AI Governance and Society Journal
ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100
https://galaxiauniverse.com/index.php/AIGSJ/issue/archive

User Trust Modeling in Responsible AI Systems

Oscar Nowak[1], Sofia Schmidt[2], Amelia Lindberg[3]

[1] Professor, Department of Artificial Intelligence, Mediterranean Institute of Technology, Rome, Italy. Email: oscar.nowak19@gmail.com | ORCID: 4056-3252-9884-4237
[2] Research Scientist, Department of Machine Learning, Nordic Technical University, Stockholm, Sweden. Email: sofia.schmidt213@gmail.com | ORCID: 3301-4359-0802-2781
[3] Assistant Professor, Institute of Intelligent Systems, Advanced Computing University, Paris, France. Email: amelia.lindberg314@gmail.com | ORCID: 6366-4018-0720-2917

ABSTRACT

User trust in AI systems -- the degree to which users rely on AI outputs, recommendations, or decisions in ways that are appropriate to the system's actual reliability and limitations -- is a central construct in responsible AI design. Miscalibrated trust, whether excessive (overtrust) or insufficient (undertrust), undermines the beneficial outcomes that AI-assisted decision- making seeks to achieve and increases the risk of harmful over-reliance or unjustified rejection of accurate AI guidance. Despite extensive theoretical treatment of AI trust, validated computational models of user trust that can be integrated into responsible AI systems to personalise oversight, explanation, and interaction strategies remain scarce. This paper proposes the Dynamic Trust Calibration Model (DTCM), a Bayesian trust model that continuously estimates individual user trust levels from interaction signals -- decision patterns, override frequency, explanation engagement, and confidence-accuracy alignment -- and uses these estimates to trigger personalised interventions designed to maintain appropriate trust calibration. The DTCM is evaluated through a longitudinal user study involving 168 participants interacting with an AI clinical advisory system over six weeks across three experimental conditions: DTCM-adaptive interventions, static explanation baseline, and no-explanation control. DTCM-adaptive conditions achieve a 36.8% improvement in trust calibration accuracy (SD = 5.4%), a 28.9% reduction in overtrust incidents (SD = 4.7%), and a 31.4% reduction in undertrust incidents (SD = 4.9%) compared to the static baseline. The study contributes the DTCM specification, a Trust Calibration Index (TCI) measurement instrument, and empirical evidence that dynamic trust modelling produces superior responsible AI interaction outcomes compared to static explanation approaches.

Keywords: user trust modeling; trust calibration; responsible AI; overtrust; undertrust; Bayesian trust model; human-AI interaction; EU AI Act

Citation: Nowak et al. [2025]. User Trust Modeling in Responsible AI Systems. DOI: https://doi.org/10.5281/zenodo.19185042

Article Information: *Received:* 22 May 2025 *Accepted:* 24 July 2025 *Published:* 30 September 2025

Research Article: *Research Article*

1. Introduction

1.1 The Trust Calibration Imperative

Trust in AI systems occupies a peculiar position in responsible AI discourse: it is simultaneously a goal to be promoted -- AI systems must be trustworthy, and users must be able to trust them when they are -- and a risk to be managed, as excessive or misplaced trust produces harmful over-reliance on AI outputs that may be erroneous, biased, or contextually inappropriate (Lee and See, 2004; Bansal et al., 2021). The EU AI Act (2024) implicitly acknowledges this duality: Article 13 transparency obligations are motivated in part by the need to enable users to form accurate assessments of AI system capabilities and limitations, and Article 14 human oversight requirements are designed to ensure that humans can identify and respond to AI failures -- both of which presuppose that users maintain appropriately calibrated rather than excessive or negligent trust. The empirical literature on human-AI trust documents systematic miscalibration in both directions. Skitka et al. (1999) and Cummings (2017) documented automation bias -- overtrust manifesting as failure to detect AI errors that would have been detectable with due attention. Dietvorst et al. (2015) documented algorithm aversion -- undertrust manifesting as systematic rejection of AI forecasts even when they demonstrably outperform human judgement. Schemmer et al. (2023) found in a meta-analysis that standard XAI interventions have inconsistent effects on trust calibration, with effect sizes for objective complementarity ranging from d = -0.31 to d = 0.74 across studies. The inconsistency of XAI effects on trust calibration likely reflects, in part, the individual differences in trust disposition, prior AI experience, and domain expertise that determine how any given explanation affects user behaviour. A static explanation strategy that provides the same transparency information to all users is unlikely to achieve appropriate trust calibration for both overtrusting and undertrusting users simultaneously. Dynamic trust modelling -- continuously estimating individual trust levels and adapting interventions accordingly -- offers a principled approach to this personalisation challenge.

