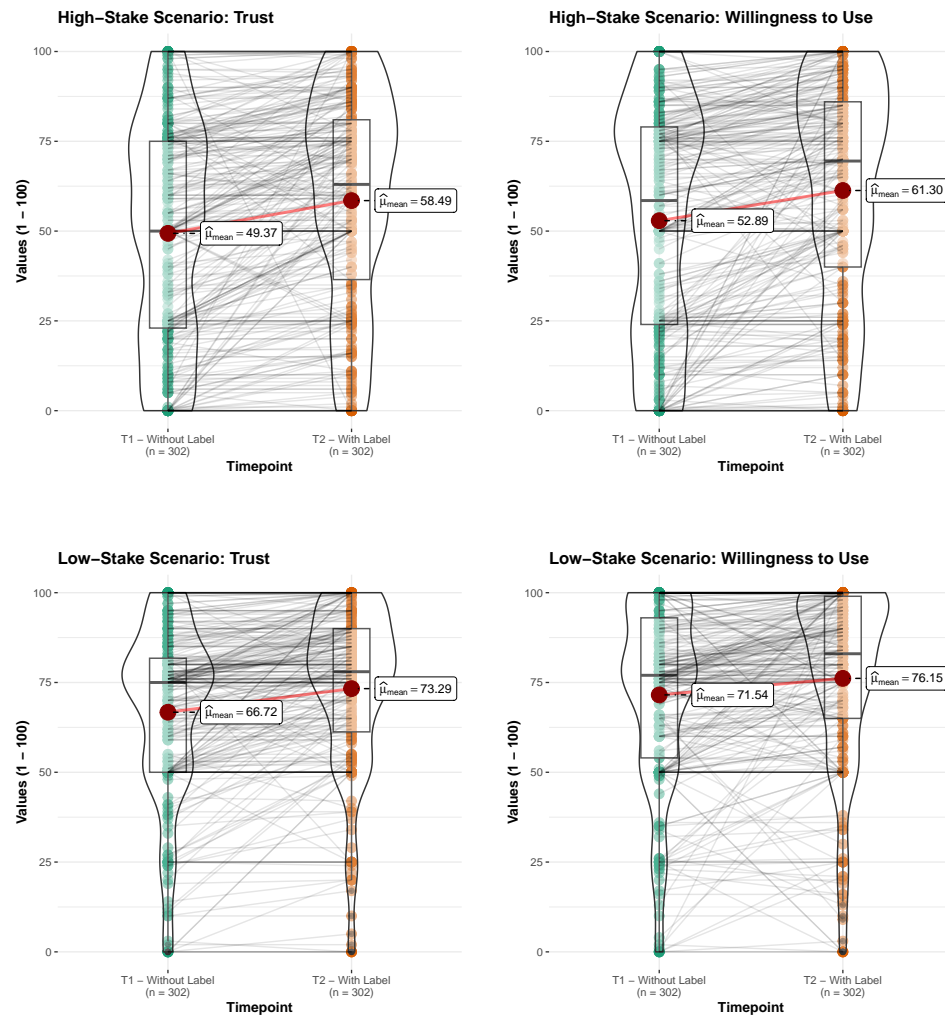
Figure 2: Plots showing the individual scores for trust and willingness to use and their respective changes from T1 (without label) to T2 (with label). The plots also depict the medians, means, and distribution of the aggregated low- and high-stake scenarios. All comparisons revealed statistically significant differences.

