The Trustworthiness Assessment Model – A Micro and Macro Level Perspective

Nadine SCHLICKER[a,1] and Markus LANGER[a]
[a] *Philipps-University of Marburg, Germany*
ORCiD ID: Nadine Schlicker https://orcid.org/0000-0003-0368-3081, Markus Langer https://orcid.org/0000-0002-8165-1803

Abstract. In this paper we shed light on the trustworthiness assessment in human-AI teams by introducing the Trustworthiness Assessment Model which explicates the transition from a system and its characteristics to the human's perception of this system at a micro and at a macro level.

Keywords. Trustworthy AI, Trust in Automation, HCI, Human-AI Teams

1. Introduction and Related Work

For a successful interaction between human AI teams (HATs), the right amount of trust in AI is important [17]. Trust is based on the human trustor's assessment of an AI's trustworthiness for a particular task and context. This trustworthiness assessment forms the basis for subsequent trust and trusting behavior in an AI system. Especially in larger HAT's this trustworthiness assessment may differ between team members [15] and affect the team's collective performance. The trustworthiness assessment is an important starting point to better understand where differences may arise.

The concept of trustworthiness has been used to refer to two sides of the same coin [14,17–19,29,33,34]. First, trustworthiness has been referred to as an „objective attribute of the trustee" [35]; see also [9,12,20]. Second, trustworthiness has been referred to as a subjective perception of a trustee's attributes [19]. Research has shown that the latter, the trustor's perceived trustworthiness (PT), does not necessarily accurately reflect the former, the system's actual trustworthiness (AT) [3,23]. Frequently cited trust models often start at the point where humans have already built up their PT [17,19]. With the Trustworthiness Assessment Model (TrAM), we add the system to these models. By explicating the transition from AT to PT, we aim to understand which factors influence the accuracy of the trustworthiness assessment at the micro level (Fig. 1, see also [30] for more details) and at the macro level (see Fig. 2).

2. Micro Level of the Trustworthiness Assessment Model (TrAM)

The TrAM consists of the AT of a system, system characteristics, trustor's individual standards, PT of a system, and cues, which are specified in the following.

[1] Corresponding Author: Nadine Schlicker, nadine.schlicker@uni-marburg.de

2.1. Actual Trustworthiness

We define a system's AT as a latent construct indicating the true value of a system's trustworthiness for a specific trustor [8]. AT is user- and context-specific and depends on the system characteristics and on the individual standards of trustworthiness. In our model *system characteristics* are (only theoretically available) context-free facts that subsume everything that could theoretically be ascertained about a system and which answer the question: "What are the characteristics of this system?" *Individual standards* answer the question: "What makes a system trustworthy for me?" We can think of individual standards as a requirement list that contains all factors constituting a perfectly trustworthy system for the trustor (Fig. 1). Individual standards are influenced by trustors' goals and interests, their cultural background , ethical values [31], as well as the normative and regulatory frame in which trustors operate. To conclude, AT reflects the degree to which the *system characteristics* match a trustor's *individual standards* for trustworthiness with respect to a specific task and point in time. AT answers the question: "*How trustworthy is the system actually with respect to my individual standards*?" In the requirement list metaphor, AT reflects how many checkboxes are marked on the "individual standards requirement list" if perfect assessment was possible - which never is and which highlights the importance of PT (Fig. 1).

