On a Scale of 1 to 5, How Reliable Are AI User Studies? A Call for Developing Validated, Meaningful Scales and Metrics about User Perceptions of AI Systems

Jan Tolsdorf ⋆, Alan F. Luo ◇, Monica Kodwani ⋆, Junho Eum ⋆,
Mahmood Sharif ◁, Michelle L. Mazurek ◇, Adam J. Aviv ⋆

⋆ *The George Washington University*, ◇ *University of Maryland, College Park*, ◁ *Tel Aviv University*

***Abstract*—Public discourse around trust, safety, and bias in AI systems intensifies, and as AI systems increasingly impact consumers' daily lives, there is a growing need for empirical research to measure psychological constructs underlying the human-AI relationship. By reviewing literature, we identified a gap in the availability of validated instruments. Instead, researchers seem to adapt, reuse, or develop measures in an ad hoc manner without much systematic validation. Through piloting different instruments, we identified limitations with this approach but also with existing validated instruments. To enable more robust and impactful research on user perceptions of AI systems, we advocate for a community-driven initiative to discuss, exchange, and develop validated, meaningful scales and metrics for human-centered AI research.**

1. Introduction

Recent advancements in AI and its widespread deployment have intensified discussions about risks and harms that AI poses to both individuals and society [1]–[5]. If we hope to build ethical and inclusive AI systems that align with diverse user needs and expectations, we must ensure these discussions remain inclusive and consumer-centered rather than purely technocratic; understanding user perceptions is one critical piece of that process. As a result, human factors research has begun exploring user perceptions of different facets of AI systems in recent years, including, but not limited to, fairness [6]–[19], trust [17], [20]–[25], bias [26]–[33], harms and risks [34]–[36], and also privacy and security [37], [38]. Thus, the research community is in need of measurement instruments that reliably measure human perceptions.

For empirical studies aiming at generalizable results, it is best practice to use psychometric scales [39]–[41]. In preparation for such an empirical study, we searched the literature for psychometric scales to gauge AI-user perceptions of fairness, risk, trust, and AI literacy. However, we found a scarcity of available psychometric scales. Like many other researchers, we dealt with this situation through a combination of adapting scales from non AI-contexts and creating new scale-like questions in a somewhat ad hoc manner. In piloting previously validated scales and our new questions, we identified several issues, which likely apply to a wide range of human-centered research on AI today. To address this challenge, we advocate for community-driven initiatives and propose discussion points to advance the development of reliable and validated measurement instruments.

2. Background

Psychometrics Psychometrics involves creating, validating, and ensuring the reliability and validity of measurement instruments for psychological concepts, called constructs [39], [42]. Constructs are abstract concepts or theoretical ideas that cannot be directly observed but are measured through observable indicators. Oftentimes, constructs comprise sub-constructs, i.e., distinct dimensions that collectively represent a broader construct being measured. Sub-constructs are often identified through theoretical frameworks or data-driven approaches like exploratory factor analysis (EFA) [43], which groups related items into factors. Scale validity is demonstrated through content, criterion, and construct methods, while reliability ensures consistency and stability [39]. Validated instruments empower cross-disciplinary researchers to measure, interpret, and compare perceptions in human factors studies. They also enable the verification of correlations between human perceptions of AI systems and technical metrics.

Scales on AI Perceptions To search for psychometric scales to gauge AI chatbot user perceptions of fairness, trust, risk, and AI literacy, we conducted a literature review and screened available systematic literature reviews (SLRs) on fairness, trust, and AI literacy [44]–[48]. The reviews unveil several issues with the current availability of measures.

Among these topics, we found that most validated scales exist for AI literacy (N=11), as summarized in a recent SLR [48]. However, some were developed to capture AI literacy in specific contexts, such as in the workplace or in education. In addition, some of the scales include up to 38 items, posing challenges for practical application in human-centered studies. The 12-item Artificial Intelligence Literacy Scale (AILS) [49] is by far the most cited of these validated scales and is not subject to a specific context.

