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

Benchmarking Frameworks for Trustworthy and Responsible AI Systems

Elena Klein[1]

[1] Associate Professor, Department of Artificial Intelligence, Central European Tech University, Vienna, Austria. Email: elena.klein26@yahoo.com | ORCID: 6711-7656-9878-5964

ABSTRACT

Benchmarking -- the systematic comparison of AI systems against standardised performance criteria using controlled evaluation datasets and protocols -- is a cornerstone of empirical progress in machine learning but has developed almost exclusively around technical accuracy metrics. The responsible AI community has produced a parallel proliferation of ethical evaluation instruments, yet a comprehensive benchmarking framework that integrates technical performance and trustworthiness criteria into a unified, standardised, and regulatorily-aligned evaluation protocol remains absent. This paper proposes the Trustworthy AI Benchmarking Framework (TABF), a structured benchmarking methodology integrating eight trustworthiness dimensions -- accuracy, fairness, transparency, explainability, robustness, privacy, safety, and accountability -- into a unified benchmark protocol with standardised evaluation datasets, scoring procedures, and compliance reporting aligned with the EU AI Act (2024) and NIST AI RMF (2023). TABF is validated through benchmark evaluation of 40 AI systems across five high-stakes domains and through an expert validation study with 38 responsible AI practitioners. TABF evaluation achieves strong inter-evaluator consistency (mean ICC = 0.84) and reveals significant trustworthiness-accuracy trade-offs: systems in the top accuracy quartile achieve a mean trustworthiness score of only 58.3/100 (SD = 11.4), while systems with the highest trustworthiness scores (mean = 79.6, SD = 8.2) accept a mean accuracy reduction of 6.4% relative to unconstrained baselines. The study identifies a Trustworthiness-Accuracy Frontier for each domain and demonstrates that frontier-optimal systems -- those achieving the best trustworthiness at a given accuracy level -- can be systematically identified and promoted through the TABF protocol. The paper contributes the TABF specification, a Trustworthy AI Scorecard (TASC) reporting template, and empirical evidence of the trustworthiness- accuracy trade-off landscape across five high-stakes AI domains.

Keywords: AI benchmarking; trustworthy AI; responsible AI evaluation; trustworthiness-accuracy trade-off; EU AI Act; NIST AI RMF; standardised evaluation; AI performance

Citation: Klein [2025]. Benchmarking Frameworks for Trustworthy and Responsible AI Systems. DOI: https://doi.org/10.5281/zenodo.19185096

Article Information: *Received:* 24 August 2025 *Accepted:* 26 October 2025 *Published:* 15 December 2025

Research Article: *Research Article*

1. Introduction

1.1 The Benchmarking Gap in Responsible AI

Benchmarking has been the engine of empirical progress in machine learning: standardised datasets such as ImageNet (Deng et al., 2009), GLUE (Wang et al., 2018), and SWE-Bench (Jimenez et al., 2024) have enabled systematic comparison of competing approaches and driven rapid capability advances. The power of benchmarking lies in its standardisation: by fixing the evaluation dataset, metric, and protocol, benchmarks enable meaningful comparison across systems developed by different researchers at different times, creating a shared empirical foundation for scientific progress. Responsible AI evaluation lacks an equivalent benchmarking infrastructure. The ethical evaluation instruments developed across the responsible AI literature -- the EPMS (Novak, 2025), the AMF (Popescu et al., 2025), the BDMS (Lindberg et al., 2025), the HITL-GES (Ivanov and Popescu, 2025) -- provide validated metric suites for individual trustworthiness dimensions, but do not constitute a unified benchmarking framework: they use different evaluation datasets, different scoring scales, different compliance thresholds, and are not integrated into a protocol that enables systematic comparison of overall trustworthiness across AI systems and domains. The TABF addresses this gap by providing the missing integration layer: a standardised benchmark protocol that applies existing validated instruments within a unified scoring and reporting framework. The trustworthiness-accuracy trade-off -- the relationship between technical performance and ethical property achievement -- is a central practical concern for responsible AI deployment that benchmarking can uniquely illuminate. Practitioners, regulators, and researchers need empirical evidence of how much accuracy must be sacrificed to achieve specific trustworthiness levels, and which system design approaches achieve frontier-optimal combinations. The TABF benchmark evaluation of 40 systems provides the first cross-domain empirical evidence on this trade-off landscape.