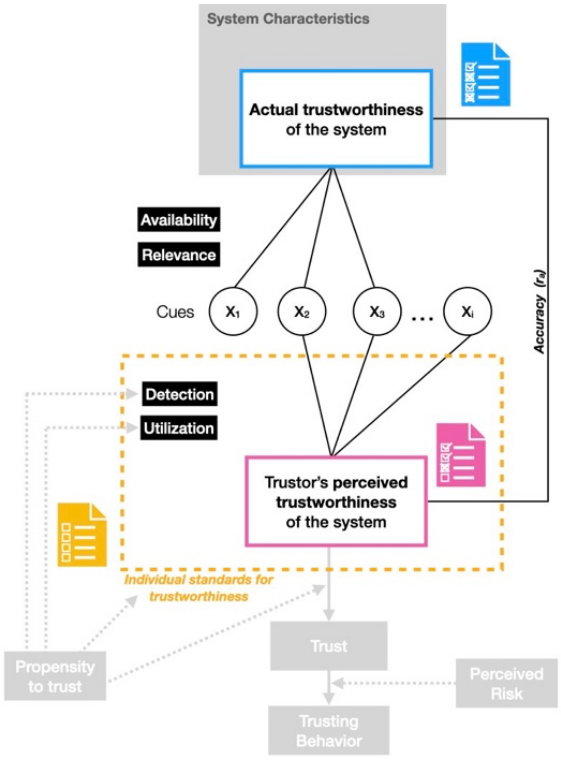

Fig. 1. The micro level of the TrAM showing the relation between AT and PT. The lines between cues and AT, and between cues and PT represent correlations. Cues are everything a trustor uses to infer the system's AT; they are thus not necessarily (positively) correlated with the AT. The list icons reflect the relation between individual standards (bottom left, yellow), AT (top right, blue), and PT (center right, purple). Grey boxes (bottom) show how TrAM integrates into the existing trust by Mayer et al. [19].

2.2. Perceived Trustworthiness

PT is the result of a trustor's assessment of actual system trustworthiness. Consequently, PT is fundamental for answering the question "*How trustworthy do I think the system is with respect to my individual standards?*". PT corresponds to another latent construct that is not directly observable, but that can be measured indirectly (e.g. by asking people to report on their PT of a system, [1,25] or by observing human-system interactions [28,36]).

2.3. Cues

Cues are pieces of information that presumably provide insights regarding the AT of a system [4,18,33]. Single cues only provide narrow or even misleading insights regarding a system's AT, and each cue relates to a certain degree to the system's AT. Thus, trustors are constantly (consciously or unconsciously) searching for, confronted with, using, and interpreting cues to assess systems' AT. Cues can e.g. be the aesthetics of a user interface, indicated predictive power of a classifier, a "Trustworthy AI" seal, or a company's logo. Cues can also stem from other people (e.g. testimonies of co-workers, see Fig. 2).

2.4. Relations Between the Model Components and Influencing Factors

The trustworthiness assessment is accurate when the trustor's PT matches the system's AT. Building on Funder's [8] model explaining how humans assess the characteristics of other humans, this accuracy depends on *relevance* and *availability* of cues on the system's side, and on *detection* and *utilization* of cues on the trustor's side. On the system's side, *cue relevance* defines how indicative a cue is for the AT of a system. Relevant cues (e.g. a system's performance in a task) correlate strongly with the AT of a system. Less relevant cues (e.g. popularity of a brand) correlate low with the AT [12,27]. *Cue availability* refers to the fact that cues can only be detected when they are accessible to the trustor. For example, the training data quality might be strongly related to the AT of a system. However, users might not have access to such information without digging deep into the system's technical documentation. On the trustor's side*, cue detection* means that relevant and available cues must be detected. Potential influencing factors of cue detection are interface properties such as low contrast or brightness, but also trustor's mood [2,7,21], attention capacities [11], situation awareness [6], time pressure [24], or experience with a system [32]. *Cue utilization* means that trustors need to correctly interpret a relevant, available, and detected cue. So, trustors need to weigh the detected information appropriately. Accuracy of trustworthiness assessment should be high when *relevant* cues are *available*, *detected* and *utilized* adequately.

3. Macro Level of the Trustworthiness Assessment Model

The micro level of TrAM explicates the process through which humans arrive from a system's AT at their PT. However, especially in larger HATs the trustworthiness assessment does not happen in isolation, but is embedded in a societal and social context [5,16,18,22]. Thus, we propose a) that there exists a chain of stakeholders who form their own PT of the system through micro-level trustworthiness assessment processes, and b)

that those stakeholders produce, what we call, *secondary cues* which can be used by other stakeholders to form their PT of the system. Secondary cues result from a stakeholder's trustworthiness assessment; they provide information about the system, but they do not stem from the system, e.g. testimonies of colleagues or a seal indicating "Trustworthy AI". *Primary* cues, on the other hand, stem directly from the system, e.g. performance indicators of the AI or training and test data.