For trust in AI, available SLRs report a lack of consensus on the definition of "trust," which unsurprisingly maps to inconsistent operationalization [46], [47]. For this topic, researchers commonly rely on self-developed or adapted instruments. The validated 12-item Human-Computer Trust

Scale (HCTS) [50], specifically developed for "intelligent systems," represents the most rigorous approach to date.

SLRs on perceived fairness highlight issues with the lack of unified definitions and conceptual clarity [44], [45]. Existing measures lack validation and alignment with fairness concepts specific to human-AI interaction. To understand prior efforts, we reviewed 43 studies identified by Starke et al. [45]. Of these, 19 used single-item measures, a practice that does not account for psychological sub-constructs of fairness. Only 15 studies reused or adapted instruments from previous literature, with 12 drawing upon organizational fairness literature [51], i.e., literature outside the AI scope. Only one study provided a partial validation [52].

3. Prototyping and Piloting Scales

In preparation for an empirical study on AI chatbots, we piloted measurement instruments with a sample of 122 Prolific participants from the U.S., balanced between male and female gender. We adopted the AILS [49] for AI literacy and the HCTS [50] for trust. In the absence of validated scales, we created fairness and risk measures by adapting organizational fairness scales [51] and building on a taxonomy of AI risks and user studies [5], [53] (cf. Appendix A).

Pre-Validation Steps For each measurement instrument, we conducted an EFA following guidelines and best practices for ordinal data [43], [54]. Our analysis evaluated a single-factor solution, replicated the original dimensions, and explored alternative structures using parallel analysis with promax rotation to account for potential factor correlations. Descriptive statistics identified items with low variance or extreme skewness. Items with inter-item correlations $|r| \geq 0.8$ or $|r| < 0.3$ and item-total correlations $|r| \leq 0.5$, as well as with factor loadings < 0.4 or cross-loadings were flagged for redundancy or misalignment. Internal consistency was assessed using Cronbach's alpha (α).

Results For AILS, we found low correlations for 47 out of 66 pairs of items, overall low item-total correlations, and low factor loadings and cross loadings. The constructs yielded poor reliability. In conclusion, we could not replicate the original factor structure despite its previous validation, leading us to discard the scale.

For HCTS, three pairs of items showed low pairwise correlations, but the factor solutions had adequate loadings, communalities, and α-reliabilities. However, we could not replicate the original four-factor structure, but only found support for a three-factor solution.

For the adapted fairness scale, two pairs of items showed low correlations, but the single-factor solution had adequate loadings, communalities, and α-reliabilities. However, multidimensionality was lost. Specifically, the scale no longer reflected the original sub-constructs from organizational fairness research, indicating that sub-constructs cannot be meaningfully interpreted. We believe this limitation may extend to other studies that have adapted the same scale.

For our self-generated risk scale, we identified meaningful single- and three-factor solutions with good factor loadings, communalities, and α-reliabilities. Four pairs of items had low correlations. Nonetheless, it may serve as a foundation for future instrument development.

4. Call for a Community-Driven Initiative

Our findings highlight several issues with the availability of reliable measurement instruments for human-centerd AI research. In particular, widely-used scales may show poor reliability and fail to replicate their original factor structure, indicating a need for stronger foundational development and validation. The lack of consensus on the definition of constructs [44]–[48] makes it difficult to assess and compare findings across studies. Our findings also highlight issues with the common practice of adapting scales from outside the AI domain to the AI domain, which may be leading to a loss of multidimensionality and therefore to a loss of meaning and interpretability.

Developing rigorous scales is resource-intensive [39]. We advocate for a collaborative, community-driven initiative to raise awareness and mobilize efforts to enhance transparency and reproducibility in scale development, fostering cross-study comparisons and integration.

Identification of Key Factors and Contexts While developing new measurement tools for emerging contexts is sometimes necessary, certain fundamental factors consistently require robust measurement, providing the chance to reduce duplication of effort. Constructs such as trust, fairness, and risk are areas of significant interest. Similarly, frequently studied contexts and application areas, such as conversational assistants, autonomous decision-making, and generative AI have overlapping but also distinct research needs. We believe that community discussions can play a pivotal role in identifying and prioritizing the most important factors for which systematic measurement is needed.

- What are the key factors, aspects, and dimensions requiring standardized measurement scales?
- To what extent should scales account for the contextual variations of different AI applications?