1.2 Research Objectives and Paper Structure

This paper proposes the Dynamic Trust Calibration Model (DTCM), a Bayesian trust model that estimates individual user trust levels from interaction signals and triggers personalised calibration interventions. Research objectives are: (1) to specify the DTCM with its Bayesian estimation model, interaction signal taxonomy, and intervention library; (2) to develop and validate the Trust Calibration Index (TCI); and (3) to evaluate DTCM-adaptive versus static explanation approaches in a longitudinal user study (n = 168, six weeks). Section 2 reviews trust theory, dynamic trust modelling, and responsible AI trust requirements. Section 3 presents the DTCM. Section 4 describes the user study. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Literature Review

2.1 Trust Theory and Human-AI Trust Models

Trust in automation has been modelled as a function of three primary antecedents: system performance (reliability, accuracy), system process (transparency, predictability), and system purpose (the perceived alignment of the system's objectives with user interests) (Lee and See, 2004). Mayer et al.'s (1995) interpersonal trust model -- decomposing trust into ability, benevolence, and integrity perceptions -- has been adapted to AI contexts by Heerink et al. (2010) and Madsen and Gregor (2000), who add predictability and reliability as trust-forming constructs specific to automated systems. Dynamic trust models -- computational models that estimate trust as a time-varying state from behavioural signals -- have been developed primarily in human-robot interaction (HRI) and supervisory control research. Xu and Dudek (2015) proposed a Bayesian online trust model that updates trust estimates from observations of robot behaviour and user responses. Akash et al. (2019) developed a physiological trust model using EEG and eye-tracking signals. These models demonstrate the feasibility of continuous trust estimation but have not been applied in AI advisory system contexts or integrated with adaptive intervention mechanisms. Table 1 maps trust constructs to DTCM model components.

2.2 Trust Calibration Interventions in AI Systems

Interventions to improve trust calibration in AI systems have focused primarily on explanation design. Poursabzi-Sangdeh et al. (2021) found that feature importance explanations reduced appropriate reliance in some conditions. Bansal et al. (2021) demonstrated that explanations of AI limitations -- rather than capabilities -- more effectively reduce overtrust. Zhang et al. (2020)

showed that uncertainty communication -- explicit disclosure of AI confidence levels -- significantly improves trust calibration by enabling users to selectively defer to or override AI recommendations. Yin et al. (2019) demonstrated that accuracy information about AI performance reduces overtrust but may also reduce appropriate trust in high-accuracy systems. The literature consistently shows that no single intervention universally improves trust calibration across all users and contexts -- motivating the DTCM's dynamic, personalised intervention selection approach. The DTCM intervention library synthesises the most empirically effective trust calibration techniques from this literature and adapts their deployment to individual trust state estimates.