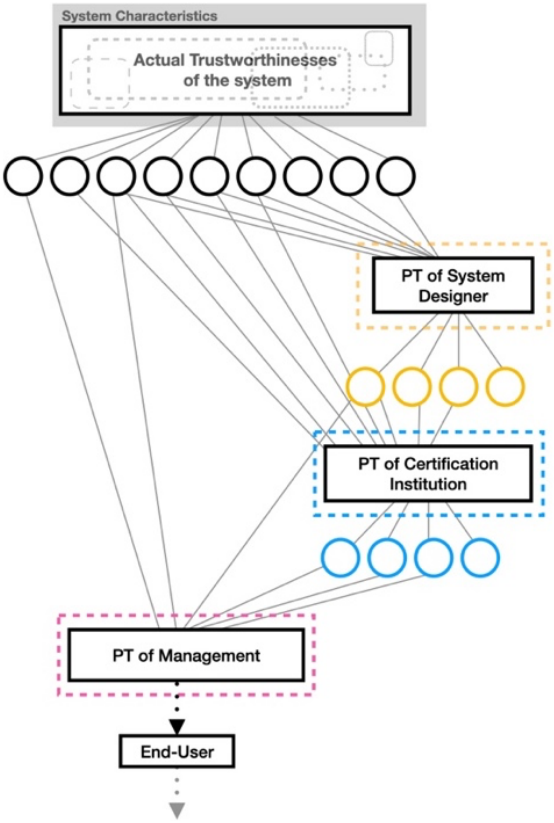

Fig. 2. The macro level of TrAM showing a sample trustworthiness propagation process that reflects a chain of micro-level trustworthiness assessments. Different trustors assess a system's AT to arrive at their PT and produce secondary cues that can be used by other trustors. The dashed lines in the AT-box indicate that there are as many actual trustworthinesses as there are trustors along the trustworthiness propagation process.

In this *trustworthiness propagation process,* different stakeholders (e.g. system designers, certification institutions, auditors, team members, users) assess their PT of the system with regard to their individual standards based on cues available to them. This may produce new secondary cues (e.g. labels, system handbooks, testimonials) for other stakeholders assessing system trustworthiness (Fig. 2). The trustworthiness propagation process is thus a sequence of different trustors assessing system trustworthiness. It is neither a pipeline nor does it involve a strict hierarchy; rather it forms a complex social network of stakeholders. The provision of secondary cues may neither be a completely intentional nor planned process nor an always accurate process. A prototypical example for an intentional secondary cue is the provision of (trustworthiness) labels. Less

intentional secondary cues might emerge in social interactions (e.g. a colleague reporting about system use). Such cues may suffer from pitfalls of (mis)communication between stakeholders [10,26]. Secondary cues can also be intentionally incomplete (e.g. due to adhering to data protection regulations), unintentionally wrong due to an inaccurate assessment of system trustworthiness, or intentionally wrong in terms of deception and fraud, such as faked data in the Volkswagen emission scandal [13]. Stakeholders' individual standards shape their assessment of system trustworthiness and their provision of secondary cues. Consequently, for each stakeholder involved, it is important to understand their individual standards and the secondary cues they may produce, since they might influence other stakeholders' trustworthiness assessments.

4. Implications and Future Research

At the micro level, TrAM explicates factors that are important for an accurate trustworthiness assessment. Especially, in HATs it is crucial to analyze whether there are enough relevant cues available that allow an accurate assessment in light of different individual standards, tasks, and application contexts. User research might help to map different stakeholders as well as their individual standards. Detection might be enhanced through human-centered system design. Furthermore, TrAM helps to investigate whether the members of HATs are sufficiently trained to detect and utilize the cues appropriately.

At the macro level TrAM can help to systematically analyze the trustworthiness propagation process and the primary and secondary cues that might occur in the system life cycle. Those cues might be analyzed in terms of their relevance in light of different individual standards and how heavily they influence the accuracy of the trustworthiness assessment of other stakeholders. It might enable to analyze whether and where the individual standards and goals of different involved stakeholders may (mis)align.

In HATs the TrAM might help to identify different individual standards of human and artificial team members. Furthermore, TrAM might help to investigate, which cues, primary and secondary, are detected and utilized to what extend to assess the AT of an artificial team member in larger HATs. The comparison between trustworthiness assessments in larger HATs and dyads might provide insights on how a joint team assessment influences the accuracy of the trustworthiness assessment.

5. Conclusion

In this paper, we introduced the TrAM, described its micro-level trustworthiness assessment process explaining how trustors assess a system's AT to arrive at their PT of the system, and we described the model's macro-level trustworthiness propagation process that emphasizes that different stakeholders prompt cues that other stakeholders may use to assess system trustworthiness. With this, we explicate a process that focuses on trustworthiness as a basis for trust and a model that augments existing trust models by adding the trustee and their characteristics. For HATs we hope, that the TrAM provides a fruitful starting point to focus on the trustworthiness assessment and differences that occur between team members regarding their individual standards and the detection and utilization of cues. Future research could investigate how individual trustworthiness assessments differ between dyads and larger HAT's.