Improving Rigor Improving the rigor of measurement tools involves not only developing robust instruments but also documenting failed validation attempts and the lessons learned from these efforts. For instance, the OECD.AI policy repository [55] demonstrates how centralized resources can facilitate access to tools and metrics for building trustworthy AI systems, but does not provide for documentation of failed attempts. Meanwhile, similar centralized resources for research remain notably absent.

- What features would make a repository for validation efforts most useful to the research community?
- How can we combine the efforts of a diverse research community from different research traditions, incentivizing interdisciplinary approaches?
- What incentives could encourage researchers to contribute to and maintain such a resource?

We believe raising awareness and mobilizing community efforts will lead to more standardized, validated, and contextually-relevant instruments for assessing user perceptions of AI systems with greater rigor and reliability.

Acknowledgment

This material is based on work supported in part by the Institute for Trustworthy AI and Law and Society (TRAILS), which is supported by the National Science Foundation under Grant No. 2229885.

References

[1] J. Dastin, "Amazon scraps secret ai recruiting tool that showed bias against women," *Reuters*, 2018, accessed: Oct. 03, 2023. [Online]. Available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[2] Z. B. Wolf, "Ai can be racist, sexist and creepy. what should we do about it?" *CNN Politics*, 2023, accessed: Oct. 03, 2023. [Online]. Available: https://www.cnn.com/2023/03/18/politics/ai-chatgpt-racist-what-matters/index.html

[3] P. Verma, "These robots were trained on ai. they became racist and sexist." *Washington Post*, 2022, accessed: Oct. 03, 2023. [Online]. Available: https://www.washingtonpost.com/technology/2022/07/16/racist-robots-ai/

[4] K. Hao, "Ai is sending people to jail—and getting it wrong," *MIT Technology Review*, 2019, accessed: Oct. 03, 2023. [Online]. Available: https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/

[5] L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P.-S. Huang, J. Mellor, A. Glaese, M. Cheng, B. Balle, A. Kasirzadeh, C. Biles, S. Brown, Z. Kenton, W. Hawkins, T. Stepleton, A. Birhane, L. A. Hendricks, L. Rimell, W. Isaac, J. Haas, S. Legassick, G. Irving, and I. Gabriel, "Taxonomy of Risks posed by Language Models," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 214–229.

[6] G. Juijn, N. Stoimenova, J. Reis, and D. Nguyen, "Perceived Algorithmic Fairness using Organizational Justice Theory: An Empirical Case Study on Algorithmic Hiring," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 2023, pp. 775–785.

[7] J. Salac, R. Landesman, S. Druga, and A. J. Ko, "Scaffolding Children's Sensemaking around Algorithmic Fairness," in *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*, 2023, pp. 137–149.

[8] T. Schmude, L. Koesten, T. Möller, and S. Tschiatschek, "On the Impact of Explanations on Understanding of Algorithmic Decision-Making," in *2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 959–970.

[9] N. Grgić-Hlača, G. Lima, A. Weller, and E. M. Redmiles, "Dimensions of Diversity in Human Perceptions of Algorithmic Fairness," in *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 2022, pp. 1–12.

[10] N. Sonboli, J. J. Smith, F. Cabral Berenfus, R. Burke, and C. Fiesler, "Fairness and Transparency in Recommendation: The Users' Perspective," in *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, 2021, pp. 274–279.

[11] M. H. Lee and C. J. Chew, "Understanding the Effect of Counterfactual Explanations on Trust and Reliance on AI for Human-AI Collaborative Clinical Decision Making," *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, no. CSCW2, pp. 369:1–369:22, 2023.

[12] H.-F. Cheng, L. Stapleton, R. Wang, P. Bullock, A. Chouldechova, Z. S. S. Wu, and H. Zhu, "Soliciting Stakeholders' Fairness Notions in Child Maltreatment Predictive Systems," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–17.

[13] R. Wang, F. M. Harper, and H. Zhu, "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*, 2020, pp. 1–14.

[14] M. Jakesch, Z. Buçinca, S. Amershi, and A. Olteanu, "How Different Groups Prioritize Ethical Values for Responsible AI," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022.