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

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

7 DISCUSSION

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

use AI in *both* low- and high-stake scenarios, thereby answering our second research question. Most participants (81%) of the census-representative survey preferred using AI with a certification label, and a large proportion of participants (71%) responded that they would be more likely to accept an AI's decision or recommendation if it had been awarded a certification label. The results further show that a majority of participants (63%) not only indicated a preference for certification labels in high-stake scenarios, but that certification labels also had a larger effect on trust and willingness to use AI in high-stake scenarios. For example, willingness to use ratings for the "hiring procedure" scenario increased from 36 to 64 points, compared to an increase from 75 to 80 points for the "price comparison" scenario. While Stuurman and Lachaud and the EU's "white paper on artificial intelligence" distinguish between regulating high-stake
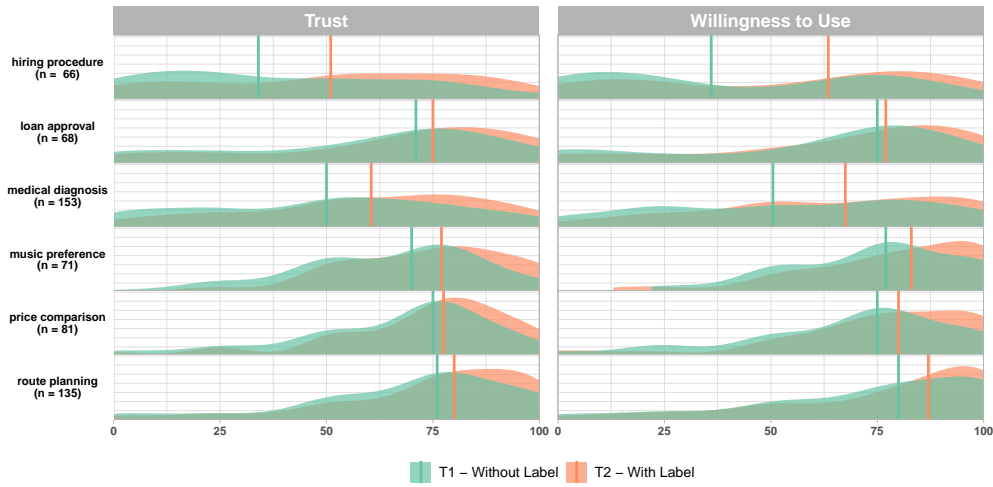
Figure 3: Plots showing the different distributions for trust and willingness to use ratings for the different high-stake (hiring procedure, loan approval, medical diagnosis) and low-stake (music preference, price comparison, route planning) without a label at T1 and with a label at T2.

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

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

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

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

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

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

8 LIMITATIONS AND FUTURE WORK

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

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

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

9 CONCLUSION

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

10 FUNDING, DECLARATION OF CONFLICTING INTERESTS AND DATA AVAILABILITY

REFERENCES

[1] M. Arnold, R. K. E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilović, R. Nair, K. Natesan Ramamurthy, A. Olteanu, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, and K. R. Varshney. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4/5 (2019), 6:1–6:13. https://doi.org/10.1147/JRD.2019.2942288

[2] Shahar Avin, Haydn Belfield, Miles Brundage, Gretchen Krueger, Jasmine Wang, Adrian Weller, Markus Anderljung, Igor Krawczuk, David Krueger, Jonathan Lebensold, et al. 2021. Filling gaps in trustworthy development of AI. *Science* 374, 6573 (2021), 1327–1329. https://doi.org/10.1126/science.abi7176

[3] Jack Bandy. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 74 (apr 2021), 34 pages. https://doi.org/10.1145/3449148

[4] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yun-han Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. 2020. Explainable Machine Learning in Deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20). Association for Computing Machinery, New York, NY, USA, 648–657. https://doi.org/10.1145/3351095.3375624

[5] Reuben Binns, Max van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage'. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, Regan Mandryk, Mark Hancock, Mark Perry, and Anna Cox (Eds.). ACM, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173951

[6] Florian Brühlmann, Serge Petralito, Lena Aeschbach, and Klaus Opwis. 2020. The Quality of Data Collected Online: An Investigation of Careless Responding in a Crowdsourced Sample. *Methods in Psychology* 2 (2020), 100022. https://doi.org/10.1016/j.metip.2020.100022

[7] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian K. Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, Tegan Maharaj, Pang Wei Koh, Sara Hooker, Jade Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O'Keefe, Mark Koren, Théo Ryffel, J. B. Rubinovitz, Tamay Besiroglu, Federica Carugati, Jack Clark, Peter Eckersley, Sarah de Haas, Maritza Johnson, Ben Laurie, Alex Ingerman, Igor Krawczuk, Amanda Askell, Rosario Cammarota, Andrew Lohn, David Krueger, Charlotte Stix, Peter Henderson, Logan Graham, Carina Prunkl, Bianca Martin, Elizabeth Seger, Noa Zilberman, Seán Ó hÉigeartaigh, Frens Kroeger, Girish Sastry, Rebecca Kagan, Adrian Weller, Brian Tse, Elizabeth Barnes, Allan Dafoe, Paul Scharre, Ariel Herbert-Voss, Martijn Rasser, Shagun Sodhani, Carrick Flynn, Thomas Krendl Gilbert, Lisa Dyer, Saif Khan, Yoshua Bengio, and Markus Anderljung. 2020. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. *CoRR* abs/2004.07213 (2020). arXiv:2004.07213 https://arxiv.org/abs/2004.07213