References

[1] Lamia Alam and Shane Mueller. 2021. Examining the effect of explanation on satisfaction and trust in AI diagnostic systems. *BMC Medical Informatics and Decision Making* 21, 1 (June 2021), 178. DOI:https://doi.org/10.1186/s12911-021-01542-6

[2] Antoine Bechara, Hanna Damasio, Daniel Tranel, and Antonio R Damasio. 1997. Deciding advantageously before knowing the advantageous strategy. *Science* 275, 5304 (February 1997), 1293–1295. DOI:https://doi.org/10.1126/science.275.5304.1293

[3] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, ACM, Cagliari Italy, 454–464. DOI:https://doi.org/10.1145/3377325.3377498

[4] George J. Cancro, Shimei Pan, and James Foulds. 2022. Tell me something that will help me trust you: A survey of trust calibration in human-agent interaction. Retrieved July 5, 2022 from http://arxiv.org/abs/2205.02987

[5] Erin K. Chiou and John D. Lee. 2021. Trusting automation: Designing for responsivity and resilience. *Hum Factors* (April 2021), 001872082110099. DOI:https://doi.org/10.1177/00187208211009995

[6] Mica R. Endsley. 2017. From here to autonomy: Lessons learned from human–automation research. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 59, 1 (February 2017), 5–27. DOI:https://doi.org/10.1177/0018720816681350

[7] Joseph P. Forgas. 1995. Mood and judgment: The affect infusion model (AIM). *Psychological Bulletin* 117, (1995), 39–66. DOI:https://doi.org/10.1037/0033-2909.117.1.39

[8] David C Funder. 1995. On the accuracy of personality judgment:A realistic approach. *Psychological Review* 102, 4 (1995), 652–670. DOI:https://doi.org/10.1037/0033-295X.102.4.652

[9] Ben Green. 2022. The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review* 45, (July 2022), 105681. DOI:https://doi.org/10.1016/j.clsr.2022.105681

[10] Mary Ellen Guffey and Dana Loewy. 2012. *Essentials of Business Communication*. Cengage Learning.

[11] Harold Hawkins, Steven Hillyard, Steven Luck, Mustapha Mouloua, Cathryn Downing, and Donald Woodward. 1990. Visual attention modulates signal detectability. *Journal of experimental psychology. Human perception and performance* 16, (December 1990), 802–11. DOI:https://doi.org/10.1037/0096-1523.16.4.802

[12] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ACM, Virtual Event Canada, 624–635. DOI:https://doi.org/10.1145/3442188.3445923

[13] Jae C. Jung and Elizabeth Sharon. 2019. The Volkswagen emissions scandal and its aftermath. *Global Business and Organizational Excellence* 38, 4 (2019), 6–15. DOI:https://doi.org/10.1002/joe.21930

[14] Matthew Kay, Shwetak N. Patel, and Julie A. Kientz. 2015. How good is 85%? A survey tool to connect classifier evaluation to acceptability of accuracy. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), Association for Computing Machinery, New York, NY, USA, 347–356. DOI:https://doi.org/10.1145/2702123.2702603

[15] Taenyun Kim, Maria D. Molina, Minjin (MJ) Rheu, Emily S. Zhan, and Wei Peng. 2023. One AI Does Not Fit All: A Cluster Analysis of the Laypeople's Perception of AI Roles. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (CHI '23), Association for Computing Machinery, New York, NY, USA, 1–20. DOI:https://doi.org/10.1145/3544548.3581340

[16] Bran Knowles and John T. Richards. 2021. The sanction of authority: Promoting public trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ACM, Virtual Event Canada, 262–271. DOI:https://doi.org/10.1145/3442188.3445890

[17] John D. Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors* 46, 1 (2004), 50–80. DOI:https://doi.org/10.1518/hfes.46.1.50_30392

[18] Q. Vera Liao and S. Shyam Sundar. 2022. *Designing for responsible trust in AI systems: A communication perspective*. DOI:https://doi.org/10.1145/3531146.3533182

[19] Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An integrative model of organizational trust. *The Academy of Management Review* 20, 3 (1995), 709–734. DOI:https://doi.org/10.2307/258792

[20] Carolyn McLeod. 2021. Trust. In *The Stanford Encyclopedia of Philosophy* (Fall 2021), Edward N. Zalta (ed.). Metaphysics Research Lab, Stanford University. Retrieved September 7, 2022 from https://plato.stanford.edu/archives/fall2021/entriesrust/

[21] Stephanie M. Merritt. 2011. Affective Processes in Human–Automation Interactions. *Hum Factors* 53, 4 (August 2011), 356–370. DOI:https://doi.org/10.1177/0018720811411912