[15] N. Sambasivan, E. Arnesen, B. Hutchinson, T. Doshi, and V. Prabhakaran, "Re-imagining Algorithmic Fairness in India and Beyond," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (Facct)*, 2021, p. 328.

[16] M. Yurrita, T. Draws, A. Balayn, D. Murray-Rust, N. Tintarev, and A. Bozzon, "Disentangling Fairness Perceptions in Algorithmic Decision-Making: The Effects of Explanations, Human Oversight, and Contestability," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.

[17] J. Schoeffer, N. Kuehl, and Y. Machowski, ""There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022.

[18] T. van Nuenen, J. Such, and M. Cote, "Intersectional Experiences of Unfair Treatment Caused by Automated Computational Systems," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, pp. 1–30, 2022.

[19] M. Lünich, B. Keller, and F. Marcinkowski, "Fairness of Academic Performance Prediction for the Distribution of Support Measures for Students: Differences in Perceived Fairness of Distributive Justice Norms," *Technology, Knowledge and Learning*, 2023.

[20] S. Mehrotra, C. C. Jorge, C. M. Jonker, and M. L. Tielman, "Integrity-based Explanations for Fostering Appropriate Trust in AI Agents," *ACM Transactions on Interactive Intelligent Systems*, vol. 14, no. 1, pp. 4:1–4:36, 2024.

[21] M. Le Guillou, L. Prévot, and B. Berberian, "Trusting Artificial Agents: Communication Trumps Performance," in *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, 2023, pp. 299–306.

[22] C. Degachi, M. L. Tielman, and M. Al Owayyed, "Trust and Perceived Control in Burnout Support Chatbots," in *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–10.

[23] P. K. Kahr, G. Rooks, M. C. Willemsen, and C. C. Snijders, "It Seems Smart, but It Acts Stupid: Development of Trust in AI Advice in a Repeated Legal Decision-Making Task," in *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 2023, pp. 528–539.

[24] S. Ma, Y. Lei, X. Wang, C. Zheng, C. Shi, M. Yin, and X. Ma, "Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–19.

[25] G. Beltrão, S. Sousa, and D. Lamas, "Unmasking Trust: Examining Users' Perspectives of Facial Recognition Systems in Mozambique," in *Proceedings of the 4th African Human Computer Interaction Conference*, 2024, pp. 38–43.

[26] S. E. Walsh and K. M. Feigh, "Mental Models of AI Performance and Bias of Nontechnical Users," in *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2023, pp. 4116–4121.

[27] C. Wang, K. Wang, A. Y. Bian, R. Islam, K. N. Keya, J. Foulds, and S. Pan, "When Biased Humans Meet Debiased AI: A Case Study in College Major Recommendation," *ACM Transactions on Interactive Intelligent Systems*, vol. 13, no. 3, pp. 17:1–17:28, 2023.

[28] C. W. T. Yuan, N. Bi, Y.-F. Lin, and Y.-H. Tseng, "Contextualizing User Perceptions about Biases for Human-Centered Explainable Artificial Intelligence," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–15.

[29] C. Rastogi, Y. Zhang, D. Wei, K. R. Varshney, A. Dhurandhar, and R. Tomsett, "Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW1, pp. 83:1–83:22, 2022.

[30] M. Ragot, N. Martin, and S. Cojean, "AI-generated vs. Human Artworks. A Perception Bias Towards Artificial Intelligence?" in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–10.

[31] S. Kim, P. Oh, and J. Lee, "Algorithmic gender bias: Investigating perceptions of discrimination in automated decision-making," *Behaviour & Information Technology*, pp. 1–14, 2024.

[32] P. Narayanan Venkit, S. Gautam, R. Panchanadikar, T.-H. Huang, and S. Wilson, "Unmasking Nationality Bias: A Study of Human Perception of Nationalities in AI-Generated Articles," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 2023.

[33] G. I. Melsión, I. Torre, E. Vidal, and I. Leite, "Using Explainability to Help Children UnderstandGender Bias in AI," in *Interaction Design and Children (IDC)*, 2021, pp. 87–99.