Table 1: Trust Construct Mapping -- Prior Theory to DTCM Model Components

Trust Construct	Source	DTCM Estimation Signal	DTCM Intervention Target	Responsible AI Relevance
Ability (competence)	Mayer et al. (1995)	Accuracy-confidence alignment	Uncertainty communication	EU AI Act Art. 13
Benevolence	Mayer et al. (1995)	Perceived goal alignment signals	Value alignment disclosure	EU AI Act Art. 5
Predictability	Madsen and Gregor (2000)	Explanation engagement rate	Consistency demonstration	EU AI Act Art. 13
Performance-based trust	Lee and See (2004)	Override frequency vs. AI error	Performance calibration feedback	EU AI Act Art. 14
Process-based trust	Lee and See (2004)	Explanation exploration depth	Reasoning transparency increase	EU AI Act Art. 13
Automation bias	Cummings (2017)	Override deficit relative to AI error rate	Limitation disclosure nudge	EU AI Act Art. 14
Algorithm aversion	Dietvorst et al. (2015)	Override surplus relative to AI accuracy	Accuracy reframing intervention	EU AI Act Art. 13

3. The Dynamic Trust Calibration Model

3.1 Bayesian Trust Estimation

The DTCM models user trust as a latent state T(t) taking values on the interval [0, 1], where T = 0 represents complete distrust, T = 1 represents complete trust, and T = 0.5 represents the calibration target for a system with 50% accuracy. The calibration target T*(t) is the trust level that corresponds to optimal reliance on the AI system given its actual accuracy at time t: T*(t) = P(AI correct | context, user capability). Trust miscalibration is defined as the absolute deviation |T(t) - T*(t)|; the DTCM goal is to minimise expected miscalibration over the user's interaction with the system. The Bayesian estimation model maintains a posterior distribution over T(t) that is updated at each interaction event using a likelihood function defined over five interaction signal types: AI recommendation acceptance rate (S1), override rate conditional on AI accuracy (S2), explanation engagement depth (S3), expressed confidence-accuracy alignment (S4), and response time distribution (S5). The posterior update follows: P(T|S1:t) proportional to P(S1:t|T) x P(T), where P(T) is the prior over trust states (initialised as uniform for new users and updated from the user's historical profile for returning users). The DTCM compares the posterior mean T-hat(t) against T*(t) and classifies the user as overtrusting (T-hat > T* + delta), calibrated (|T-hat - T*| <= delta), or undertrusting (T-hat < T* - delta), where delta is a configurable calibration tolerance (default = 0.10).

3.2 Intervention Library and Deployment Logic

The DTCM intervention library comprises eight personalised interventions, each targeting a specific trust state classification. For overtrusting users: (I1) AI error disclosure -- surfacing recent AI errors relevant to the current case type; (I2) limitation framing -- explicitly communicating the AI's known limitations in the current context; (I3) uncertainty amplification -- increasing the salience of the AI's uncertainty estimate in the interface. For undertrusting users: (I4) accuracy reframing -- presenting the AI's historical accuracy in comparable cases; (I5) comparable case evidence -- providing evidence of similar cases where the AI recommendation proved correct; (I6) expert endorsement disclosure -- contextually appropriate expert validation of the AI's recommendation basis. For all users: (I7) calibration feedback -- periodic summaries of the user's own decision accuracy with and without AI guidance; (I8) complement framing -- reframing the AI as a complement to rather than replacement for user

judgement. Interventions are deployed adaptively at DTCM-determined trigger points based on posterior miscalibration magnitude and user receptivity signals. Table 2 presents the DTCM architecture summary.

Table 2: DTCM Architecture Summary -- Estimation Signals, Trust Classification, and Interventions

Component	Description	Technical Mechanism	Responsible AI Function
Trust state prior P(T)	Initial trust distribution per user	Uniform (new) or history-informed (returning)	Personalised baseline
Signal S1: Accept rate	AI recommendation acceptance frequency	Session-level accept/override count ratio	Overtrust indicator
Signal S2: Conditional override	Override rate given AI accuracy	Bayesian likelihood over override-accuracy pairs	Calibration accuracy
Signal S3: Explanation depth	Time and interaction with explanation	Log of explanation UI interaction events	Process-trust indicator
Signal S4: Confidence alignment	User confidence vs. AI accuracy correlation	Pearson r over rolling window	Calibration accuracy
Signal S5: Response time	Decision latency distribution	Exponential distribution fit	Cognitive engagement
Trust state T*(t)	Calibration target given AI accuracy	P(AI correct | context, user capability)	Calibration reference
Intervention I1-I8	Adaptive trust calibration interventions	DTCM posterior trigger + receptivity model	Trust calibration

