Towards an AI Safety Literacy Scale

Vanessa Bracamonte
KDDI Research, Inc.

Yukiko Sawaya
KDDI Research, Inc.

Yukihide Kohama
KDDI Research, Inc.

Abstract

As part of an overall objective to improve communication and education on AI safety, we develop and preliminarily validate an AI Safety Literacy Scale to measure self-perceived competence (knowledge, attitudes, skills) regarding AI-related risks. Based on research on AI security, risk, and trustworthiness, we established the scope of AI safety literacy, conducted a literature review and selected candidate items from existing work on AI literacy. We then conducted a user survey and validated the scale using exploratory and confirmatory factor analysis on split sample data from 740 participants. The result is a 3-factor AI Safety Literacy Scale consisting of Safe AI Use Knowledge, AI Output Vigilance and Responsible AI Use factors.

1 Introduction

As AI-based applications have become more widely used, incidents involving them have also increased [19, 22]. Although improving AI security and trustworthiness is required to address these issues, user education on AI-related risks is also important [10, 14, 22, 28]. Understanding current AI safety-related competencies, and how those competencies affect how people interact with AI, is important for designing effective communication and education on this topic. Having a way to measure AI safety literacy would allow us to investigate what factors influence it and how it relates to other variables such as intention to behave safely in AI-related situations. However, although a number of AI literacy scales exist, to the best of our knowledge there is currently no scale specifically on AI safety literacy. By AI safety literacy, we refer in particular to self-perceived competencies [18] (knowledge, attitudes, and skills) for recognizing and managing risks that people can encounter in situations where AI is involved, rather than technical aspects of AI security. To address this gap, our objective was to develop a scale for evaluating self-perceived competence in AI safety.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2024.
August 11–13, 2024, Philadelphia, PA, United States.

2 Methodology

Selection of Candidate Items Our requirements for the scale were that it should cover competencies in aspects of AI safety that involve the perspective of what people are able to exert some degree of control on, regardless of whether they directly interact with the AI: inputs to the AI, handling of output from AI, and the decision to (or how to) use the AI itself. We started by identifying the AI aspects to measure, based on existing work in AI security, risk and trustworthiness guidelines and frameworks [20]. Next, we conducted a search on the literature on AI literacy to select items for the scale (Appendix Table 2). Although there are no scales for AI safety literacy, AI literacy scales often include aspects that can be considered as part of AI safety (e.g. ethical considerations [1, 4, 13, 27]). We identified papers that proposed AI literacy or competency scales and listed all items (269 items). From the list, we selected items that addressed safety-related aspects identified in the previous step. We then adapted the item descriptions to standardize the way AI was mentioned in the description, since some of the items used different terms for AI, or to adapt it more clearly to general AI. At every stage, the resulting items were verified independently by two of the authors and the results discussed until an agreement was reached. We verified that there were at least 3 items representing each of the AI safety aspects of interest. The final candidate pool contained 69 items.

Survey design We designed a user survey to validate the items and create the scale. We prepared a questionnaire that

included the candidate AI Safety Literacy Scale items, with a 5-point Likert response scale from Strongly disagree to Strongly agree, with an option for "did not understand". The questionnaire also included questions on AI use, IT skills and demographics, as well as attention check questions. Start and end times were recorded. The user survey was conducted online using a third-party survey company, among Japanese participants. The items were translated using a back-translation process. We used machine translation using two different LLM models for the initial translation draft, back-translation and result comparison [6]. The results were revised by one of the authors, and then reviewed by another author independently. Remaining issues were discussed and resolved. The completed and translated questionnaire was then checked by two non-experts, who identified words that were difficult to understand, and the questionnaire was revised based on that feedback.

Ethical Considerations The user survey was exempt from ethical review according to the guidelines of our institution. Nevertheless we considered ethical guidelines for designing the survey. Participation was voluntary and we did not receive any personally identifiable information.

3 Results

Data The survey was conducted at the end of March and obtained 1,200 responses, excluding those who failed the attention check questions. We checked the data for straightlining within the scale candidate items and for low response times (duration of less than half of the median time). 156 low-quality cases were removed. The mean item missingness ("did not understand" response) rate was 0.072. Since this was not random missingness, imputation was not appropriate and cases with missing values were removed, resulting in 740 cases for analysis. The sample had a mean age of 46.5 years old (std = 13.3). Gender distribution was 45.8% female and 54.2% male respondents (Appendix Table 3).

