Certification Labels for Trustworthy AI

Nicolas Scharowski[1], Michaela Benk[2], Swen J. Kühne[3], Léane Wettstein[1] and Florian Brühlmann[1]

[1]University of Basel, Center for General Psychology and Methodology
[2]ETH Zurich, Mobiliar Lab for Analytics
[3]Zurich University of Applied Sciences, School of Applied Psychology

Abstract

This study is the first to empirically investigate the use of *certification labels* as a solution to communicating the trustworthiness of AI to end-users. Through interviews ($N = 12$) and a Swiss census-representative survey ($N = 302$) we investigated attitudes towards certification labels and their effectiveness in conveying trustworthiness in low- and high-stakes AI scenarios. The results showed that certification labels can significantly increase end-users' trust and willingness to use AI, particularly in high-stakes scenarios. However, the study also highlighted opportunities and limitations in addressing end-users concerns regarding the use of AI. The research provides insights and recommendations for designing and implementing certification labels as a promising constituent of trustworthy AI.

Keywords

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

1. Introduction

In recent years, the use of artificial intelligence (AI) has increased dramatically in various sectors of society and profoundly impacted human lives. This rise of AI has led to increased discussions with various institutions, and communities engaged in a discourse about how to ensure trustworthy AI. As a result, various principles on trustworthy AI have been identified, such as mitigating bias and unfairness in AI systems, explaining AI decisions, and ensuring user privacy [1, 2, 3, 4]. Despite these efforts to design trustworthy AI, the challenge of conveying trustworthiness to different stakeholders remains. Especially end-users (i.e., laypersons in data science or machine learning that may be directly or indirectly affected by AI decisions) may not be in a position to trust AI as they lack the necessary knowledge to assess AI trustworthiness criteria (e.g., privacy, fairness, robustness) [5]. To address this challenge, recent research has focused on the pivotal role of auditability as a key enabler of trust in AI and the importance of creating an "AI trustworthiness ecosystem."

However, current research mainly focuses on auditable documentation like model cards and datasheets [6, 7]. While AI documentations are valuable artifacts to inform auditors in their decision-making, they are tailored to experts and regulators, not end-users. Therefore, the challenge remains how to effectively communicate to end-users that an AI auditing process has

EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland
nicolas.scharowski@unibas.ch (N. Scharowski)
0000-0001-5983-346X (N. Scharowski); 0000-0002-8171-320X (M. Benk); 0000-0001-8326-9168 (S. J. Kühne); 0009-0003-5068-7007 (L. Wettstein); 0000-0001-8945-3273 (F. Brühlmann)

deemed the AI trustworthy. This work addressed this challenge by focusing on communicating the outcomes of auditing processes to end-users using *certification labels*, commonly used in other domains to certify that a product meets certain criteria and promotes trustworthiness [8]. Certification labels can be designed to be accessible to end-users, and if reflecting a credible auditing process, they can serve as a trustworthiness cue for end-users [9]. However, end-users' attitudes toward AI certification labels and their effectiveness in communicating AI trustworthiness remain unknown. This study was the first to empirically investigate this.

2. Method

To explore the effectiveness of certification labels in communicating AI trustworthiness, we conducted a mixed-method study with both interviews ($N = 12$) and a Swiss census-representative survey ($N = 302$) with end-users. The final sample for the interviews consisted of students (P2, P3, P4, P8, P11), 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. Following Kapania et al., we used low-stake (music preference, route planning, price comparison) and high-stake (medical diagnosis, hiring procedure, loan approval) AI scenarios and measured both *trust* and *willingness to use AI* before and after presenting end-users with a certification label. The certification label used in this study was an existing label developed by the Swiss Digital Initative (SDI) for the broader context of digital trust.