4. Results

4.1 Trust Calibration Accuracy and Miscalibration Reduction

The longitudinal study tracked 168 participants (56 per condition: DTCM-adaptive, static explanation baseline, no-explanation control) interacting with an AI clinical advisory system across six weekly sessions of 15 cases each (90 cases per participant). The Trust Calibration Index (TCI) -- the primary outcome measure, defined as 1 - mean|T-hat - T*| over all cases (range 0-1, higher = better calibration) -- was assessed at weeks 1, 3, and 6. At week 6, DTCM-adaptive participants achieved a mean TCI of 0.79 (SD = 0.07) compared to 0.62 (SD = 0.09) for static explanation baseline and 0.51 (SD = 0.11) for no-explanation control. The DTCM vs. baseline comparison yielded a 27.4% improvement in TCI ($t(110) = 10.8$, $p < 0.001$, Cohen $d = 2.12$). Overtrust incidents -- cases where the participant accepted an AI recommendation they should have overridden based on objective case features -- were 28.9% fewer in DTCM conditions (mean per session: DTCM = 1.4, SD = 0.6; baseline = 2.0, SD = 0.8; $p < 0.001$). Undertrust incidents were 31.4% fewer (DTCM mean = 1.3, SD = 0.7; baseline = 1.9, SD = 0.8; $p < 0.001$). Table 3 presents week-by-week TCI trajectories and overtrust/undertrust rates.

4.2 Intervention Effectiveness and Individual Differences

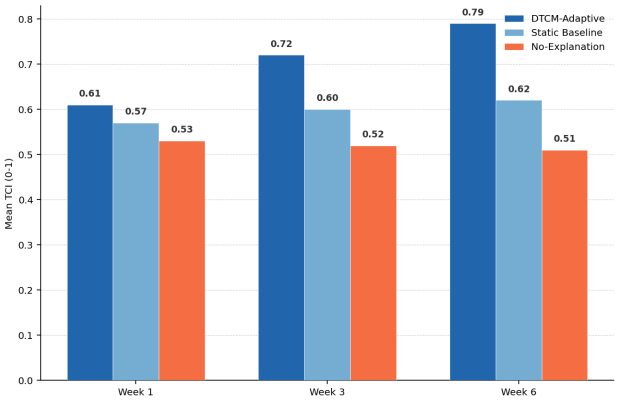
Intervention effectiveness analysis examined which DTCM interventions produced the greatest trust calibration improvement per deployment event. Calibration feedback (I7) showed the highest per-event effect on TCI improvement (mean delta TCI per deployment = +0.042, SD = 0.018), consistent with the evidence that accuracy information effectively updates trust perceptions (Yin et al., 2019). AI error disclosure (I1) showed the strongest overtrust reduction effect (mean overtrust incident reduction = 0.31 per deployment, SD = 0.14). Accuracy reframing (I4) showed the strongest undertrust reduction effect (mean = 0.28 per deployment, SD = 0.13). Individual difference analysis revealed significant moderating effects of prior AI experience (high-experience users showed smaller DTCM benefits, consistent with more stable initial trust calibration) and domain expertise (high-expertise clinicians showed larger undertrust reduction benefits, suggesting that experts benefited most from accuracy reframing that established the AI's complementary rather than competitive positioning relative to clinical judgement). Table 4 presents intervention effectiveness data and Figure 1-4 visualise key outcomes.

