Multi-Level AI Trustworthiness Labels Scale Potential Users' Perceptions and Evaluations of AI Products

Christina U. Pfeuffer
Catholic University of Eichstätt-Ingolstadt, Germany

Katholische Universität Eichstätt-Ingolstadt

Theoretical Background

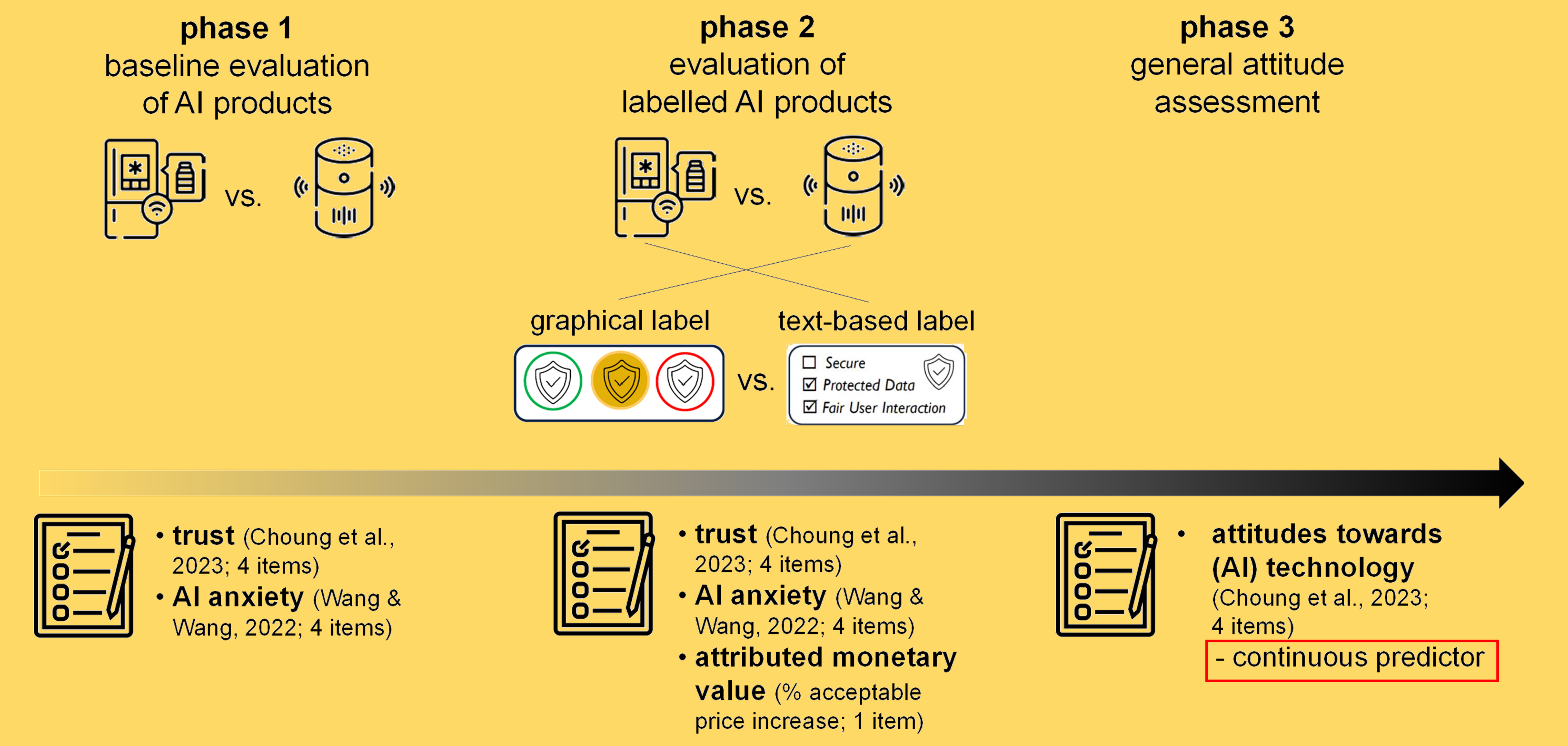
(Potential) users of artificial intelligence (AI) products (e.g., smart fridges, voice assistants) are rarely able to evaluate the trustworthiness of AI products accurately (Gillespie et al., 2023; Schlicker et al., 2025), as corresponding information is commonly not easily accessible. Concerns regarding AI trustworthiness, in particular, data security and data privacy (DSDP) concerns (Gillespie et al., 2023) jeopardize a further widespread acceptance and broader adoption of AI products (see e.g., Marangunić & Granić, 2015; Venkatesh et al., 2003, for theories on technology acceptance). Trust is an essential precursor of technology acceptance and adoption (e.g., Vorm & Combs, 2022).