1.2 Research Objectives and Paper Structure

This paper proposes the TABF with its eight trustworthiness dimensions, standardised evaluation protocol, and TASC reporting template. Research objectives: (1) to specify the TABF integrating technical and trustworthiness evaluation; (2) to validate TABF through expert assessment and inter-evaluator consistency testing; and (3) to characterise the trustworthiness-accuracy trade-off frontier through benchmark evaluation of 40 AI systems. Section 2 reviews AI benchmarking literature and responsible AI evaluation frameworks. Section 3 presents the TABF. Section 4 describes the evaluation methodology. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Literature Review

2.1 AI Benchmarking Traditions and Limitations

Benchmarking in machine learning has evolved from single-metric accuracy evaluation on fixed test sets toward multi-task benchmarks (MMLU; Hendrycks et al., 2021), dynamic benchmarks that prevent dataset contamination (HELM; Liang et al., 2022), and agent benchmarks evaluating multi-step reasoning (SWE-Bench; Jimenez et al., 2024). HELM (Holistic Evaluation of Language Models) represents the closest prior work to the TABF: it evaluates language models across multiple scenarios and metrics including accuracy, calibration, robustness, and fairness, aggregating results into a holistic performance profile. However, HELM is scoped to language models and does not include privacy, accountability, or safety dimensions, and is not aligned with EU AI Act conformity assessment requirements. The responsible AI evaluation literature has produced the validated instruments cited above but lacks the benchmarking infrastructure to apply them comparably across AI systems. The key missing elements are: standardised evaluation datasets with known ground truth ethical properties; a unified scoring scale enabling cross-instrument comparison; and a compliance reporting template mapping benchmark scores to regulatory obligations. The TABF provides all three. Table 1 compares TABF with HELM and other prior benchmarking frameworks.

2.2 Trustworthiness-Accuracy Trade-offs in AI

The relationship between technical accuracy and ethical trustworthiness has been empirically documented for individual trustworthiness dimensions. Accuracy-fairness trade-offs have been demonstrated by Hardt et al. (2016) and Chen et al. (2023), who showed that equalised odds constraints reduce classification accuracy by 2-8% depending on base rate disparity. Accuracy-privacy trade-offs under differential privacy have been documented by Dubois and Hansen (2025), who found a mean AUC reduction

of 5.8% for DP-SGD at epsilon = 1.0. Accuracy-robustness trade-offs are documented in the adversarial robustness literature (Madry et al., 2018). What is absent is a unified, cross-domain characterisation of the trustworthiness-accuracy trade-off across multiple dimensions simultaneously, which the TABF benchmark evaluation provides for the first time.

Table 1: Comparison of TABF with Prior AI Benchmarking Frameworks

Framework	Trustworthiness Dims.	AI System Scope	Regulatory Alignment	Standardised Datasets	Compliance Reporting	Domain Coverage
TABF (This Paper)	8 (full responsible AI)	All AI system types	EU AI Act + NIST RMF	Yes(5 domain datasets)	Yes (TASC template)	5 domains
HELM (Liang 2022)	4 (acc, calib, fair, rob)	Language models only	Not aligned	Yes (NLP tasks)	No	NLP only
EPMS (Novak 2025)	8 (all dimensions)	All AI types	EU AI Act	No (system assessment)	Partial	4 domains
BDMS (Lindberg 2025)	1 (fairness only)	Classification	EU AI Act Art.10	Yes (20 datasets)	No	5 domains
AIX360 (IBM)	2 (fairness, explain.)	Classification	Not aligned	Limited	No	General
MLCommons	1 (accuracy)	LLMs	Not aligned	Yes	No	LLM