[8] Christiano Castelfranchi and Rino Falcone. 2010. *Trust theory: A socio-cognitive and computational model.* John Wiley & Sons. https://doi.org/10.1002/9780470519851

[9] Erin K. Chiou and John D. Lee. 2021. Trusting Automation: Designing for Responsivity and Resilience. *Human Factors: The Journal of Human Factors and Ergonomics Society* 65 (2021), 137 – 165. https://doi.org/10.1177/00187208211009995

[10] European Comission. 2022. *EU Ecolabel facts and figures.* Retrieved February 2, 2023 from https://environment.ec.europa.eu/topics/circular-economy/eu-ecolabel-home/business/ecolabel-facts-and-figures_en

[11] European Commission. 2020. White Paper on Artificial Intelligence: a European approach to excellence and trust. *Official Journal of European Union* L COM(2020) 65 final (19-02- 2020). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0065&qid=1675254609974

[12] European Commission. 2021. Proposal for a Regulation of the European Parliament and of the council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. *Official Journal of European Union* L COM (2021) 206 final (21-04- 2021). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

[13] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. 2022. Who Audits the Auditors? Recommendations from a Field Scan of the Algorithmic Auditing Ecosystem. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1571–1583. https://doi.org/10.1145/3531146.3533213

[14] Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. 2022. Interactive Model Cards: A Human-Centered Approach to Model Documentation. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 427–439. https://doi.org/10.1145/3531146.3533108

[15] Denmark. 2019. National Strategy for Artificial Intelligence. *Ministry of Finance and Ministry of Industry Business Financial Affairs* (March 2019). https://eng.em.dk/media/13081/305755-gb-version_4k.pdf

[16] Group Ecocert. 2018. *Organic agriculture Europe.* Group Ecocert. Retrieved Jan 14, 2023 from https://www.ecocert.com/en/certification-detail/organic-farming-europe-eu-n-848-2018

[17] Gregory Falco, Ben Shneiderman, Julia Badger, Ryan Carrier, Anton Dahbura, David Danks, Martin Eling, Alwyn Goodloe, Jerry Gupta, Christopher Hart, et al. 2021. Governing AI safety through independent audits. *Nature Machine Intelligence* 3, 7 (2021), 566–571. https://doi.org/10.1038/s42256-021-00370-7

[18] Luciano Floridi, Matthias Holweg, Mariarosaria Taddeo, Javier Amaya Silva, Jakob Mökander, and Yuni Wen. 2022. capAI-A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act. *Available at SSRN: https://ssrn.com/abstract=4064091* (March 23, 2022). https://dx.doi.org/10.2139/ssrn.4064091

[19] Fraunhofer Institute for Telecommunications and HHI Heinrich Hertz Institute. [n.d.]. *Auditing and Certification of AI Systems.* Retrieved Jan 25 2023 from https://www.hhi.fraunhofer.de/en/departments/ai/technologies-and-solutions/auditing-and-certification-of-ai-systems.html

[20] ForHumanity. 2016. *Independent Audit of AI Systems (IAAIS).* ForHumanity. Retrieved Jan 25 2023 from https://forhumanity.center/independent-audit-of-ai-systems/

[21] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (nov 2021), 86–92. https://doi.org/10.1145/3458723

[22] Samantha Goodman, Lana Vanderlee, Rachel Acton, Syed Mahamad, and David Hammond. 2018. The impact of front-of-package label design on consumer understanding of nutrient amounts. *Nutrients* 10, 11 (2018), 1624. https://doi.org/10.3390/nu10111624

