Human Factors in Cybersecurity, Vol. 168, 2025, 12–21
https://doi.org/10.54941/ahfe1006137

Charting Trustworthiness: A Socio-Technical Perspective on AI and Human Factors

Theofanis Fotis[1,2], Kitty Kioskli[1], and Eleni Seralidou[1]

[1]Trustilio B.V., Vijzelstraat 68, 1017 HL Amsterdam, The Netherlands
[2]University of Brighton, School of Education, Sport & Health Sciences, Brighton, BN19PH, United Kingdom

ABSTRACT

The integration of AI into critical decision-making environments, including cybersecurity, highlights the importance of understanding human factors in fostering trust and ensuring safe human-AI collaboration. Existing research emphasizes that personality traits, such as openness, trust propensity, and affinity for technology, significantly influence user interaction with AI systems, impacting both trustworthiness and reliance behaviours. Furthermore, studies in cybersecurity underscore the socio-technical nature of threats, with human behaviour contributing to a significant portion of breaches. Addressing these insights, the study discusses the development and validation of a questionnaire designed to assess personality-driven factors in AI trustworthiness, advancing tools to mitigate human-centric risks in cybersecurity. Building on interdisciplinary foundations from cyberpsychology, human-computer interaction, and behavioural sciences, the questionnaire evaluates dimensions including ethical responsibility, collaboration, technical competence, and adaptability. Items were systematically reviewed by subject matter experts to ensure face and content validity, reflecting theoretical and empirical insights from prior studies on human behaviour and cybersecurity resilience. The tool's scoring system employs weighted Likert-scale responses, facilitating nuanced assessments of trust dynamics and highlighting areas for intervention. By bridging theoretical and applied perspectives, this research contributes to advancing the role of human factors in cybersecurity, offering actionable insights for the design of trustworthy AI systems and calibrated trust practices.

Keywords: Artificial intelligence, Human factors, Trustworthiness, Sociotechnical, Human traits, Cybersecurity, Co-creation

INTRODUCTION

The rapid integration of Artificial Intelligence (AI) into critical decision-making environments has introduced both opportunities and challenges, particularly in promoting trust and ensuring effective human-AI collaboration. AI-enabled systems are increasingly deployed in sectors such as cybersecurity, healthcare, port management, journalism, finance, and governance, where trust in AI is essential for adoption and reliability (Bach et al., 2022). However, research highlights that trust in AI is influenced

by multiple socio-technical factors, including user personality traits, system transparency, and the socio-ethical considerations surrounding AI deployment (Riedl, 2023).

Existing frameworks for AI trustworthiness emphasize the importance of moving beyond purely technical approaches to embrace a human-centred perspective. The field of Human-Computer Interaction (HCI) provides valuable insights into designing AI systems that align with user needs and expectations, ensuring that AI technologies support appropriate reliance rather than blind trust or scepticism (Bach et al., 2022).

Furthermore, studies in cybersecurity emphasize the socio-technical nature of threats, where human behaviour contributes significantly to breaches. Research indicates that personality traits, such as openness to experience, trust propensity, and affinity for technology, shape how users interact with AI, impacting trust formation and reliance behaviours (Kuper & Kramer, 2024). Addressing these human factors is essential to designing AI systems that mitigate risks while enhancing user confidence.

As AI continues to permeate high-stakes domains such as cybersecurity, healthcare, port management, journalism, finance, and governance, the issue of trustworthiness has emerged as a critical concern. Trust in AI systems determines not only user adoption but also appropriate reliance on AI-generated recommendations, particularly in decision-support environments (Bach et al., 2022). Unlike human-to-human trust, AI trust is influenced by a combination of socio-ethical considerations, technical and design features, and user psychological traits. The challenge lies in cultivating balanced trust, ensuring that users neither over-rely on AI to the point of complacency nor dismiss it out of scepticism (Kuper & Kramer, 2024).

AI trustworthiness is inherently a socio-technical construct, meaning that it depends not only on the algorithmic accuracy of AI models but also on how these models interact with human cognitive, social, and ethical expectations. Research HCI suggests that AI systems should go beyond technical-centric approaches and embrace a human-centred design to ensure trustworthiness. For instance, AI transparency and explainability are essential for achieving user confidence, as black-box models often fail to gain user trust (Bach et al., 2022).

