Published: 10 November 2025

Citation in BibTeX format

HAI '25: International Conference on Human-Agent Interaction
November 10 - 13, 2025
Yokohama, Japan

POSTER

Instilling (Dis-)Trust in AI Products: Recommendations for the Design of Data Security and Data Privacy Labels

CHRISTINA U PFEUFFER, Catholic University of Eichstätt-Ingolstadt, Eichstatt, Bayern, Germany

Open Access Support provided by:

Catholic University of Eichstätt-Ingolstadt

HAI '25: Proceedings of the 13th International Conference on Human-Agent Interaction (November 2025)
https://doi.org/10.1145/3765766.3765868
ISBN: 9798400721786

Instilling (Dis-)Trust in AI Products: Recommendations for the Design of Data Security and Data Privacy Labels

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

Abstract

Rarely are we fully informed about the data security and data privacy (DSDP) of artificial intelligence (AI) products and services we use. Providing DSDP information on AI products in an easily accessible and quick-to-process format could help instill appropriate levels of (dis-)trust in (potential) users. Here, participants were presented with hypothetical AI products paired with different labels (graphical vs. text-based) conveying low to high DSDP levels. Expectedly, trust increased and anxiety decreased when an AI product reached a higher DSDP level. That is, labels effectively communicated DSDP differences. Text-based labels were associated with increased trust and decreased anxiety compared to graphical labels. Interestingly, when not provided with DSDP information via a label, participants attributed an intermediate level of (dis-)trust to AI products. These findings illustrate the importance and potential of introducing easy-to-process labels to convey information about AI products, for instance, DSDP information.

CCS Concepts

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

Keywords

artificial intelligence, data security and data privacy, label, regulation, trust, AI anxiety

ACM Reference Format:
Christina U. Pfeuffer. 2025. Instilling (Dis-)Trust in AI Products: Recommendations for the Design of Data Security and Data Privacy Labels. In *13th International Conference on Human-Agent Interaction (HAI '25), November 10–13, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3765766.3765868

1 Introduction

Artificial intelligence (AI) and corresponding AI products hold the potential to benefit both individuals, organizations, and society at large by, for instance, optimizing products and services, enhancing productivity and efficiency, or lowering costs [11]. This potential can only be realized when human-AI interactions are appropriately shaped [2, 10]. Concerns regarding AI trustworthiness, in particular, data security and data privacy concerns [9, 11], jeopardize a further widespread acceptance and broader adoption of AI products (see e.g., [6, 18, 23, 24], for prominent theories of technology acceptance and adoption). Recent theorizing emphasizes especially the role of trust (e.g., [25], linked to transparency and derived from a trustworthiness assessment [21]) as an essential precursors of technology acceptance and adoption. As such, establishing the public's trust in AI appears paramount to its further acceptance and adoption.

Users, however, are hardly able to evaluate the trustworthiness of AI accurately [11, 21], as corresponding information is commonly not easily accessible. They therefore (dis-)trust mainly based on heuristics [3, 16, 17] and strong, often unjustified AI endorsement [11], is coupled with low understanding of AI in the general public [11, 15]. Discrepancies between objective trustworthiness (e.g., adherence to criteria like those proposed by the European Commission [8, 9]) and how trustworthy individuals perceive AI to be call for corresponding affirmative action. Both misplaced distrust [5, 25] and misplaced trust (due to expectancy violations, [13, 19]) prevent the further acceptance and adoption of new (and trustworthy) AI technologies and obstruct corresponding benefits of AI usage. I propose that informative, multi-level labels (e.g., similar to the Nutri-Score indicating the nutritional value of food, e.g., [16]; for prior studies on technology/AI certification labels see [1, 12, 20, 27]) constitute the best-suited means of achieving accurate assessments of AI trustworthiness with very limited (potential) user effort across varying levels of AI literacy.

Here, I communicated the data security and data privacy (DSDP; i.e., AI trustworthiness criteria) level of hypothetical AI products using three-level labels (graphical vs. text-based label). I expected trust and attributed monetary value to increase and AI anxiety to decrease for AI products with higher DSDP levels communicated by a corresponding DSDP/trustworthiness label. Furthermore, I expected to observe differences between the two label types.

2 Experimental Methods

An extended preprint (https://osf.io/preprints/psyarxiv/q25nr_v1; see for extended descriptions), a preregistration (https://osf.io/vbxqy), and all study materials (https://doi.org/10.17605/OSF.IO/HD3NA) are available online.