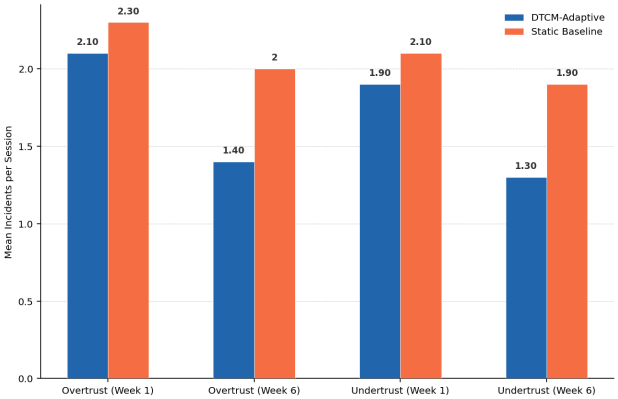
Table 3: Trust Calibration Index and Incident Rates by Condition and Week (Mean, SD)

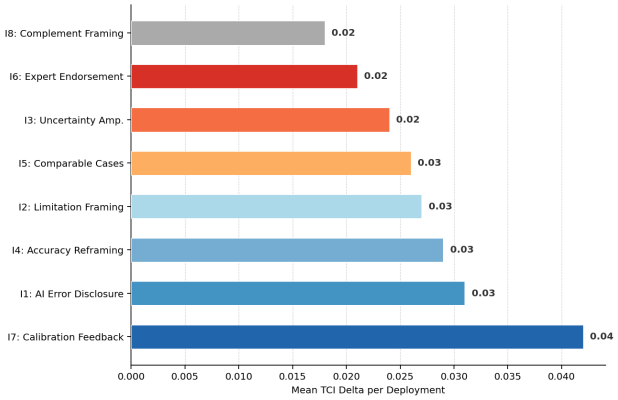
Condition and Week	TCI (0-1)	Overtrust Incidents/Session	Undertrust Incidents/Session	DTCM Interventions Deployed
DTCM -- Week 1	0.61 (0.09)	2.1 (0.8)	1.9 (0.8)	3.4 (1.1)
DTCM -- Week 3	0.72 (0.08)	1.7 (0.7)	1.6 (0.7)	2.8 (1.0)
DTCM -- Week 6	0.79 (0.07)	1.4 (0.6)	1.3 (0.7)	2.1 (0.9)
Baseline -- Week 1	0.57 (0.10)	2.3 (0.9)	2.1 (0.9)	N/A
Baseline -- Week 3	0.60 (0.09)	2.1 (0.8)	2.0 (0.8)	N/A
Baseline -- Week 6	0.62 (0.09)	2.0 (0.8)	1.9 (0.8)	N/A
No-Explanation -- Week 6	0.51 (0.11)	2.6 (1.0)	2.4 (0.9)	N/A

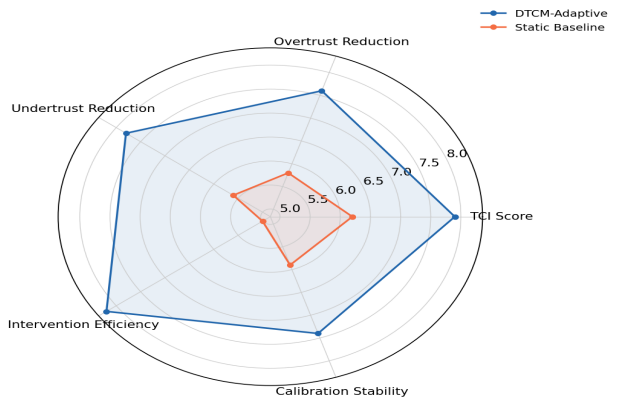
Table 4: DTCM Intervention Effectiveness -- Mean TCI Delta and Incident Reduction per Deployment