Personality Traits and AI Trust

Trust in AI systems is a crucial factor influencing their acceptance and effective utilization across different sectors. The extent to which users trust AI is shaped by multiple human factors, primarily categorized into socio-ethical considerations, technical and design attributes, and individual user characteristics (Lee et al., 2021). User characteristics play a significant role in shaping perceptions of AI trustworthiness. Personality traits have been found to influence trust levels, with certain individuals demonstrating a higher propensity to trust AI than others (Zhou et al., 2020). For instance, individuals with lower openness to new experiences tend to exhibit greater trust in AI, as do those with a higher inclination towards neuroticism. Moreover, gender differences have been observed, with women generally

displaying higher trust levels (Morana et al., 2020). User attitudes, including their willingness to adopt AI, expectations, and perceptions, further shape trust. Overcoming the digital divide, which refers to the disparity in technical competence and motivation to engage with AI, is vital for trust-building (Klumpp & Zijm, 2019).

Reflecting on the important role of the human traits as discussed above and given that AI trust is neither uniform nor static, incorporating personality-driven insights into AI trustworthiness frameworks is essential for designing adaptive, human-centred systems that promote appropriate reliance rather than blind acceptance or unwarranted scepticism. By capturing personality traits and linking them to AI trust and risk assessment, this study aims to enhance the development of AI technologies that align with diverse user expectations, foster confidence, and mitigate socio-technical risks in critical decision-making environmentA key contribution of this study is the development of an AI Trustworthiness Questionnaire *'TrustSense'*, designed to measure how personality-driven factors influence AI trust. A well-designed AI Trustworthiness Questionnaire can provide empirical insights into how personality-driven trust behaviours shape AI interactions, allowing researchers and practitioners to refine AI interfaces, enhance transparency, and develop adaptive trust-building strategies. By integrating psychological, ethical, and technical considerations, such a tool would contribute to a more responsible and user-aligned AI ecosystem, ultimately fostering appropriate reliance and risk-aware adoption.

METHODS

The development of the initial questionnaire was guided by a comprehensive literature review, incorporating insights from previous studies on AI trustworthiness, cybersecurity risk assessment, and human-AI interaction (Kioskli & Polemi, 2020, 2021, 2022) and expert panel discussions (n = 4). The first version of the questionnaire was structured around individual psychological and behavioural traits that impact AI trust by identifying relevant constructs. These constructs were derived through the literature review as above, from theories of trust in AI and automation, which emphasize personality traits like openness, conscientiousness, and trust propensity in shaping user attitudes towards AI adoption. Human computer interaction principles, particularly regarding explainability, transparency, and user adaptability in AI-based decision-making. Cybersecurity behaviour models, focusing on how individuals manage risks associated with AI-driven systems, including adherence to security protocols and ethical AI use. Based on the identified constructs, an initial pool of 24 items was generated, covering 13 key dimensions: Proactivity and Threat Awareness, Responsibility and Ethics, Innovation and Adaptability, Resilience, Collaboration, Integrity, Technical Proficiency, Policy Adherence, Openness to Interventions. Each item was formatted as a 5-point Likert scale

statement, with reverse-coded questions included to identify inconsistencies in responses.

The initial version of the questionnaire underwent face and content validity assessment through a dedicated workshop. This workshop was part of the regular meetings conducted within the FAITH project (see Acknowledgments) and involved project partners (n = 35), including AI experts, developers, and researchers from the fields of social sciences and ethics, as well as professionals from the healthcare, port, and media industries.

Participants were given access to an online version of the questionnaire in advance, allowing them sufficient time to review its content. During the workshop, the authors facilitated discussions, soliciting both general and item-specific feedback regarding the questionnaire's relevance, clarity, practical utility, comprehensiveness, and acceptability. Additionally, participants deliberated on the contexts in which the questionnaire might be administered.

To ensure a thorough evaluation, discussions encouraged participants to explore emerging perspectives in greater depth, propose refinements, and interconnect their insights, encouraging a more holistic and well-rounded critique of the questionnaire.

More specifically, the participants commented reviewed the questionnaire for Relevance (relevant content and questions), Clarity (easy to understand questions), Completeness (missing key dimensions) and Redundancy (unnecessary or repetitive items),

All their comments have been considered for the revision of the questionnaire, in the thematic areas of a) clarification of ambiguous wording to improve user comprehension, b) reorganization of dimensions to better reflect AI team roles and responsibilities and c) refinement of scoring and interpretation methods.