[22] Guglielmo Papagni, Jesse de Pagter, Setareh Zafari, Michael Filzmoser, and Sabine T. Koeszegi. 2022. Artificial agents' explainability to support trust: considerations on timing and context. *AI & Soc* (June 2022). DOI:https://doi.org/10.1007/s00146-022-01462-7

[23] Andrea Papenmeier, Dagmar Kern, Daniel Hienert, Yvonne Kammerer, and Christin Seifert. 2022. How accurate does it feel? Human perception of different types of classification mistakes. In *CHI Conference on Human Factors in Computing Systems* (CHI '22), Association for Computing Machinery, New York, NY, USA, 1–13. DOI:https://doi.org/10.1145/3491102.3501915

[24] Tobias Rieger and Dietrich Manzey. 2022. Human Performance Consequences of Automated Decision Aids: The Impact of Time Pressure. *Hum Factors* 64, 4 (June 2022), 617–634. DOI:https://doi.org/10.1177/0018720820965019

[25] Tobias Rieger, Eileen Roesler, and Dietrich Manzey. 2022. Challenging presumed technological superiority when working with (artificial) colleagues. *Sci Rep* 12, 1 (March 2022), 3768. DOI:https://doi.org/10.1038/s41598-022-07808-x

[26] Lee Ross, David Greene, and Pamela House. 1977. The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology* 13, 3 (May 1977), 279–301. DOI:https://doi.org/10.1016/0022-1031(77)90049-X

[27] Marie Roy Christine, Olivier Dewit, and Benoit A. Aubert. 2001. The impact of interface usability on trust in Web retailers. *Internet Research* 11, 5 (January 2001), 388–398. DOI:https://doi.org/10.1108/10662240110410165

[28] Max Schemmer, Patrick Hemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. 2022. Should I Follow AI-based Advice? Measuring Appropriate Reliance in Human-AI Decision-Making. *arXiv preprint arXiv:2204.06916}* (2022), 10. DOI:https://doi.org/DOI: 10.5445/IR/1000145647

[29] Nadine Schlicker and Markus Langer. 2021. Towards warranted trust: A model on the relation between actual and perceived system trustworthiness. In *Mensch und Computer 2021*, ACM, Ingolstadt Germany, 325–329. DOI:https://doi.org/10.1145/3473856.3474018

[30] Nadine Schlicker, Alarith Uhde, Kevin Baum, Martin C. Hirsch, and Markus Langer. 2022. Calibrated Trust as a Result of Accurate Trustworthiness Assessment – Introducing the Trustworthiness Assessment Model. DOI:https://doi.org/10.31234/osf.io/qhwvx

[31] Claire Textor, Rui Zhang, Jeremy Lopez, Beau G. Schelble, Nathan J. McNeese, Guo Freeman, Richard Pak, Chad Tossell, and Ewart J. de Visser. 2022. Exploring the Relationship Between Ethics and Trust in Human–Artificial Intelligence Teaming: A Mixed Methods Approach. *Journal of Cognitive Engineering and Decision Making* 16, 4 (December 2022), 252–281. DOI:https://doi.org/10.1177/15553434221113964

[32] Carl Thompson, Len Dalgleish, Tracey Bucknall, Carole Estabrooks, Alison M. Hutchinson, Kim Fraser, Rien de Vos, Jan Binnekade, Gez Barrett, and Jane Saunders. 2008. The Effects of Time Pressure and Experience on Nurses' Risk Assessment Decisions: A Signal Detection Analysis. *Nursing Research* 57, 5 (October 2008), 302–311. DOI:https://doi.org/10.1097/01.NNR.0000313504.37970.f9

[33] Ewart J. de Visser, Marvin Cohen, Amos Freedy, and Raja Parasuraman. 2014. A design methodology for trust cue calibration in cognitive agents. In *Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments*, Randall Shumaker and Stephanie Lackey (eds.). Springer International Publishing, Cham, 251–262. DOI:https://doi.org/10.1007/978-3-319-07458-0_24

[34] Ewart J. de Visser, Marieke M.M. Peeters, Malte Jung, Spencer Kohn, Tyler Shaw, Richard Pak, and Mark Neerincx. 2020. Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics* 12, (May 2020). DOI:https://doi.org/10.1007/s12369-019-00596-x

[35] John Zerilli, Umang Bhatt, and Adrian Weller. 2022. How transparency modulates trust in artificial intelligence. *Patterns* (February 2022), 100455. DOI:https://doi.org/10.1016/j.patter.2022.100455

[36] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (FAT* '20), Association for Computing Machinery, New York, NY, USA, 295–305. DOI:https://doi.org/10.1145/3351095.3372852