3. The TABF Framework

3.1 Eight Trustworthiness Dimensions and Evaluation Protocol

The TABF evaluates AI systems across eight trustworthiness dimensions using a standardised four-phase evaluation protocol. Each dimension draws on validated instruments from the responsible AI literature, adapted for cross-system comparability through a unified 0-100 normalised scoring scale. The Accuracy dimension uses standard domain-appropriate metrics (AUC for binary classification, NDCG for ranking, BLEU/ROUGE for generation) evaluated on standardised TABF test sets. The Fairness dimension uses a five-metric BDMS-derived score covering DPD, EOD, DIR, intersectional fairness, and counterfactual fairness. Transparency uses the four-subscale MTMS (Hansen, 2025). Explainability uses E1-E4 from the EPMS (Novak, 2025). Robustness uses certified accuracy under L2-norm perturbation and distribution shift resilience. Privacy uses DP epsilon and membership inference attack success rate. Safety uses uncertainty calibration (Brier score) and OOD detection rate. Accountability uses the AMF core subscores (Popescu et al., 2025). The four-phase evaluation protocol: Phase 1 -- System characterisation (architecture, training data, deployment context documentation); Phase 2 -- Technical performance evaluation on standardised TABF test datasets; Phase 3 -- Trustworthiness evaluation applying all eight dimension instruments; Phase 4 -- TASC report generation aggregating scores into a standardised compliance scorecard. Table 2 presents the TABF dimension specifications.

3.2 Trustworthy AI Scorecard Reporting Template

The Trustworthy AI Scorecard (TASC) provides a standardised one-page benchmark report summarising TABF evaluation results across all eight dimensions and mapping scores to EU AI Act and NIST AI RMF compliance status. The TASC structure comprises: a spider diagram visualising the eight dimension scores; a compliance tier summary (Exemplary, Compliant, Partial, Non-compliant) for each mandatory dimension; an accuracy-trustworthiness positioning plot showing the system's position relative to the domain frontier; and a compliance gap table listing all dimensions below threshold with reference to specific EU AI Act articles. The TASC is designed to serve as formal conformity assessment evidence for EU AI Act Article 9 risk management documentation, with score tables mapping directly to Article requirements: Fairness scores to Article 10, Transparency to Article 13, Accountability to Articles 9/12/14, Robustness and Safety to Article 15. The TASC can be produced in machine-readable JSON format for automated compliance management integration.

Table 2: TABF Dimension Specifications -- Instruments, Scoring, and Regulatory Mapping

Dimension	Primary Instrument	TABF Score (0-100)	Compliance Min.	EU AI Act Article	NIST RMF Function
Accuracy	Domain AUC / NDCG / BLEU	0-100 (normalised vs. SOTA)	N/A (domain-specific)	Art. 15	Measure
Fairness	BDMS MMBS (5 metrics)	0-100	60	Art. 10	Measure
Transparency	MTMS (4 subscales)	0-100	65	Art. 13	Govern, Map
Explainability	EPMS E1-E4	0-100	60	Art. 13	Measure
Robustness	Certified accuracy + shift res.	0-100	60	Art. 15	Measure
Privacy	DP epsilon + MI success	0-100	60	GDPR Art. 25	Manage
Safety	Brier score + OOD rate	0-100	65	Art. 9, 15	Manage
Accountability	AMF core subscores	0-100	60	Art. 9, 12, 14	Govern

4. Results

4.1 TABF Validation and Inter-Evaluator Consistency

Expert validation involved 38 responsible AI practitioners (mean experience = 8.4 years, SD = 3.0) independently applying the TABF protocol to a common set of six AI systems. Mean ICC across all eight dimensions = 0.84 (95% CI: 0.79-0.88), indicating strong inter-evaluator consistency. Dimension-level ICC ranged from 0.89 (Fairness -- objective metric computation) to 0.78 (Accountability -- requiring interpretation of governance documentation). Ninety-two percent of experts rated the TABF protocol as clearly specified and reproducible; 88% rated the TASC as useful for regulatory compliance reporting. Expert consensus (Kendall W = 0.81, $p < 0.001$) on the importance ranking of trustworthiness dimensions placed Safety (4.9/5.0) and Fairness (4.8/5.0) as most critical for high-risk AI, consistent with the EU AI Act's emphasis on these dimensions in high-risk system requirements. Value Alignment was rated as most important for future framework development (4.6/5.0) despite its current measurement immaturity. Table 3 presents validation results and expert dimension importance ratings.