In addition, the consensus of the participants was that the benefit of assessing the trustworthiness of the AI participants may be particularly relevant for some sectors or application domains, and less relevant for others. For example, in sectors or contexts with a broad range of AI users that have not undergone initial filtering or assessment, such assessments may be particularly useful. In sectors or contexts with trained and selected personnel, requirement assessments may already be in place through organizational measures and there is, hence, not required to conduct this specifically for an AI trustworthiness assessment. In such cases, an assessment of AI readiness at a team or organizational level may be more relevant.

As a result, the final questionnaire transitioned from an individual self-assessment to a team-wide AI trustworthiness maturity model, assessing the collective responsibility of AI teams in ensuring ethical, transparent, and secure AI operations with key changes (Table 1).

Table 1: Key changes in questionnaire focus post-validation.

Pre-Validation (Individual Trust Factors)	Post-Validation (Team-Wide AI Trustworthiness)
Personal ethical behaviours	Team-wide ethics enforcement and AI governance
Individual risk awareness	Collective threat assessment and incident response
Individual AI adaptability	Organizational AI maturity and governance models
Individual cybersecurity hygiene	Team-wide AI compliance and security readiness
Personal collaboration habits	Structured team knowledge sharing and resilience strategies

Following the workshop and revision of the initial questionnaire, the final proposed TrustSense questionnaire Measuring Organizational AI Trust Maturity is presented.

The final validated questionnaire measures team-wide AI trustworthiness maturity across the following dimensions: Proactivity and Threat Awareness: how well the team identifies, assesses, and mitigates AI-related risks. Responsibility and Ethics: the extent to which the team collectively upholds AI ethics and compliance with legal standards. Innovation and Adaptability: the organization's capacity to continuously improve AI trustworthiness through iterative learning and technological enhancements. Resilience: how effectively the team recovers from AI-related incidents and maintains operational stability. Collaboration and Knowledge Sharing: how well teams exchange insights, train personnel, and strengthen AI security measures. Technical Proficiency: the team's ability to critically assess AI outputs, recognize biases, and maintain AI system integrity. Policy Adherence: organizational compliance with AI regulatory frameworks, internal governance policies, and industry best practices.

Proactivity and Threat Awareness

1. The team understands the technological, social and compliance requirements of the multidimensional aspects of AI trustworthiness (cybersecurity, privacy, quality, robustness, transparency, explicability etc).
2. The team routinely identifies at potential technological, operational or social AI threats.
3. In scenarios where an AI threat is exploited or an AI incident occurs, the team acts to mitigate it in line with the requirements of their quality management system.
4. The team understands the organization's security and AI policy, including its commitments and objectives; the team is aware of the quality objectives that apply to their specific roles and responsibilities; and understands how nonconformities can negatively affect the AI operations and how these impact their business.

Responsibility and Ethics

5. The team collectively ensures that all members understand their roles in maintaining AI trustworthiness, holding routine review sessions as part of internal quality reviews.
6. The team consistently prioritizes adherence to AI trustworthiness best practices, directives, standards and guidelines, even during high-pressure scenarios.
7. The team knows the intended use of the AI systems that they operate, their normal operation and their expected outcomes.

Innovation and Adaptability

8. The team has an established routine for implementing or enhancing new mitigation actions (e.g., technological control, policy, procedure) to address AI trustworthiness challenges creatively.
9. If faced with a significant error, the team collectively develops a revised process to prevent recurrence, shared in accordance with the requirements of the quality management system.

Resilience

10. The team ensures effective recovery and organizational business continuity within the first 24 hours after an incident, consistently meeting project deadlines and maintaining a success rate above 90%.
11. Technological failures do not lead to a drop in team performance metrics (Reverse-coded).

Collaboration

12. The team collaborates effectively and has an established routine for sharing new key insights or data points that may address technological threats.
13. The team builds professional relationships with internal and external partners, encouraging meetings to enhance coordination and enhance the threat intelligence.

Integrity

14. The team consistently upholds ethical principles, legal compliance and adheres to professional codes of conduct in its operations.

Technical Proficiency (Questionnaire 1: Technical Users)

15. The team demonstrates proficiency in managing data quality (e.g. data wrangling, distributed databases for handling large datasets, understanding how to create high-quality, unbiased synthetic datasets) by conducting routine audits of datasets and their use.
16. The team applies advanced technological tools (e.g. optimising AI models; knowledge and tools to ensure model transparency, interpretable models, protected models from adversarial attacks; reduction of algorithmic biases, privacy, auditability, robustness) in ongoing projects where this is required.