[34] T. Araujo, N. Helberger, S. Kruikemeier, and C. H. de Vreese, "In AI we trust? Perceptions about automated decision-making by artificial intelligence," *AI & SOCIETY*, vol. 35, no. 3, pp. 611–623, 2020.

[35] N. Dennler, A. Ovalle, A. Singh, L. Soldaini, A. Subramonian, H. Tu, W. Agnew, A. Ghosh, K. Yee, I. F. Peradejordi, Z. Talat, M. Russo, and J. D. J. D. P. Pinhal, "Bound by the Bounty: Collaboratively Shaping Evaluation Processes for Queer AI Harms," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 2023, pp. 375–386.

[36] G. Lima, N. Grgić-Hlača, and M. Cha, "Blaming Humans and Machines: What Shapes People's Reactions to Algorithmic Harm," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.

[37] P. Chametka, S. Maqsood, and S. Chiasson, "Security and Privacy Perceptions of Mental Health Chatbots," in *2023 20th Annual International Conference on Privacy, Security and Trust (PST)*, 2023, pp. 1–7.

[38] P. G. Kelley, C. Cornejo, L. Hayes, E. S. Jin, A. Sedley, K. Thomas, Y. Yang, and A. Woodruff, ""There will be less privacy, of course": How and why people in 10 countries expect AI will affect privacy in the future," in *Proceedings of the 19th Symposium on Usable Privacy and Security (SOUPS)*, 2023.

[39] G. O. Boateng, T. B. Neilands, E. A. Frongillo, H. R. Melgar-Quiñonez, and S. L. Young, "Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer," *Frontiers in Public Health*, vol. 6, 2018.

[40] A. T. Ginty, "Psychometric properties," in *Encyclopedia of Behavioral Medicine*, M. D. Gellman and J. R. Turner, Eds., 2013, pp. 1563–1564.

[41] "APA Dictionary of Psychology," accessed: 21-Apr-2024. [Online]. Available: https://dictionary.apa.org/psychometric-scaling

[42] J. D. Wasserman and B. A. Bracken, "Fundamental Psychometric Considerations in Assessment," in *Handbook of Psychology, Second Edition*, 2012, ch. 3.

[43] M. W. Watkins, "Exploratory Factor Analysis: A Guide to Best Practice," *Journal of Black Psychology*, vol. 44, no. 3, pp. 219–246, 2018.

[44] D. Narayanan, M. Nagpal, J. McGuire, S. Schweitzer, and D. De Cremer, "Fairness Perceptions of Artificial Intelligence: A Review and Path Forward," *International Journal of Human–Computer Interaction*, vol. 40, no. 1, pp. 4–23, 2024.

[45] C. Starke, J. Baleis, B. Keller, and F. Marcinkowski, "Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature," *Big Data & Society*, vol. 9, no. 2, p. 205395172211151, 2022.

[46] T. A. Bach, A. Khan, H. Hallock, G. Beltrão, and S. Sousa, "A Systematic Literature Review of User Trust in AI-Enabled Systems: An HCI Perspective," *International Journal of Human–Computer Interaction*, vol. 40, no. 5, pp. 1251–1266, 2024.

[47] S. W. T. Ng and R. Zhang, "Trust in AI chatbots: A systematic review," *Telematics and Informatics*, vol. 97, p. 102240, 2025.

[48] O. Almatrafi, A. Johri, and H. Lee, "A systematic review of AI literacy conceptualization, constructs, and implementation and assessment efforts (2019–2023)," *Computers and Education Open*, vol. 6, p. 100173, 2024.

[49] B. Wang, P.-L. P. Rau, and T. Yuan, "Measuring user competence in using artificial intelligence: Validity and reliability of artificial intelligence literacy scale," *Behaviour & Information Technology*, vol. 42, no. 9, pp. 1324–1337, 2023.

[50] S. Gulati, S. Sousa, and D. Lamas, "Design, development and evaluation of a human-computer trust scale," *Behaviour & Information Technology*, vol. 38, no. 10, pp. 1004–1015, 2019.

[51] J. A. Colquitt and J. B. Rodell, "Measuring justice and fairness," in *The Oxford Handbook of Justice in the Workplace*, 2015, pp. 187–202.