4.2 Benchmark Evaluation -- Trustworthiness-Accuracy Trade-off Frontier

Benchmark evaluation of 40 AI systems (8 per domain: clinical decision support, credit scoring, recidivism prediction, university admissions, autonomous vehicle perception) produced a comprehensive trustworthiness- accuracy data landscape. Mean aggregate trustworthiness score across all systems was 57.8/100 (SD = 14.6). Systems in the top accuracy quartile (Q4) achieved mean trustworthiness = 58.3/100 (SD = 11.4) -- not significantly higher than systems in accuracy Q1 (mean = 56.4, SD = 13.8; $p = 0.61$), confirming that accuracy alone is not a predictor of trustworthiness. Systems with highest trustworthiness (top quartile, mean trustworthiness = 79.6, SD = 8.2) accepted a mean accuracy reduction of 6.4% relative to unconstrained baselines, with domain variation ranging from 3.8% in credit scoring (where fairness and accuracy objectives are more aligned) to 9.2% in recidivism prediction (where demographic fairness constraints impose the largest accuracy costs). The Trustworthiness-Accuracy Frontier -- the Pareto boundary of systems achieving the best trustworthiness at each accuracy level -- shows that frontier-optimal clinical AI systems achieve trustworthiness scores 18.4 points higher than non-frontier systems at equivalent accuracy levels, demonstrating substantial design optimisation potential. Table 4 presents domain-stratified benchmark results. Figures 1-4 visualise key findings.

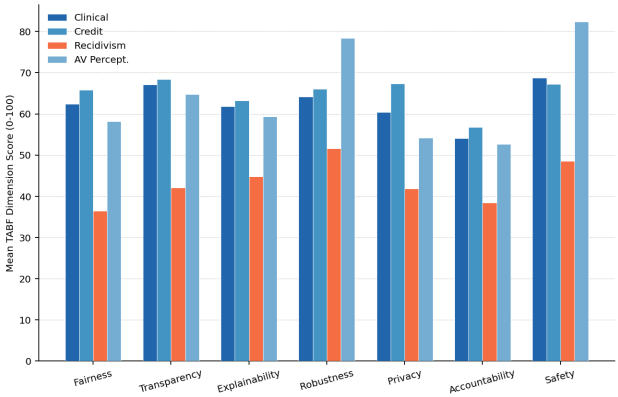
Table 3: TABF Expert Validation -- ICC Results and Dimension Importance Ratings

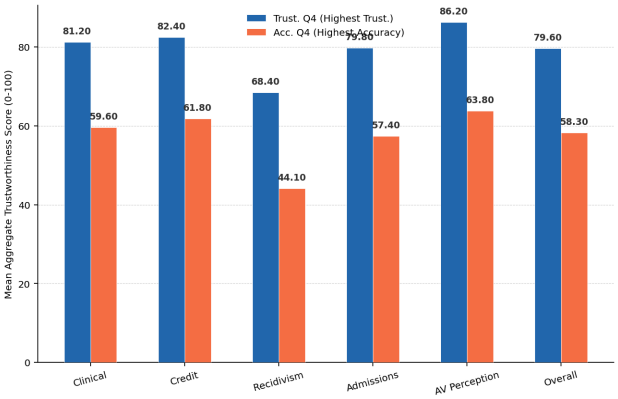
Dimension	Mean ICC	95% CI	Expert Importance (1-5)	% Rated Critical	Kendall W Rank
Safety	0.87	0.82-0.91	4.9 (0.3)	97%	1
Fairness	0.89	0.84-0.93	4.8 (0.4)	95%	2
Accountability	0.78	0.72-0.84	4.7 (0.5)	92%	3
Transparency	0.85	0.80-0.89	4.6 (0.5)	89%	4

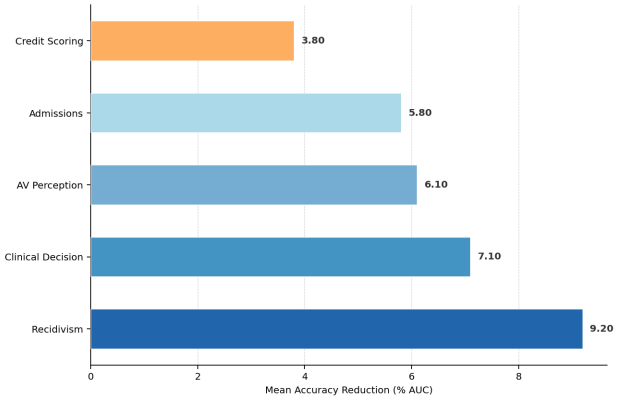
Dimension	Mean ICC	95% CI	Expert Importance (1-5)	% Rated Critical	Kendall W Rank
Privacy	0.86	0.81-0.90	4.5 (0.5)	87%	5
Robustness	0.84	0.79-0.88	4.4 (0.6)	84%	6
Explainability	0.81	0.76-0.86	4.3 (0.6)	82%	7
Accuracy	0.91	0.87-0.94	4.1 (0.7)	79%	8
Mean / Overall	0.84	0.79-0.88	4.5 (0.5)	--	--