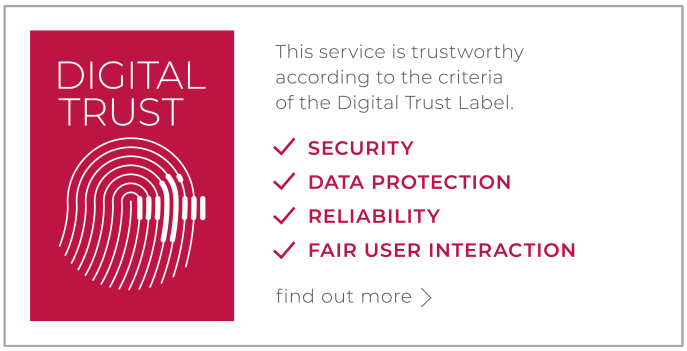

Figure 1: The "Digital Trust Label" from the non-profit foundation Swiss Digital Initiative, which we adapted for this study in the context of certification labels for AI as stimuli. ©Swiss Digital Initiative

The label is based on a catalog of verifiable and auditable criteria, co-developed by an academic expert group based on a user study on digital trust. Independent third-party audits are conducted following the catalog, and if all criteria are fulfilled, the label is awarded to the service. The audits are conducted following a catalog containing, at the time of the study, 35 criteria organized in 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.

Supplementary materials, the data, and corresponding R-scripts are available on OSF: https://osf.io/gzp5k/?view_only=709e50a07d2f46c3a10474f1d125b32f.

3. Results

The results of our study, from both the interviews and the survey, suggest that certification labels can effectively communicate the trustworthiness of AI. Quantitative findings of the census-representative survey demonstrate that presenting end-users a certification label significantly increases end-users' trust and willingness to use AI in both low- and high-stake scenarios (see Appendix). End-users were found to have a higher preference for certification labels in high-stake scenarios, and the impact of a certification label on trust and willingness to use AI was more pronounced in high-stakes 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, 12].

Qualitative findings of the interviews show that end-users have positive views of AI certification labels and that they provide the opportunity to increase trust, perceived transparency and fairness, and provide standardization for AI. Furthermore, participants indicated that certification labels can mitigate their data-related concerns regarding privacy and data protection. Certification labels were perceived by participants as effective tools for addressing their data-related concerns, by holding certified parties accountable for complying with the standards set in the catalog. For participants, the standards set by the certification label regarding security and data protection represent the acceptable minimum to consider using an AI system. However, some limitations concerning the use of AI remain (see Appendix). The interviews also provide inhibitors and facilitators for the effective use of certification labels in the context of AI. End-users expressed a preference for independent audits and the difficulty of communicating subjective criteria such as "fairness," the meaning of which can be ambiguous. Certification labels may not address all end-user's concerns (e.g., AI performance measures) and should be considered one component of a larger effort to ensure trustworthy AI.

4. Discussion

Our study demonstrates the potential of certification labels as a promising approach to communicating AI trustworthiness to end-users. The quantitative results showed that certification labels can significantly increase both trust and willingness to use AI in low- and high-stake 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.

References

[1] B. Lepri, N. Oliver, E. Letouzé, A. S. Pentland, P. Vinck, Fair, transparent, and accountable algorithmic decision-making processes, Philosophy & Technology 31 (2018) 611–627. URL: https://doi.org/10.1007/s13347-017-0279-x.

[2] M. Langer, D. Oster, T. Speith, L. Kästner, H. Hermanns, E. Schmidt, A. Sesing, K. Baum, 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 (2021) 103473. doi:10.1016/j.artint.2021.103473.

[3] D. Kaur, S. Uslu, K. J. Rittichier, A. Durresi, Trustworthy artificial intelligence: A review, ACM Comput. Surv. 55 (2022). URL: https://doi.org/10.1145/3491209. doi:10.1145/3491209.

[4] B. C. Stahl, D. Wright, Ethics and privacy in ai and big data: Implementing responsible research and innovation, IEEE Security & Privacy 16 (2018) 26–33. URL: https://doi.org/10.1109/MSP.2018.270116.

[5] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. F. Moura, P. Eckersley, Explainable machine learning in deployment, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 648–657. URL: https://doi.org/10.1145/3351095.3375624. doi:10.1145/3351095.3375624.

[6] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, T. Gebru, Model cards for model reporting, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 220–229. URL: https://doi.org/10.1145/3287560.3287596. doi:10.1145/3287560.3287596.

[7] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. III, K. Crawford, Datasheets for datasets, Commun. ACM 64 (2021) 86–92. URL: https://doi.org/10.1145/3458723. doi:10.1145/3458723.

[8] E. Tonkin, A. M. Wilson, J. Coveney, T. Webb, S. B. Meyer, Trust in and through labelling–a systematic review and critique, British Food Journal 117 (2015) 318–338. URL: https://doi.org/10.1108/BFJ-07-2014-0244.