102 participants (35 male, 64 female, 3 diverse; age: M = 26.7 years, SD = 8.9; attitude towards technology [5]: M = 14.4, SD = 2.86, [4;20]) took part after providing informed consent. First, participants were informed about the features and functions of

ACM ISBN 979-8-4007-2178-6/2025/11
https://doi.org/10.1145/3765766.3765868

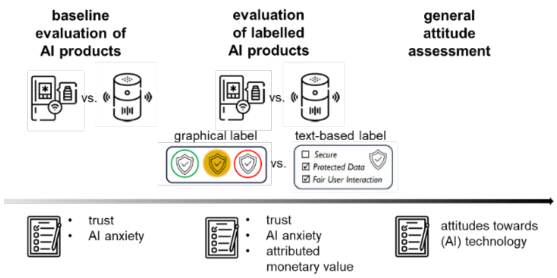

Figure 1: Study Design and Time Course

two hypothetical AI product types (smart fridge, voice assistant). They then rated their trust (4 items [5]; 1 = strongly disagree to 5 = strongly agree) and (state) anxiety (4 items [26]; 1 = strongly disagree to 5 = strongly agree) regarding each AI product (represented by an icon), first, when presented without further information (baseline) and, second (after an introduction of the labels; criteria adapted from [22]; compare [8, 9]), when presented with a DSDP label (label type: graphical vs. text-based; within) indicating a low, intermediate, or high level of trustworthiness (DSDP level; within; see Figure 1 for experimental design and procedure). Then, I assessed the monetary value participants attributed to the respective labelled AI products by showing two different levels of the same label type per AI product and trial (level comparison: low-intermediate vs. intermediate-high vs. low-high; within) and asking how much more (% price) participants were willing to pay for the AI product with the higher DSDP level. Participants then rated their attitude towards (AI) technology ([5]; 1 = strongly disagree to 5 = strongly agree) and were debriefed.

3 RESULTS

A Bayesian linear mixed model analysis approach (criterion: $BF_{10} > 3$ or $< 1/3$) was used. To account for differences between a person's ratings of the respective AI product type at baseline (i.e., without a DSDP label) and when presented with a DSDP label, I analyzed corresponding difference scores (trust/AI anxiety condition – trust/AI anxiety baseline).

Baseline. Trust at baseline was 9.8 (SD = 2.7; [0;20])/10.8 (SD = 3.0) for the AI voice assistant/smart fridge and AI anxiety at baseline was 13.1 (SD = 3.7; [0;20])/12.5 (SD = 3.5) for the AI voice assistant/smart fridge.

Trust. Trust ratings increased with increasing DSDP levels, $BF_{10} = 1.36 \times 10^{44} \pm 1.16\%$ (see Figure 2, left). Moreover, trust ratings were higher for text-based as compared to graphical labels, $BF_{10} = 7.18 \times 10^{7} \pm 0.88\%$. Label type and DSDP level interacted, $BF_{10} = 4.03 \pm 1.61\%$.

AI Anxiety. AI anxiety ratings decreased with increasing DSDP levels, $BF_{10} = 8.4 \times 10^{25} \pm 1.76\%$ (see Figure 2, middle). AI anxiety ratings were lower for text-based as compared to graphical labels, $BF_{10} = 9.5 \pm 1.75\%$. There was evidence against an interaction of label type and DSDP level, $BF_{10} = 0.1 \pm 2.23\%$.

Attributed Value. Attributed monetary value (acceptable percentage of price increase for a higher DSDP level) increased across DSDP level comparisons, $BF_{10} = 5.8 \times 10^{29} \pm 1.26\%$ (see Figure 2, right). Higher monetary value was attributed to AI products labelled with graphical as compared to text-based labels, $BF_{10} = 8.1 \pm 1.15\%$. There was inconclusive evidence against an interaction of label type and DSDP level comparison, $BF_{10} = 0.44 \pm 1.81\%$.

4 DISCUSSION

Participants' trust and AI anxiety as well as the monetary value they attributed to AI products scaled with the DSDP label level (low vs. intermediate vs. high). This shows that DSDP labels effectively communicated AI trustworthiness, affecting (potential) user's perception and evaluation of AI products. Importantly, trust and AI anxiety ratings were baseline-adjusted (i.e., a value of 0 corresponded to a participant's respective baseline rating). This comparison of labelled AI products against the baseline revealed that trust and AI anxiety ratings at baseline corresponded to ratings for AI products labelled with an intermediate DSDP level. It thus appears that participants unjustifiedly attributed an intermediate