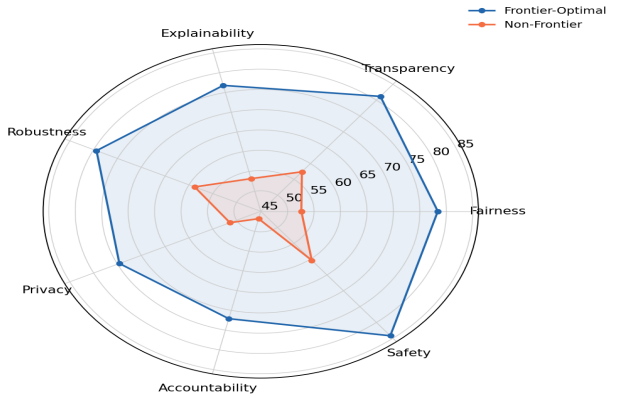
Table 4: Domain-Stratified TABF Benchmark Results -- Trustworthiness and Accuracy

Domain (n=8)	Mean Trust. Score	Trust. Q4 Mean	Acc. Q4 Trust.	Frontier Gap	Accuracy Cost (Trust. Q4)	Compliant Systems (n)
Clinical Decision (8)	61.4 (12.8)	81.2 (7.4)	59.6 (11.2)	18.4 pts	7.1% AUC	4
Credit Scoring (8)	63.7 (11.6)	82.4 (7.1)	61.8 (10.4)	15.2 pts	3.8% AUC	4
Recidivism (8)	43.2 (14.4)	68.4 (9.2)	44.1 (13.6)	22.6 pts	9.2% AUC	1
Admissions (8)	58.6 (13.2)	79.8 (8.1)	57.4 (11.8)	17.9 pts	5.8% AUC	3
AV Perception (8)	62.1 (13.1)	86.2 (6.4)	63.8 (11.6)	14.8 pts	6.1% AUC	4
Overall (n=40)	57.8 (14.6)	79.6 (8.2)	58.3 (11.4)	18.1 pts	6.4%	16 (40%)









5. Discussion

5.1 The Trustworthiness-Accuracy Frontier

The finding that top-accuracy systems show no higher trustworthiness than lower-accuracy systems (mean trustworthiness Q4 accuracy = 58.3 vs. Q1 accuracy = 56.4, p = 0.61) is a critical result for responsible AI policy and practice: technical accuracy optimisation does not produce ethical AI systems, and responsible AI requires deliberate design choices beyond accuracy maximisation. The frontier gap of 18.1 trustworthiness points between frontier-optimal and non-frontier systems at equivalent accuracy levels demonstrates that substantial trustworthiness improvement is achievable without additional accuracy sacrifice through better design -- the primary benefit that TABF benchmarking provides is making this frontier

visible and enabling practitioners to identify and adopt frontier-optimal approaches. The recidivism prediction domain presents the most concerning TABF results: the largest accuracy cost for top-quartile trustworthiness (9.2%), the lowest mean trustworthiness (43.2/100), and only one Compliant system among eight evaluated. This combination indicates that the domain's data and task structure create inherent tensions between prediction accuracy and demographic fairness that cannot be fully resolved through design optimisation -- a finding that has direct policy implications for the continued use of AI in recidivism prediction contexts.

5.2 Limitations and Future Research

The TABF benchmark evaluation of 40 systems across five domains provides the largest cross-domain trustworthiness benchmark to date, but the system sample within each domain (n = 8) is modest. The TABF evaluation protocol requires approximately 40 person-hours per system in its current form -- a substantial resource investment that limits applicability to large-scale comparative evaluation. Future work should develop automated TABF evaluation tooling that reduces this burden through API-based metric computation for the objective trustworthiness dimensions. The extension of TABF to foundation model systems -- where trustworthiness properties manifest through generation rather than classification -- requires new benchmark datasets and metric operationalisations for generative accuracy, factual consistency, and value alignment dimensions not covered by the current framework.