As such, both users' misplaced distrust (e.g., Choung et al., 2023; Schlicker et al., 2025) as well as users' misplaced trust (due to expectancy violations, e.g., Hong, 2021; Rheu et al., 2024) prevent the further acceptance and adoption of new (and trustworthy) AI technologies and obstruct corresponding benefits of AI usage.

Here, I investigated whether multi-level AI trustworthiness labels communicating DSDP information suitably scale users' (dis-)trust in AI products – in particular when additionally considering individual attitudes towards (AI) technology. Importantly, participants rated both labeled as well as unlabeled (baseline) AI products, allowing for an additional assessment of biases.

Methods

N = 102



phase 1
baseline evaluation of AI products

vs.

phase 2
evaluation of labelled AI products

vs.

graphical label

text-based label

vs.

☐ Secure
☑ Protected Data
☑ Fair User Interaction

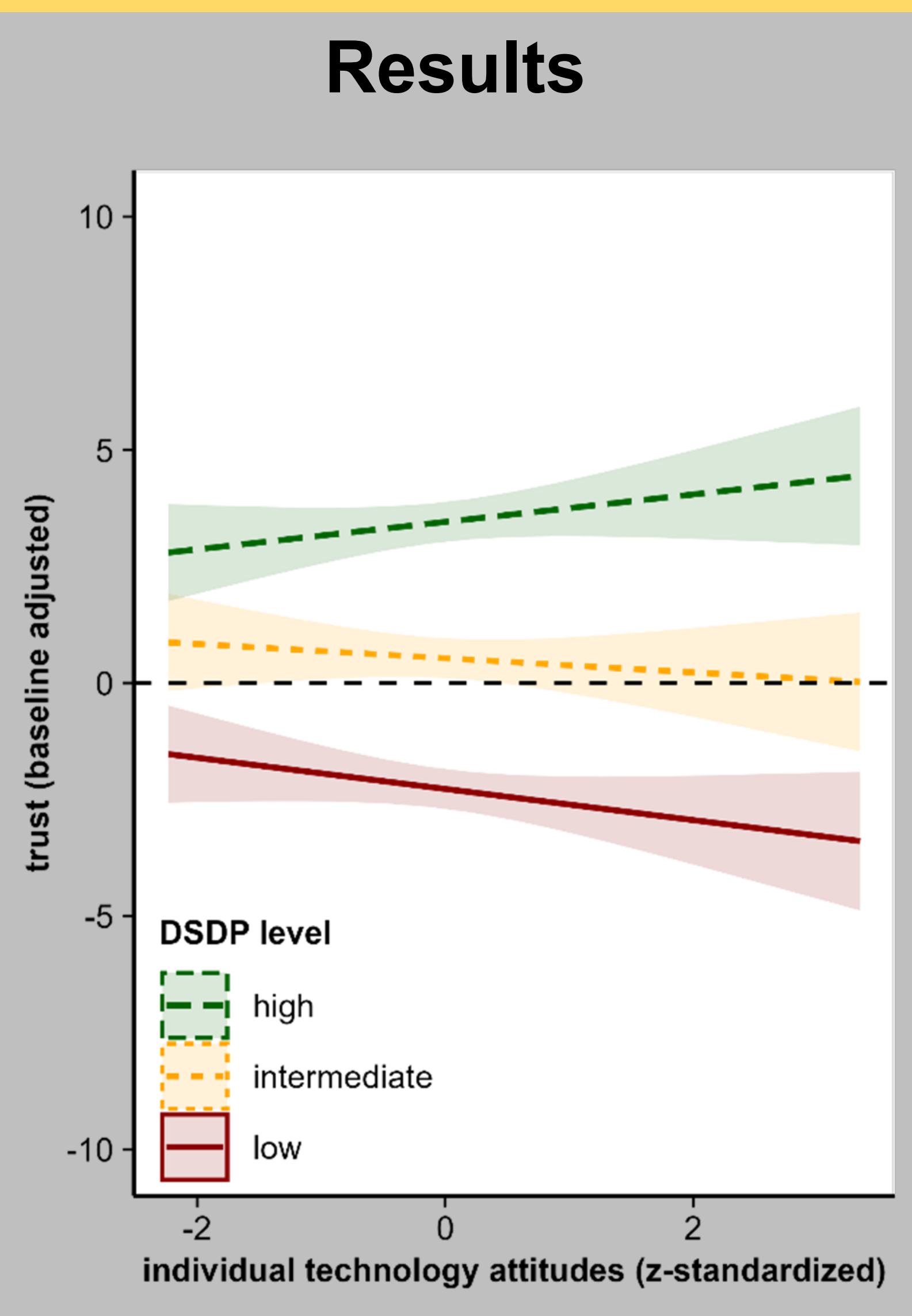
phase 3
general attitude assessment

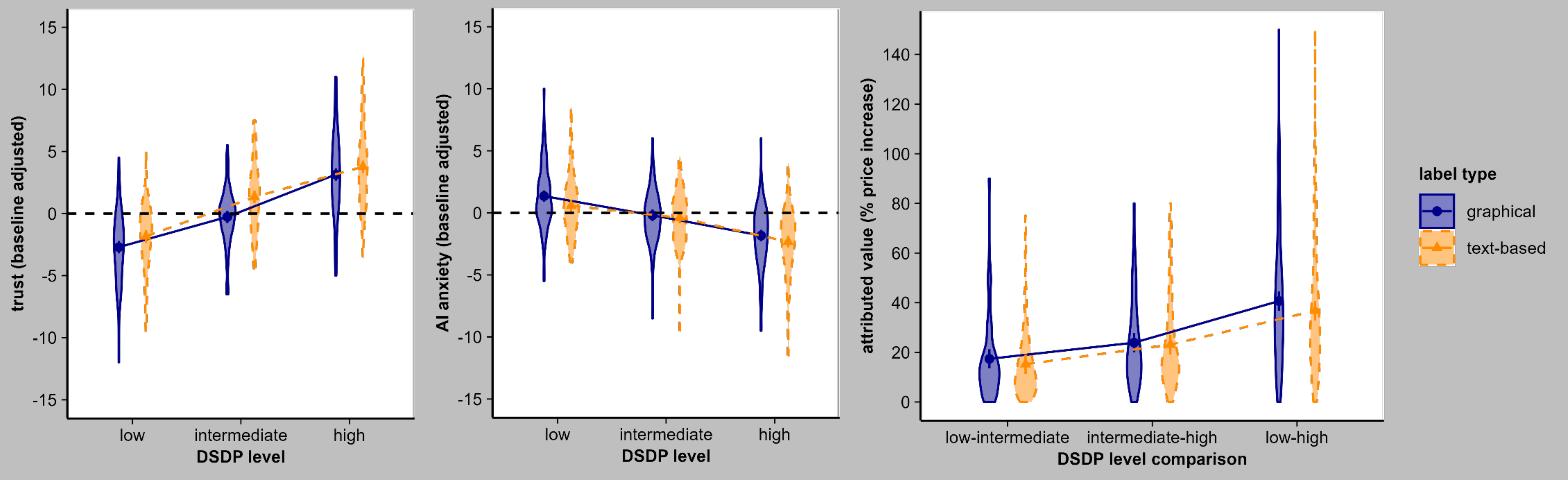
- **trust** (Choung et al., 2023; 4 items)
- **AI anxiety** (Wang & Wang, 2022; 4 items)

- **trust** (Choung et al., 2023; 4 items)
- **AI anxiety** (Wang & Wang, 2022; 4 items)
- **attributed monetary value** (% acceptable price increase; 1 item)

- **attitudes towards (AI) technology** (Choung et al., 2023; 4 items)
  - continuous predictor

Results





trust	AI anxiety	attributed value	
BF_{10} = 0.16 ±1.58%	BF_{10} = 0.20 ±3.47%	BF_{10} = 0.35 ±3.31%	**tech attitude**
BF_{10} = 1.12 x 10^{147} ±1.88% BF_{10} = 6.73 ±3.57%	BF_{10} = 8.3 x 10^{25} ±3.05% BF_{10} = 0.05 ±9.07%	BF_{10} = 8.0 x 10^{111} ±2.66% BF_{10} = 0.06 ±9.04%	**DSDP level (comparison)** **x tech attitude**