For the analysis, we conducted exploratory (EFA) (N=372) and confirmatory factor analysis (CFA) (N=368) following guidelines included in [11], using a stratified random sample split within age x gender groups. Before conducting the EFA we checked item correlations (Spearman); all were under 0.8. VIF values were under 10. The Bartlett's test of sphericity was significant (p <0.001), indicating sufficiently correlated items. The overall Kaiser-Meyer-Olkin (KMO) measure was 0.984. Based on the scree plot (4 factors) and parallel analysis (2 factor) results, we tested 2, 3 and 4-factor models. We used maximum likelihood with a promax rotation. The primary loading threshold was 0.4, with a cross-loading cutoff of 0.32 and a minimum gap of 0.2 between primary and secondary loading [2]. For item removal, in addition to these thresholds we also considered the content of the item and its missingness rate. The process resulted in 3-factor model (F1:8 items, F2: 4 items, F3: 3 items), with a cumulative variance explained of 62.42% (Appendix Table 4). CFA was conducted using semopy [12] with the model structure identified in the EFA. The results indicated a generally acceptable fit, although RMSEA was higher than 0.06 [11]: RMSEA = 0.064, SRMR = 0.041, CFI = 0.971, TLI = 0.965. The lowest Cronbach's α among the 3 factors was 0.831 [2]. The minimum standardized factor loading was .716 [2] (Appendix Table 5).

Table 1: AI Safety Literacy Scale.

Item	Description
Safe AI Use Knowledge (items adapted from [1, 7, 13, 24])	
SAIU1	I can explain the limitations of AI.
SAIU2	I have knowledge of the risks AI processing poses.
SAIU3	I can describe risks that may arise when using AI.
SAIU4	I can explain why data security must be considered when using AI.
SAIU5	I can explain why data privacy must be considered when using AI.
SAIU6	I can describe potential legal problems that may arise when using AI.
SAIU7	I know that the government introduced security and safety guidelines for AI.
SAIU8	I can evaluate AI for their ethical risks and implications.
AI Output Vigilance (items adapted from [1, 5, 7])	
AIOV1	I know that some information output by AI may include errors.
AIOV2	I understand AI can produce output that is factually inaccurate.
AIOV3	I understand AI can produce output that is out of context or inappropriate.
AIOV4	I can prevent an AI from influencing me in my everyday decisions.
Responsible AI Use (items adapted from [7, 13])	
RAIU1	I can follow the legal regulations regarding the use of AI.
RAIU2	I am careful to follow the rules about copyrights and licenses of content that AI produces.
RAIU3	I am aware that the time spent on AI should be managed.

Based on the competencies represented by the items in the 3 factors and our definition of AI safety, we identified dimensions related to safe use of AI (Safe AI Use Knowledge), the need to validate the output of the AI (AI Output Vigilance) and aspects to consider for using AI responsibly (Responsible AI Use) (Table 1).

4 Conclusion and Future Work

We developed and preliminarily validated an AI Safety Literacy Scale, consisting of the dimensions of Safe AI Use Knowledge, AI Output Vigilance, and Responsible AI Use. The scale is intended to support research on improving communication and education on AI safety by measuring self-perceived competence for recognizing and managing AI-related risks. There are a number of limitations, including that the scale covers the AI safety aspects and user perspective established in this study, but it is not comprehensive; in particular, input safety was not well represented. In addition, the scale has only been validated in a Japanese sample. Future work will address these limitations and also examine the extent to which the scale correlates with actual AI safety knowledge and behavior.

References

[1] Astrid Carolus, Martin J. Koch, Samantha Straka, Marc E. Latoschik, and Carolin Wienrich. 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):100014, 2023.

[2] Serena Carpenter. Ten steps in scale development and reporting: A guide for researchers. *Communication Methods and Measures*, 12(1):25–44, 2018.

[3] Lorena Casal-Otero, Alejandro Catala, Carmen Fernández-Morante, María Taboada, Beatriz Cebreiro, and Senén Barro. AI literacy in K-12: A systematic literature review. *International Journal of STEM Education*, 10(1):29, 2023.