[52] D. Shin, B. Zhong, and F. A. Biocca, "Beyond user experience: What constitutes algorithmic experiences?" *International Journal of Information Management*, vol. 52, p. 102061, 2020.

[53] N. Goyal, I. D. Kivlichan, R. Rosen, and L. Vasserman, "Is Your Toxicity My Toxicity? Exploring the Impact of Rater Identity on Toxicity Annotation," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, pp. 1–28, 2022.

[54] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis*, 8th ed., 2019.

[55] OECD.AI, "OECD.AI: The OECD Artificial Intelligence Policy Observatory," 2025, accessed: 2025-01-20. [Online]. Available: https://oecd.ai/en/

Appendix A.
Scales

TABLE 1. DEVELOPED INSTRUMENT ON PERCEIVED RISK IN AI CHATBOTS INFORMED BY WEIDINGER ET AL. [5] AND GOYAL ET AL. [53].

Item Number	Item
risks_scale_1	Outputs that reproduce, contain, or reinforce harmful stereotypes of specific groups of people?
risks_scale_2	Outputs that include threats or language inciting violence?
risks_scale_3	Outputs that are less helpful in certain languages or dialects?
risks_scale_4	Outputs that disseminate or reproduce false or misleading information?
risks_scale_5	Outputs that cause real-world harm by sharing incorrect information about important topics, such as medicine or the law?
risks_scale_6	Outputs that promote harmful stereotypes by implying gender or ethnic identity?
risks_scale_7	Outputs that include profanities, identity attacks, insults, or offensive language?
risks_scale_8	Outputs that reproduce or reinforce norms and values that exclude specific groups of people, such as exclusionary language?
risks_scale_9	Outputs that are less helpful for different social groups?
risks_scale_10	Outputs that are negative, discriminatory, or hateful against a group of people based on criteria including (but not limited to) race or ethnicity, religion, nationality or citizenship, disability, age, or sexual orientation?
risks_scale_11	Outputs that contain swear words, curse words, or other obscene or profane language?
risks_scale_12	Outputs that are inflammatory, stereotyping, insulting, or negative towards a person or a group of people?
risks_scale_13	Outputs that contain threatening language, such as encouraging violence or harm, including self-harm?

TABLE 2. DEVELOPED INSTRUMENT ON PERCEIVED FAIRNESS IN AI CHATBOTS INFORMED BY COLQUITT AND RODELL [51].

Item Number	Item
fairness_scale_1	Outcomes generated by AI chatbots are neutral and unbiased.
fairness_scale_2	Outcomes generated by AI chatbots are based on accurate information.
fairness_scale_3	Outcomes generated by AI chatbots take into account concerns of a wide range of different people.
fairness_scale_4	Outcomes generated by AI chatbots uphold ethical and moral standards.
fairness_scale_5	Outcomes generated by AI chatbots are just.
fairness_scale_6	Outcomes generated by AI chatbots are fair.
fairness_scale_7	Outcomes generated by AI chatbots are polite and respectful.
fairness_scale_8	Outcomes generated by AI chatbots refrain from improper remarks or comments.
fairness_scale_9	Explanations provided about outcomes generated by AI chatbots are honest.
fairness_scale_10	Explanations provided about outcomes generated by AI chatbots are thorough.

TABLE 3. HUMAN-COMPUTER TRUST SCALE (HCTS) [50] ADAPTED TO AI CHATBOTS.

Item Number	Item
hcts_scale_1	I believe that there could be negative consequences when using AI chatbots.
hcts_scale_2	I feel I must be cautious when using AI chatbots.
hcts_scale_3	It is risky to interact with AI chatbots.
hcts_scale_4	I believe that AI chatbots will act in my best interest.
hcts_scale_5	I believe that AI chatbots will do its best to help me if I need help.
hcts_scale_6	I believe that AI chatbots are interested in understanding my needs and preferences.
hcts_scale_7	I think that AI chatbots are competent and effective in helping me with what I use them for.
hcts_scale_8	I believe that AI chatbots have all the functionalities I would expect from them.
hcts_scale_9	If I use AI chatbots, I think I would be able to depend on them completely.
hcts_scale_10	I can always rely on AI chatbots for the things I use them for.
hcts_scale_11	I can trust the information presented to me by AI chatbots.