BF_{10} = 4.6 x 10^{5} ±1.51% BF_{10} = 0.20 ±2.88%	BF_{10} = 9.5 ±2.93% BF_{10} = 0.18 ±1.67%	BF_{10} = 9.9 ±3.09% BF_{10} = 0.19 ±9.27%	**label type** **x tech attitude**
BF_{10} = 0.70 ±2.07% BF_{10} = 0.13 ±2.14%	BF_{10} = 0.1 ±2.23% BF_{10} = 0.04 ±1.67%	BF_{10} = 0.07 ±7.92% BF_{10} = 0.05 ±14.96%	**DSDP level (comparison) x label type** **x tech attitude**

impact of tech attitude on trust
DSDP level: low vs. intermediate BF_{10} = 0.27 ±2.06%
DSDP level: intermediate vs. high BF_{10} = 5.47 ±4.48%

Discussion

Trust, AI anxiety, as well as the monetary value attributed to AI products scaled with an AI trustworthiness label's DSDP level. Importantly, participants' ratings of unlabeled AI products corresponded to their perceptions of AI products labeled with an intermediate DSDP level.

This apparent bias towards intermediate DSDP judgements in the absence of information on AI products underscores the relevance of explicitly communicating AI trustworthiness to (potential) users. Interestingly, differences in trust for AI products with a high as compared to intermediate DSDP level further increased with more positive attitudes towards AI technology.

Contact: christina.pfeuffer@ku.de

- 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
- 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
- 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
- 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
- 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
- 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
- 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
- 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
- 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