Problem Solving

17. The team is skilled in resolving issues, completing 90% of identified challenges through interdisciplinary collaboration.

Resource Accessibility

18. The team has access to high-performance computing tools and networks, routinely engaging in sessions with external experts to enhance capabilities.
19. Limited interaction with external technological communities is detrimental to team progress (Reverse-coded).

Policy Adherence

20. The team adheres to policies by maintaining a compliance score on Trustworthy AI above 95% during regular audits.

Motivation and Commitment

21. The team consistently demonstrates a commitment to trustworthy AI by organizing regular ethical reviews and discussions and attend professional trainings.

Privacy and Compliance

22. The team prioritizes privacy and legal compliance for trustworthy AI by achieving at least 95% adherence in internal audits.

Openness to Interventions

23. The team welcomes external feedback, routinely attending training sessions annually to refine practices.
24. Resistance to changes in workflows that enhance trustworthiness is a challenge (Reverse-coded).

Scoring and Interpretation

Each question is scored on a Likert scale (Strongly Disagree = 1, Disagree = 2, Neutral = 3, Agree = 4, Strongly Agree = 5). Reverse-coded items are adjusted to ensure consistency.

The calculation methodology ensures that individual responses contribute to a collective organizational score. The process involves the following steps:

1. Weight Assignment: Each dimension of trustworthiness is assigned a weight based on its criticality to the organization. Example weights:

- Responsibility and Ethics: 25%
- Technical Proficiency: 20%
- Collaboration: 20%
- Proactivity and Threat Awareness: 15%
- Privacy and Compliance: 20%

2. Normalization: To standardize the scores, a logarithmic transformation is applied if the data distribution shows significant skewness. This step ensures comparability across dimensions.

3. Overall Organizational Score: The weighted scores for all Questions are summed to produce the overall organizational trustworthiness score.

4. Categorization: The final score is categorized into trustworthiness levels using predefined thresholds: 4.5–5: Very High, 3.5–4.49: High, 2.5–3.49: Moderate, 1.5–2.49: Low, 1–1.49: Very Low, > 1: Negligible.

The overall score can then be utilized by the organization by following the mitigation recommendations of the TrustSense questionnaire (Table 2):

Table 2: Mitigation recommendations of the TrustSense questionnaire.

Likelihood	Scoring	Interpretation of Results	Trustworthy AI Maturity for Teams	Mitigation Recommendations
Very High	4.5–5	The team demonstrates consistently high maturity regarding trustworthy AI	Very High	To maintain this level, organize regular team training sessions, recognize collective achievements, and promote a culture of continuous improvement.
High	3.5–4.49	The team largely adheres to requirements for trustworthy AI, with minor areas for improvement.	High	Enhance organizational training programs, encourage cross-team collaborations, and refine adherence to ethical codes and organizational policies to elevate performance.
Medium	2.5–3.49	The team shows partial adherence to requirements for trustworthy AI, indicating areas needing attention.	Moderate	Implement structured training initiatives, strengthen collaborative practices, and promote organizational mentorship to address identified gaps.
Low	1.5–2.49	Significant gaps in trustworthy AI maturity exists at the team's level.	Low	Facilitate intensive team workshops, prioritize ethical compliance, and establish policies to strengthen trustworthiness practices across teams.
Very Low	1–1.49	The team faces considerable challenges in trustworthy AI maturity.	Very Low	Commit to comprehensive retraining programs, monitor collective progress through evaluations, and establish supervised practices to rebuild foundational trustworthiness traits.
Negligible	<1		Negligible	

CONCLUSION

To the authors knowledge, this is the first attempt to measure the AI trustworthiness maturity of an organisation, reflecting on individual behaviour and human traits. The development of the questionnaire was based on strong theoretical ground and existing work of the authors, reflecting also the opinions of the target population as these were shared during the validation workshop. AI trustworthiness is not solely determined by

individual behaviours but is shaped by organizational policies, collective team responsibility, and institutional practices, necessitating a team-wide measurement framework to assess AI trustworthiness maturity at the organizational level rather than relying on personal perceptions (Schaschek & Engel, 2023). Trust in AI also depends on systemic factors such as governance structures, team dynamics, and organizational risk management (Autio et al., 2024).