TABLE 4. FACTOR LOADINGS OF THE ADAPTED HUMAN-COMPUTER TRUST SCALE (HCTS) [50] WHEN TRYING TO REPLICATE THE ORIGINAL FACTOR STRUCTURE, WITH LOADINGS $> |.40|$ INDICATED AS BOLD.

Item Number	Sub-construct	Factor 1	Factor 2	Factor 3	Factor 4	Communality
hcts_scale_1	Risk	0.22	**0.66**	-0.03	-0.05	0.49
hcts_scale_2	Risk	0.21	**0.67**	-0.04	-0.02	0.50
hcts_scale_3	Risk	-0.25	**0.86**	0.08	0.09	0.82
hcts_scale_4	Benevolence	0.11	0.16	**0.58**	-0.01	0.37
hcts_scale_5	Benevolence	-0.08	0.06	**0.89**	-0.03	0.81
hcts_scale_6	Benevolence	0.29	-0.16	**0.59**	0.07	0.46
hcts_scale_7	Competence	0.01	0.05	0.00	**0.97**	0.95
hcts_scale_8	Competence	**0.63**	-0.10	0.15	-0.05	0.43
hcts_scale_9	Reciprocity	**0.75**	0.11	-0.06	-0.09	0.58
hcts_scale_10	Reciprocity	**0.66**	-0.03	-0.04	0.22	0.49
hcts_scale_11	Reciprocity	**0.55**	0.14	0.11	0.04	0.34

TABLE 5. ARTIFICIAL INTELLIGENCE LITERACY SCALE (AILS) [49].

Item Number	Item
ails_scale_1	I can distinguish between smart devices and non-smart devices.
ails_scale_2	I do not know how AI technology can help me. R
ails_scale_3	I can identify the AI technology employed in the applications and products I use.
ails_scale_4	I can skillfully use AI applications or products to help me with my daily work.
ails_scale_5	It is usually hard for me to learn to use a new AI application or product. R
ails_scale_6	I can use AI applications or products to improve my work efficiency.
ails_scale_7	I can evaluate the capabilities and limitations of an AI application or product after using it for a while.
ails_scale_8	I can choose a proper solution from various solutions provided by a smart agent.
ails_scale_9	I can choose the most appropriate AI application or product from a variety for a particular task.
ails_scale_10	I always comply with ethical principles when using AI applications or products.
ails_scale_11	I am never alert to privacy and information security issues when using AI applications or products. R
ails_scale_12	I am always alert to the abuse of AI technology.

TABLE 6. FACTOR LOADINGS OF THE ARTIFICIAL INTELLIGENCE LITERACY SCALE (AILS) [49] WHEN TRYING TO REPLICATE THE ORIGINAL FACTOR STRUCTURE, WITH LOADINGS $> |.40|$ INDICATED AS BOLD.

Item Number	Sub-construct	Factor 1	Factor 2	Factor 3	Factor 4	Communality
ails_scale_1	Awareness	-0.13	0.23	**0.81**	-0.10	0.73
ails_scale_2	Awareness	**0.76**	-0.15	0.17	-0.02	0.63
ails_scale_3	Awareness	-0.09	0.34	0.10	0.16	0.16
ails_scale_4	Usage	**0.68**	0.37	-0.17	-0.12	0.64
ails_scale_5	Usage	0.31	-0.04	0.20	0.33	0.25
ails_scale_6	Usage	**0.71**	0.15	-0.27	0.02	0.59
ails_scale_7	Evaluation	0.16	**0.44**	0.24	0.13	0.30
ails_scale_8	Evaluation	0.06	**0.64**	-0.11	0.13	0.45
ails_scale_9	Evaluation	0.08	**0.74**	0.31	-0.27	0.72
ails_scale_10	Ethics	0.03	0.10	0.02	0.24	0.07
ails_scale_11	Ethics	-0.02	-0.11	0.11	0.26	0.09
ails_scale_12	Ethics	-0.09	0.13	-0.25	**0.70**	0.57