[4] Ismail Celik. Towards intelligent-TPACK: An empirical study on teachers' professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education. *Computers in Human Behavior*, 138:107468, 2023.

[5] Cecilia Ka Yuk Chan and Wenxin Zhou. An expectancy value theory (EVT) based instrument for measuring student perceptions of generative AI. *Smart Learning Environments*, 10:64, 2023.

[6] Ji-Bum Chung and Taehyun Kim. Leveraging large language models for enhanced back-translation: Techniques and applications. *IEEE Access*, 13:61322–61328, 2025.

[7] Ian Clifford, Stefano Kluzer, Sandra Troia, Mara Jakobsone, and Uldis Zandbergs. DigCompSat: A self-reflection tool for the european digital competence framework for citizens. Technical Report JRC123226, Joint Research Centre, Luxembourg, 2020.

[8] Judith Cosgrove and Romina Cachia. DigComp 3.0: European digital competence framework - fifth edition. Technical Report JRC144121, Joint Research Centre, Luxembourg, 2025.

[9] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations and directives (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689, 12 July 2024, 2024.

[10] European Union Agency for Cybersecurity. Securing machine learning algorithms. Technical report, European Union Agency for Cybersecurity, December 2021.

[11] Li-tze Hu and Peter M. Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1):1–55, 1999.

[12] Anna A. Igolkina and Georgy Meshcheryakov. semopy: A python package for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 0(0):1–12, 2020.

[13] Ozan Karaca, S. Ayhan Çalışkan, and Kadir Demir. Medical artificial intelligence readiness scale for medical students (MAIRS-MS) - development, validity and reliability study. *BMC Medical Education*, 21(1):112, 2021.

[14] Alexandra Klymenko, Stephen Meisenbacher, Patrick Gage Kelley, Sai Teja Peddinti, Kurt Thomas, and Florian Matthes. "we are not future-ready": Understanding AI privacy risks and existing mitigation strategies from the perspective of AI developers in europe. In *Twenty-First Symposium on Usable Privacy and Security (SOUPS 2025)*, pages 113–132, Seattle, WA, 2025. USENIX Association.

[15] Matthias Carl Laupichler, Alexandra Aster, Nicolas Haverkamp, and Tobias Raupach. Development of the "scale for the assessment of non-experts' AI literacy": An exploratory factor analysis. *Computers in Human Behavior Reports*, 12:100338, 2023.

[16] Matthias Carl Laupichler, Alexandra Aster, Jana Schirch, and Tobias Raupach. Artificial intelligence literacy in higher and adult education: A scoping literature review. *Computers and Education: Artificial Intelligence*, 3:100101, 2022.

[17] Tomáš Lintner. A systematic review of AI literacy scales. *npj Science of Learning*, 9:50, 2024.

[18] Duri Long and Brian Magerko. What is AI literacy? competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–16, New York, NY, USA, 2020. Association for Computing Machinery.

[19] MIT AI Risk Initiative. AI incident tracker. https://airisk.mit.edu/ai-incident-tracker. Accessed 2026-05-21.

[20] National Institute of Standards and Technology. Crosswalk: An illustration of how NIST AI RMF trustworthiness characteristics relate to the OECD recommendation on AI, proposed EU AI act, executive order 13960, and blueprint for an AI bill of rights. Technical report, U.S. Department of Commerce, January 2023.

[21] Davy Tsz Kit Ng, Jac Ka Lok Leung, Samuel Kai Wah Chu, and Maggie Shen Qiao. Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2:100041, 2021.

[22] Open Worldwide Application Security Project. OWASP AI exchange. https://owaspai.org/.

[23] Organisation for Economic Co-operation and Development. Recommendation of the council on artificial intelligence. OECD Legal Instruments, OECD/LEGAL/0449, 2019.

[24] Marc Pinski and Alexander Benlian. AI literacy - towards measuring human competency in artificial intelligence. In *Proceedings of the 56th Hawaii International Conference on System Sciences*, pages 165–174, 2023.