[9] Q. Liao, S. S. Sundar, Designing for responsible trust in ai systems: A communication perspective, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 1257–1268. URL: https://doi.org/10.1145/3531146.3533182. doi:10.1145/3531146.3533182.

[10] S. Kapania, O. Siy, G. Clapper, A. M. SP, N. Sambasivan, "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, CHI '22, Association for Computing Machinery, New York, NY, USA, 2022. URL: https://doi.org/10.1145/3491102.3517533. doi:10.1145/3491102.3517533.

[11] E. Commission, White paper on artificial intelligence: a european approach to excellence and trust, Official Journal of European Union L COM(2020) 65 final (2020). URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0065&qid=1675254609974.

[12] K. Stuurman, E. Lachaud, Regulating ai. a label to complete the proposed act on artificial intelligence, Computer Law & Security Review 44 (2022) 105657. URL: https://doi.org/10.1016/j.clsr.2022.105657. doi:10.1016/j.clsr.2022.105657.

A. Quantitative and qualitative Results

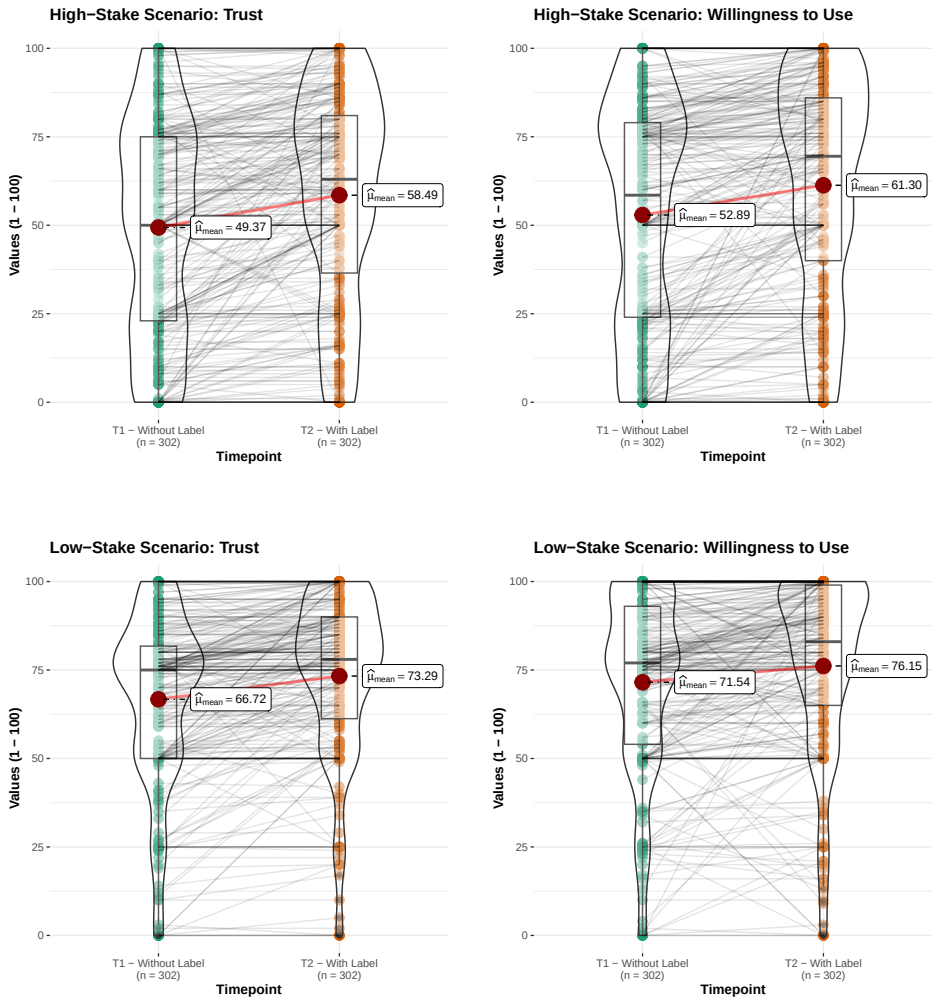

Figure 2: Plots showing the individual scores for trust and willingness to use and their respective changes from T1 (without label) to T2 (with label). The plots also depict the medians, means and distribution of the aggregated low- and high-stake scenarios. All comparisons revealed statistically significant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

Table 1
End-user's 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 4 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)*