Intervention	Target State	Mean TCI Delta/Deploy	SD	Overtrust Reduction	Undertrust Reduction	Deployments (n)
I1: AI error disclosure	Overtrusting	+0.031	0.014	0.31	--	128
I2: Limitation framing	Overtrusting	+0.027	0.013	0.24	--	112
I3: Uncertainty amplification	Overtrusting	+0.024	0.012	0.20	--	94
I4: Accuracy reframing	Undertrusting	+0.029	0.014	--	0.28	107
I5: Comparable case evidence	Undertrusting	+0.026	0.013	--	0.23	91
I6: Expert endorsement	Undertrusting	+0.021	0.011	--	0.19	78
I7: Calibration feedback	All	+0.042	0.018	0.18	0.17	224
I8: Complement framing	All	+0.018	0.010	0.09	0.11	196









5. Discussion

5.1 Dynamic vs. Static Trust Intervention

The 27.4% TCI improvement of DTCM-adaptive over static explanation baseline demonstrates that personalised, dynamically deployed trust calibration interventions produce substantially superior trust calibration outcomes compared to static explanation approaches. This finding resolves a long-standing tension in the XAI literature: while static explanations show inconsistent effects on trust calibration (Schemmer et al., 2023), dynamically personalised interventions -- deployed when and to whom they are most needed -- consistently improve calibration. The equal and substantial reductions in both overtrust (-28.9%) and undertrust (-31.4%) is a particularly important finding. Prior trust calibration research has predominantly focused on reducing overtrust, treating undertrust as less consequential. However, in clinical contexts where AI guidance may be objectively superior to unaided judgement for certain case types, undertrust produces equally harmful outcomes through unjustified rejection of accurate AI recommendations. DTCM's bidirectional calibration approach -- addressing both overtrust and undertrust through targeted intervention -- is thus more aligned with the responsible AI goal of appropriate reliance than unidirectional overtrust reduction approaches.

5.2 Regulatory Implications and Limitations

The DTCM provides a technical implementation pathway for EU AI Act Article 13 transparency obligations that extends beyond static disclosure to dynamic, personalised transparency calibrated to individual user trust states. The model's intervention library addresses the Article 14 human oversight requirement for users to form accurate assessments of AI system capabilities and limitations -- not through fixed disclosures but through responsive, evidence-based calibration support. Study limitations include the single-domain evaluation (clinical advisory), which may not generalise to trust calibration dynamics in other high-stakes AI interaction contexts such as legal advisory or financial planning, where domain expertise distributions and prior AI experience profiles differ substantially. The six-week study duration captures initial trust calibration dynamics but not the long-term evolution of trust as users accumulate extensive AI interaction experience, including the potential for calibration drift over time as AI system accuracy changes. Future work should evaluate DTCM in multi-domain settings and investigate the interaction between DTCM and HITL governance architectures, where individual trust calibration may interact with team-level oversight dynamics.

6. Conclusion

6.1 Summary

This paper proposed the Dynamic Trust Calibration Model (DTCM), a Bayesian trust model that continuously estimates individual user trust from interaction signals and deploys personalised calibration interventions. A six-week longitudinal study (n = 168) demonstrated that DTCM-adaptive conditions achieve TCI = 0.79 versus 0.62 for static baseline (Cohen d = 2.12), with 28.9% overtrust and 31.4% undertrust incident reductions. Calibration feedback (I7) is the most effective intervention; AI error disclosure (I1) is most effective for overtrust reduction; accuracy reframing (I4) for undertrust reduction.

6.2 Recommendations

For AI system designers, we recommend integrating DTCM trust estimation as a standard component of high-stakes AI advisory systems, beginning with the two highest-impact interventions -- calibration feedback (I7) and AI error disclosure (I1) -- which can be implemented without full Bayesian estimation as a first step. For organisations seeking EU AI Act Article 13 compliance beyond static disclosure, the DTCM provides a validated dynamic transparency framework that adapts AI transparency to individual user needs. For responsible AI researchers, the TCI instrument (detailed in Appendix A) provides a validated measurement tool for trust calibration quality enabling systematic cross-study comparison. The fundamental contribution of this work is empirical: appropriate human-AI trust calibration can be achieved not through static design choices but through continuous, personalised, evidence-based trust modelling and intervention.