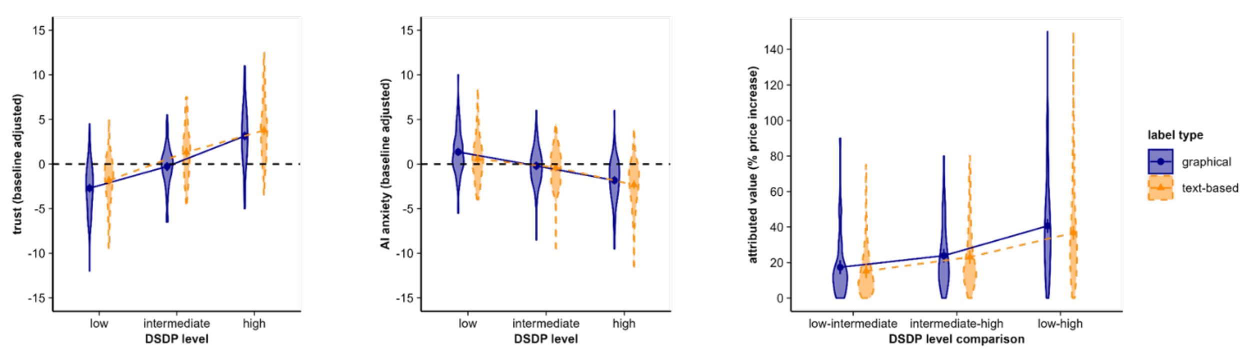

Figure 2: Effects of Data Security and Data Privacy (DSDP) Level/Level Comparison and Label Type on Trust, AI Anxiety, and Attributed Monetary Value. Trust and AI anxiety scores are displayed relative to a participant's respective baseline rating of the corresponding AI product (0 = rating equivalent to baseline). Violins around the respective mean depict the corresponding rating distribution per condition.

DSDP level to AI products in the absence of DSDP information. These findings underscore the importance of introducing corresponding DSDP labels for AI products to prevent both unjustified trust and unjustified distrust.

Moreover, 'AI products with text-based labels were associated with higher trust and lower AI anxiety than graphical labels, whereas AI products with graphical labels were attributed higher monetary value. Thus, text-based labels are better suited to increase trust [5, 7, 25] and thereby the acceptance and adoption of AI, whereas graphical labels might better serve to make AI DSDP/trustworthiness labels more appealing to AI companies and can be processed faster by (potential) users.

Future research will, for instance, need to incorporate further trustworthiness criteria (e.g., [8]), select more informed thresholds for AI trustworthiness levels, assess the potential of combined label types, and account for label effects at different AI literacy levels (e.g., [4]).

Acknowledgments

I thank Anja Ruff for her help with material preparation as well as setting up and advertising the online study.

References

[1] Martin Adam, Sebastian Lins, Ali Sunyaev, and Alexander Benlian. 2024. The Contingent Effects of IS Certifications on the Trustworthiness of Websites. *Journal of the Association for Information Systems* 25, 3 (2024), 594–617. https://doi.org/10.17705/1jais.00836

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

[3] Zana Buçinca, Maja B. Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 1–21. https://doi.org/10.1145/3449287

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

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

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

[7] Massilva Dekkal, Manon Arcand, Sandrine Prom Tep, Lova Rajaobelina, and Line Ricard. 2024. Factors affecting user trust and intention in adopting chatbots: the moderating role of technology anxiety in insurtech. *Journal of Financial Services Marketing* 29, 3 (September 2024), 699–728. https://doi.org/10.1057/s41264-023-00230-y

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

[9] European Parliament and Council of the European Union. 2021. Regulation (EU) 2021/206 of 21 April 2021 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri$=$CELEX%3A32022R0206

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

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

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

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

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

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

[16] Q. Vera Liao and S. Shyam Sundar. 2022. Designing for Responsible Trust in AI Systems: A Communication Perspective. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, June 21, 2022. 1257–1268. https://doi.org/10.1145/3531146.3533182

[17] Zhuoran Lu and Ming Yin. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, May 06, 2021. ACM, Yokohama Japan, 1–16. https://doi.org/10.1145/3411764.3445562

[18] Nikola Marangunić and Andrina Granić. 2015. Technology acceptance model: a literature review from 1986 to 2013. *Universal Access in the Information Society* 14, 1 (March 2015), 81–95. https://doi.org/10.1007/s10209-014-0348-1

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

[20] Nicolas Scharowski, Michaela Benk, Swen J. Kühne, Léane Wettstein, and Florian Brühlmann. 2023. Certification Labels for Trustworthy AI: Insights From an Empirical Mixed-Method Study. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, June 12, 2023. ACM, Chicago IL USA, 248–260. https://doi.org/10.1145/3593013.3593994

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

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

[23] Viswanath Venkatesh, Michael G. Morris, Gordon B. Davis, and Fred D. Davis. 2003. User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly* 27, 3 (2003), 425. https://doi.org/10.2307/30036540

[24] Viswanath Venkatesh and Fred D. Davis. 2000. A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. *Management Science* 46, 2 (February 2000), 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926

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

[26] Yu-Yin Wang and Yi-Shun Wang. 2022. Development and validation of an artificial intelligence anxiety scale: an initial application in predicting motivated learning behavior. *Interactive Learning Environments* 30, 4 (April 2022), 619–634. https://doi.org/10.1080/10494820.2019.1674887

[27] Magdalena Wischnewski, Nicole Krämer, Christian Janiesch, Emmanuel Müller, Theodor Schnitzler, and Carina Newen. 2024. In Seal We Trust? Investigating the Effect of Certifications on Perceived Trustworthiness of AI Systems. *Human-Machine Communication* 8, (2024), 141–162. https://doi.org/10.30658/hmc.8.7