[23] Matthew Gorton, Barbara Tocco, Ching-Hua Yeh, and Monika Hartmann. 2021. What determines consumers' use of eco-labels? Taking a close look at label trust. *Ecological Economics* 189 (2021), 107173. https://doi.org/10.1016/j.ecolecon.2021.107173

[24] Klaus G Grunert, Sophie Hieke, and Josephine Wills. 2014. Sustainability labels on food products: Consumer motivation, understanding and use. *Food policy* 44 (2014), 177–189. https://doi.org/10.1016/j.foodpol.2013.12.001

[25] Sebastian Hallensleben, Carla Hustedt, Lajla Fetic, Torsten Fleischer, Paul Grünke, Thilo Hagendorff, Marc Hauer, Andreas Hauschke, Jessica Heesen, Michael Herrmann, Rafaela Hillerbrand, Christoph Hubig, Andreas Kaminski, Tobias Krafft, Wulf Loh, Philipp Otto, and Michael Puntschuh. 2020. From Principles to Practice – An interdisciplinary framework to operationalise AI ethics. *Artificial Intelligence Ethics Impact Group* (01 April 2020). https://www.ai-ethics-impact.org/en

[26] Rick Harbaugh, John W Maxwell, and Beatrice Roussillon. 2011. Label confusion: The Groucho effect of uncertain standards. *Management science* 57, 9 (2011), 1512–1527. https://doi.org/10.1287/mnsc.1110.1412

[27] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2020. The dataset nutrition label. *Data Protection and Privacy, Volume 12: Data Protection and Democracy* 12 (2020), 1. https://doi.org/10.48550/arXiv.1805.03677

[28] Ming-Hui Huang and Roland T. Rust. 2018. Artificial Intelligence in Service. *Journal of Service Research* 21, 2 (2018), 155 – 172. https://doi.org/10.1177/1094670517752459

[29] IEEE. 2008. IEEE Standard for Software Reviews and Audits. *IEEE Std 1028-2008* vol., no. (15 Aug. 2008), 1–53. https://doi.org/10.1109/IEEESTD.2008.4601584

[30] Iina Ikonen, Francesca Sotgiu, Aylin Aydinli, and Peeter WJ Verlegh. 2020. Consumer effects of front-of-package nutrition labeling: An interdisciplinary meta-analysis. *Journal of the Academy of Marketing Science* 48, 3 (2020), 360–383. https://doi.org/10.1007/s11747-019-00663-9

[31] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. ACM, New York, NY, USA, 624–635. https://doi.org/10.1145/3442188.3445923

[32] Maurice Jakesch, Zana Buçinca, Saleema Amershi, and Alexandra Olteanu. 2022. How Different Groups Prioritize Ethical Values for Responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 310–323. https://doi.org/10.1145/3531146.3533097

[33] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1 (2019), 389–399. https://doi.org/10.1038/s42256-019-0088-2

[34] Alexandra Jones, Bruce Neal, Belinda Reeve, Cliona Ni Mhurchu, and Anne Marie Thow. 2019. Front-of-pack nutrition labelling to promote healthier diets: current practice and opportunities to strengthen regulation worldwide. *BMJ global health* 4, 6 (2019), e001882. https://doi.org/10.1136/bmjgh-2019-001882

[35] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. "Because AI is 100% Right and Safe": User Attitudes and Sources of AI Authority in India. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 158, 18 pages. https://doi.org/10.1145/3491102.3517533

[36] Davinder Kaur, Suleyman Uslu, Kaley J. Rittichier, and Arjan Durresi. 2022. Trustworthy Artificial Intelligence: A Review. *ACM Comput. Surv.* 55, 2, Article 39 (jan 2022), 38 pages. https://doi.org/10.1145/3491209

[37] Bran Knowles and John T. Richards. 2021. The Sanction of Authority: Promoting Public Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 262–271. https://doi.org/10.1145/3442188.3445890

[38] P. M. Krafft, Meg Young, Michael Katell, Jennifer E. Lee, Shankar Narayan, Micah Epstein, Dharma Dailey, Bernease Herman, Aaron Tam, Vivian Guetler, Corinne Bintz, Daniella Raz, Pa Ousman Jobe, Franziska Putz, Brian Robick, and Bissan Barghouti. 2021. An Action-Oriented AI Policy Toolkit for Technology Audits by Community Advocates and Activists. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 772–781. https://doi.org/10.1145/3442188.3445938