Organizations can utilize the TrustSense Questionnaire scoring to systematically assess their AI trustworthiness maturity at a team level, identifying strengths and areas for improvement. By interpreting the scores, organizations can implement targeted mitigation strategies that align with their current maturity level, ensuring a structured approach to enhancing AI governance, ethical compliance, and team-wide accountability. Higher-scoring teams can maintain their maturity through continuous training and reinforcement of best practices, while lower-scoring teams can benefit from intensive workshops, structured mentorship, and policy enhancements to strengthen trustworthiness. The overall score serves as a benchmark for ongoing evaluation, guiding organizations in cultivating a culture of responsible AI adoption and continuous improvement.

Moving forward, the psychometric properties of this questionnaire must be tested by administering it to pilot samples from relevant organizations. The completed questionnaires will allow for an assessment of reliability. Additionally, it is recommended to evaluate and finalize the construct validity and scoring system.

Furthermore, a feasibility study should determine whether this questionnaire can be implemented across various industries beyond those represented by the initial participants (healthcare, media, and port services). More research is also needed to identify the optimal context for assessing organizational AI trustworthiness maturity, particularly in relation to the psychological and behavioural traits of individual team members.

ACKNOWLEDGMENT

The authors would like to acknowledge the financial support provided for the following project ‘Fostering Artificial Intelligence Trust for Humans towards the Optimization of Trustworthiness through Large-scale Pilots in Critical Domains’ (FAITH) project, which has received funding from the European Union’s Horizon Programme under grant agreement No. 101135932. The views expressed in this paper represent only the views of the authors and not those of the European Commission or the partners in the above-mentioned projects. Finally, the authors declare that there are no conflicts of interest, including any financial or personal relationships, that could be perceived as potential conflicts.

REFERENCES

Alisa Küper & Nicole Krämer (2024): Psychological Traits and Appropriate Reliance: Factors Shaping Trust in AI, International Journal of Human–Computer Interaction, doi: 10.1080/10447318.2024.2348216.

Autio, C. et al. (2024) Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST. Available at: https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence (Accessed: 17 March 2025).

Kioskli, K., Polemi, N. (2022): 'Estimating attackers' profiles results in more realistic vulnerability severity scores', in Ahram, T. and Karwowski, W. (eds), Human Factors in Cybersecurity, AHFE International Conference, AHFE Open Access, Vol. 53, AHFE International, 2022, http://doi.org/10.54941/ahfe1002211.

Kioskli K, Polemi N, (2020): "A socio-technical approach to cyber risk assessment." International Journal of Electrical and Computer Engineering, 14(10), pp. 305–309.

Kioskli K, Polemi N, (2021): "Measuring psychosocial and behavioural factors improves attack potential estimates." In Proceedings of the 15th International Conference for Internet Technology and Secured Transactions, pp. 216–219."

Klumpp, M., & Zijm, H. (2019). Logistics innovation and social sustainability: How to prevent an artificial divide in human–computer interaction. Journal of Business Logistics, 40(3), 265–278. https://doi.org/10.1111/jbl.12198

Morana, Stefan; Gnewuch, Ulrich; Jung, Dominik; and Granig, Carsten, "The Effect of Anthropomorphism on Investment Decision-Making with Robo-Advisor Chatbots" (2020). In Proceedings of the 28th European Conference on Information Systems (ECIS), An Online AIS Conference, June 15–17, 2020. https://aisel.aisnet.org/ecis2020_rp/63

Lee, M., Frank, L., & IJsselsteijn, W. (2021). Brokerbot: A cryptocurrency chatbot in the social-technical gap of trust. Computer Supported Cooperative Work (CSCW), 30(1), 79–117. https://doi.org/10.1007/s10606-021-09392-6

Riedl, R. (2022): Is trust in artificial intelligence systems related to user personality? Review of empirical evidence and future research directions. Electron Markets 32, 2021–2051. https://doi.org/10.1007/s12525-022-00594-4

Schaschek, Myriam and Engel, Sarah, (2023) "Measuring Trustworthiness of AI Systems: A Holistic Maturity Model". Rising like a Phoenix: Emerging from the Pandemic and Reshaping Human Endeavours with Digital Technologies ICIS 2023. 7.

Tita Alissa Bach, Amna Khan, Harry Hallock, Gabriela Beltrão & Sonia Sousa (2022): A Systematic Literature Review of User Trust in AI-Enabled Systems: An HCI Perspective, International Journal of Human–Computer Interaction, doi: 10.1080/10447318.2022.2138826.

Zhou, J., Luo, S., & Chen, F. (2020). Effects of personality traits on user trust in human–machine collaborations. Journal on Multimodal User Interfaces, 14(4), 387–400. https://doi.org/10.1007/s12193-020-00329-9