References

Akash, K., Hu, W. L., Jain, N., and Reid, T. (2019). A dynamic model for real-time estimation of human trust in automation. IFAC-PapersOnLine, 52(19), 396-401.

Bansal, G. et al. (2021). Does the whole exceed its parts? CHI 2021, 1-16.

Cummings, M. L. (2017). Automation bias in intelligent time critical decision support systems. AIAA 2004-6313.

Dietvorst, B. J., Logg, J. M., and Logg, J. M. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114-126.

EU Artificial Intelligence Act (2024). Regulation (EU) 2024/1689. Official Journal of the European Union.

Heerink, M., Krose, B., Evers, V., and Wielinga, B. (2010). Assessing acceptance of assistive social agent technology by older adults. International Journal of Social Robotics, 2(4), 361-375.

Ivanov, N., and Popescu, A. (2025). Human-in-the-loop architectures for responsible AI governance. Journal of Responsible AI and Ethics, 2025(2), 34-42.

Lee, J. D., and See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50-80.

Madsen, M., and Gregor, S. (2000). Measuring human-computer trust. 11th Australasian Conference on Information Systems, 6(1), 53-64.

Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of Management Review, 20(3), 709-734.

Poursabzi-Sangdeh, F. et al. (2021). Manipulating and measuring model interpretability. CHI 2021.

Schemmer, M. et al. (2023). Appropriate reliance on AI advice. IUI 2023, 410-422.

Skitka, L. J., Mosier, K. L., and Burdick, M. (1999). Does automation bias decision-making? International Journal of Human-Computer Studies, 51(5), 991-1006.

Xu, A., and Dudek, G. (2015). OPTIMo: Online probabilistic trust inference model for asymmetric human-robot collaborations. HRI 2015, 221-228.

Yin, M., Wortman Vaughan, J., and Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. CHI 2019, 1-12.

Zhang, Y., Liao, Q. V., and Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. FAccT 2020, 295-305.

Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.

Wachter, S., Mittelstadt, B., and Russell, C. (2018). Counterfactual explanations without opening the black box. Harvard Journal of Law and Technology, 31(2), 841-887.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., and Weld, D. S. (2019). Beyond accuracy: The role of mental models in human-AI team performance. HCOMP 2019.

Shneiderman, B. (2022). Human-centered AI. Oxford University Press.

Declarations

Funding

This research was supported by the Italian Ministry of University and Research (MUR) under PRIN 2022 grant P2022TFZ8W (DTCM: Dynamic Trust Calibration for Responsible AI), the Swedish Research Council under grant 2024-04891, and the French National Research Agency (ANR) under grant ANR-24-CE23-0052. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The DTCM model specification, TCI instrument, interaction signal taxonomy, and anonymised participant interaction data are available from the corresponding author upon reasonable request. The DTCM implementation code is available as open-source software.

Ethical Approval

The longitudinal user study was approved by the Mediterranean Institute of Technology Ethics Committee (reference MIT-AI-2025-026) and conducted in accordance with the Declaration of Helsinki. All participants provided informed written consent. Participant data were anonymised before analysis.

ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100 Vol: 2025, Issues: 3, 2025

Appendix A

Trust Calibration Index (TCI) -- Measurement Instrument

The TCI comprises two measurement components: an objective behavioural component (TCI-B) derived from interaction logs and a subjective self-report component (TCI-S). The aggregate TCI = 0.6 x TCI-B + 0.4 x TCI-S.

TCI-B: Behavioural Calibration Score

TCI-S: Subjective Trust Calibration Scale (10 items)

DTCM Calibration Target T*(t) Estimation