[39] Markus Langer, Daniel Oster, Timo Speith, Lena Kästner, Holger Hermanns, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What Do We Want From Explainable Artificial Intelligence (XAI)? A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research. *Artificial Intelligence* 296 (Feb. 2021), 103473. https://doi.org/10.1016/j.artint.2021.103473

[40] John D. Lee and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (March 2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[41] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex 'Sandy' Pentland, and Patrick Vinck. 2018. Fair, Transparent, and Accountable Algorithmic Decision-making Processes. *Philosophy & Technology* 31 (2018), 611–627. https://doi.org/10.1007/s13347-017-0279-x

[42] Lan Li, Tina Lassiter, Joohee Oh, and Min Kyung Lee. 2021. Algorithmic Hiring in Practice: Recruiter and HR Professional's Perspectives on AI Use in Hiring. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 166–176. https://doi.org/10.1145/3461702.3462531

[43] 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* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1257–1268. https://doi.org/10.1145/3531146.3533182

[44] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. ACM, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376590

[45] Consumer Research Associates Ltd. 2007. *EFTA study on certification and marks in Europe.* Retrieved May 4, 2023 from https://www.efta.int/sites/default/files/publications/study-certification-marks/executive-summary.pdf

[46] Malta. 2019. Malta the ultimate AI Launchpad: a strategy and vision for Artificial Intelligence in Malta 2030. *Parliamentary Secretariat for Financial Services Digital Economy Innovation* (October 2019). https://malta.ai/wp-content/uploads/2019/11/Malta_The_Ultimate_AI_Launchpad_vFinal.pdf

[47] Philipp Mayring and Thomas Fenzl. 2019. Qualitative Inhaltsanalyse. In *Handbuch Methoden der empirischen Sozialforschung*, N. Baur and J. Blasius (Eds.). Springer VS, Wiesbaden, (pp. 633–648). https://doi.org/10.1007/978-3-658-21308-4_42

[48] Adam W Meade and S Bartholomew Craig. 2012. Identifying careless responses in survey data. *Psychological methods* 17, 3 (2012), 437. https://doi.org/10.1037/a0028085

[49] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596

[50] Birthe Nesset, Gnanathusharan Rajendran, José David Aguas Lopes, and Helen Hastie. 2022. Sensitivity of Trust Scales in the Face of Errors. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction* (Sapporo, Hokkaido, Japan) *(HRI '22)*. IEEE Press, 950–954. https://doi.org/10.1109/HRI53351.2022.9889427

[51] Indrajeet Patil. 2021. statsExpressions: R package for tidy dataframes and expressions with statistical details. *Journal of Open Source Software* 6, 61 (2021), 3236. https://doi.org/10.21105/joss.03236

[52] Stefano Puntoni, Rebecca Walker Reczek, Markus Giesler, and Simona Botti. 2020. Consumers and Artificial Intelligence: An Experiential Perspective. *Journal of Marketing* 85, 1 (2020), 131 – 151. https://doi.org/10.1177/0022242920953847

[53] R Core Team. 2022. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[54] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 33–44. https:

//doi.org/10.1145/3351095.3372873

[55] Shlomo S Sawilowsky. 2009. New effect size rules of thumb. *Journal of modern applied statistical methods* 8, 2 (2009), 467–474. https://doi.org/10.22237/jmasm/1257035100

[56] Nicolas Scharowski, Sebastian AC Perrig, Nick von Felten, and Florian Brühlmann. 2022. Trust and Reliance in XAI–Distinguishing Between Attitudinal and Behavioral Measures. *CHI TRAIT Workshop* (2022), 6 pages. https://doi.org/10.48550/arXiv.2203.12318

[57] Nadine Schlicker, Alarith Uhde, Kevin Baum, Martin C Hirsch, and Markus Langer. 2022. Calibrated Trust as a Result of Accurate Trustworthiness Assessment–Introducing the Trustworthiness Assessment Model. (2022). https://doi.org/10.31234/osf.io/qhwvx