[25] Jiahong Su, Davy Tsz Kit Ng, and Samuel Kai Wah Chu. Artificial intelligence (AI) literacy in early childhood education: The challenges and opportunities. *Computers and Education: Artificial Intelligence*, 4:100124, 2023.

[26] Elham Tabassi. Artificial intelligence risk management framework (AI RMF 1.0). Technical Report NIST AI 100-1, National Institute of Standards and Technology, Gaithersburg, MD, 2023.

[27] Bingcheng Wang, Pei-Luen Patrick Rau, and Tianyi Yuan. Measuring user competence in using artificial intelligence: Validity and reliability of artificial intelligence literacy scale. *Behaviour & Information Technology*, 42(9):1324–1337, 2023.

[28] Chien Wen Tina Yuan, Nanyi Bi, Ya-Fang Lin, and Yuen-Hsien 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*. Association for Computing Machinery, 2023.

[29] Shan Zhang, Priyadharshini Ganapathy Prasad, and Noah L. Schroeder. Learning about AI: A systematic review of reviews on AI literacy. *Journal of Educational Computing Research*, 63(5), 2025.

Appendices

A Acknowledgments

This work was supported by the Japan Science and Technology Agency (JST) K Program, grant number JPMJKP24C4.

B Participant Compensation

In addition to the compensation set by the survey company, all respondents were paid an additional 100 yen (~0.65 USD) due to the length of the questionnaire.

C Literature Related to Scale Development

Table 2: Literature reviewed for definition and scale development.

Purpose	References
AI literacy scales and measurement instruments	[1, 4, 5, 7, 13, 15, 24, 27]
Literature reviews of AI literacy	[3, 8, 16, 17, 21, 25, 29]
AI security, risk and trustworthiness frameworks	[9, 20, 22, 23, 26]

D Participant Demographics

Table 3: Sample demographics (N=740).

Age		**Gender**	
18–19	0.7%	Female	45.8%
20s	15.9%	Male	54.2%
30s	16.5%	**Education**	
40s	21.9%	Middle School	2.3%
50s	25.0%	High School	27.7%
60s	20.0%	Jr College/Tech/Vocational	18.6%
		University	45.3%
IT Skills		Graduate School	6.1%
Low	4.7%	**AI Use**	
Basic	23.9%	Used last year	58.0%
Medium	46.8%	Used before	2.3%
High	24.6%	Never	39.7%

E EFA and CFA Results Detail

Table 4: EFA loadings for the AI Safety Literacy Scale.

Item	F1	F2	F3	α
SAIU4	.917	.110	-.171	.949
SAIU8	.862	.040	-.050	
SAIU3	.848	.014	.047	
SAIU2	.825	-.003	.059	
SAIU1	.803	-.035	-.040	
SAIU6	.792	-.053	.077	
SAIU5	.769	.038	.085	
SAIU7	.611	-.055	.167	
AIOV1	-.078	.948	-.027	.884
AIOV3	.100	.828	-.062	
AIOV2	-.086	.794	.094	
AIOV4	.202	.522	.099	
RAIU3	-.027	.074	.804	.831
RAIU1	.132	.055	.674	
RAIU2	.202	.151	.522	

F1=SAIU, F2=AIOV, F3=RAIU

Table 5: CFA loadings for the AI Safety Literacy Scale.

Item	**Estimate**	**Std. Est.**	**SE**	z	p
SAIU7	0.873	.716	.054	16.105	$<.001$
SAIU1	1.012	.827	.050	20.143	$<.001$
SAIU8	0.984	.839	.048	20.635	$<.001$
SAIU6	1.000	.846	–	–	–
SAIU5	1.045	.850	.050	21.099	$<.001$
SAIU4	1.016	.854	.048	21.289	$<.001$
SAIU2	1.050	.868	.048	21.891	$<.001$
SAIU3	1.098	.892	.048	23.048	$<.001$
AIOV4	0.842	.740	.054	15.580	$<.001$
AIOV1	1.000	.813	–	–	–
AIOV2	1.037	.835	.057	18.317	$<.001$
AIOV3	1.070	.861	.056	19.076	$<.001$
RAIU1	1.000	.758	–	–	–
RAIU3	1.033	.775	.067	15.413	$<.001$
RAIU2	1.074	.823	.065	16.522	$<.001$