[58] Christin Seifert, Stefanie Scherzinger, and Lena Wiese. 2019. Towards Generating Consumer Labels for Machine Learning Models. In *2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI)* (Los Angeles, CA, USA). IEEE, 173–179. https://doi.org/10.1109/CogMI48466.2019.00033

[59] Marcel Stadelmann and Renate Schubert. 2018. How do different designs of energy labels influence purchases of household appliances? A field study in Switzerland. *Ecological economics* 144 (2018), 112–123. https://doi.org/10.1016/j.ecolecon.2017.07.031

[60] Bernd Carsten Stahl and David Wright. 2018. Ethics and privacy in AI and big data: Implementing responsible research and innovation. *IEEE Security & Privacy* 16, 3 (2018), 26–33. https://doi.org/10.1109/MSP.2018.270116

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

[62] Khan MR Taufique, Kristian S Nielsen, Thomas Dietz, Rachael Shwom, Paul C Stern, and Michael P Vandenbergh. 2022. Revisiting the promise of carbon labelling. *Nature Climate Change* 12, 2 (2022), 132–140. https://doi.org/10.1038/s41558-021-01271-8

[63] Frauke Mattison Thompson, Sven Tuzovic, and Corina Braun. 2019. Trustmarks: Strategies for exploiting their full potential in e-commerce. *Business Horizons* 62, 2 (2019), 237–247. https://doi.org/10.1016/j.bushor.2018.09.004

[64] Emma Tonkin, Annabelle M Wilson, John Coveney, Trevor Webb, and Samantha B Meyer. 2015. Trust in and through labelling–a systematic review and critique. *British Food Journal* 117, 1 (2015), 318–338. https://doi.org/10.1108/BFJ-07-2014-0244

[65] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad van Moorsel. 2020. The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 272–283. https://doi.org/10.1145/3351095.3372834

[66] Briana Vecchione, Karen Levy, and Solon Barocas. 2021. Algorithmic Auditing and Social Justice: Lessons from the History of Audit Studies. In *Equity and Access in Algorithms, Mechanisms, and Optimization* (–, NY, USA) *(EAAMO '21)*. Association for Computing Machinery, New York, NY, USA, Article 19, 9 pages. https://doi.org/10.1145/3465416.3483294

[67] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5 (2021), 1 – 39. https://doi.org/10.1145/3476068

[68] Ruotong Wang, F. Maxwell Harper, and Haiyi Zhu. 2020. Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376813

[69] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. 2021. Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 666–677. https://doi.org/10.1145/3442188.3445928

[70] Sarah Woods, Michael Walters, Kheng Lee Koay, and Kerstin Dautenhahn. 2006. Comparing human robot interaction scenarios using live and video based methods: towards a novel methodological approach. In *9th IEEE International Workshop on Advanced Motion Control, 2006.* (Istanbul, Türkiye). IEEE, 750–755. https://doi.org/10.1109/AMC.2006.1631754

[71] Kun-Hsing Yu, Andrew Beam, and Isaac S. Kohane. 2018. Artificial intelligence in healthcare. *Nature Biomedical Engineering* 2 (2018), 719–731. https://doi.org/10.1038/s41551-018-0305-z

[72] Mireia Yurrita, Dave Murray-Rust, Agathe Balayn, and Alessandro Bozzon. 2022. Towards a Multi-Stakeholder Value-Based Assessment Framework for Algorithmic Systems. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 535–563. https://doi.org/10.1145/3531146.3533118

A APPENDIX

A.1 High-stake Scenarios

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

A.1.2 Hiring Procedure. Consider the situation where you are applying for a new job at a company. The company is using an AI system called MyJob for evaluating job applications. You will be required to fill out a form, uploading your CV, and submit them along with personal information like address, marital status, employment status and references to MyJob. Once assessed, MyJob will determine based on the provided information whether or not you will be invited for an interview.

A.1.3 Loan Approval. Consider the situation where you are applying for a loan at a bank. The bank is using an AI system called MyLoans for evaluating loan applications. You will be required to fill out a form, specifying the loan amount, and submit them along with personal information like marital status, employment status, annual income and financial history to MyLoans. Once assessed, MyLoans will determine based on the provided information whether your loan application is successful or not.

A.2 Low-stake Scenarios

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

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

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