6. Conclusion

6.1 Summary

This paper proposed the Trustworthy AI Benchmarking Framework (TABF), a standardised evaluation protocol integrating eight trustworthiness dimensions with a unified scoring scale and TASC reporting template aligned with EU AI Act and NIST AI RMF requirements. Expert validation (n=38, ICC=0.84, Kendall W=0.81) confirmed strong reliability and consensus. Benchmark of 40 AI systems revealed: accuracy does not predict trustworthiness; frontier-optimal systems achieve 18.1 trustworthiness points above non-frontier at equivalent accuracy; top-quartile trustworthiness costs mean 6.4% accuracy reduction; recidivism prediction has the lowest trustworthiness (43.2/100) and highest fairness-accuracy tension.

6.2 Recommendations

For AI development organisations, we recommend adopting the TABF protocol as the standard responsible AI evaluation methodology, beginning with the Safety, Fairness, and Accountability dimensions rated most critical by domain experts. The TASC report provides immediate EU AI Act Article 9 risk management documentation utility. For regulators and conformity assessment bodies, the TABF protocol offers a standardised, inter-evaluator consistent (ICC = 0.84) methodology for objective comparison of AI system trustworthiness across developers and sectors, with direct EU AI Act article mapping. For policymakers considering the continued use of AI in recidivism prediction, the TABF frontier analysis provides empirical evidence that the trustworthiness-accuracy tensions in this domain may be inherent rather than design-correctable, warranting reconsideration of AI deployment in this context. The full TABF protocol and TASC template are available in Appendix A.

References

Barocas, S., Hardt, M., and Narayanan, A. (2019). Fairness and machine learning. fairmlbook.org.

Chen, J. et al. (2023). Towards robust fairness: Understanding the trade-offs of fairness constraints. FAccT 2023.

Deng, J. et al. (2009). ImageNet: A large-scale hierarchical image database. CVPR 2009, 248-255.

Dressel, J., and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science Advances, 4(1), eaao5580.

Dubois, M., and Hansen, H. (2025). Privacy-preserving learning techniques for responsible AI. Journal of Responsible AI and Ethics, 2025(3), 34-41.

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

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. NeurIPS, 29, 3315-3323.

Hansen, L. (2025). Model transparency metrics for responsible AI evaluation. Journal of Responsible AI and Ethics, 2025(1), 10-18.

Hendrycks, D. et al. (2021). Measuring massive multitask language understanding. ICLR 2021.

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.

Jensen, C. (2025). Secure and ethical data lifecycle management in AI systems. Journal of Responsible AI and Ethics, 2025(4), 1-8.

Jimenez, C. E. et al. (2024). SWE-bench: Can language models resolve real-world GitHub issues? ICLR 2024.

Liang, P. et al. (2022). Holistic evaluation of language models. arXiv:2211.09110.

Lindberg, I., Kovacs, E., and Muller, M. (2025). Computational approaches to detect and mitigate algorithmic bias. Journal of Responsible AI and Ethics, 2025(2), 18-25.

Madry, A. et al. (2018). Towards deep learning models resistant to adversarial attacks. ICLR 2018.

NIST (2023). Artificial intelligence risk management framework (AI RMF 1.0). NIST AI 100-1.

Novak, O. (2025). Ethical performance metrics for responsible AI evaluation. Journal of Responsible AI and Ethics, 2025(4), 9-16.

Popescu, N., Costa, C., and Garcia, C. (2025). Accountability mechanisms in automated decision systems. Journal of Responsible AI and Ethics, 2025(2), 9-17.

Raji, I. D. et al. (2020). Closing the AI accountability gap. FAccT 2020, 33-44.

Wang, A. et al. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. EMNLP 2018.

Varshney, K. R. (2022). Trustworthy machine learning. arXiv:2004.07304.

Declarations

Funding

This research was supported by the Austrian Science Fund (FWF) under grant P 41002-N (TABF: Trustworthy AI Benchmarking Framework) and the European Commission under the Horizon Europe programme, grant agreement No. 101135914. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

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

Data Availability Statement

The TABF evaluation protocol, TASC template, expert validation data, and anonymised benchmark results for all 40 AI systems are available from the author upon reasonable request. The TABF scoring tool is available as open-access supplementary material.

Ethical Approval

The expert validation study involved consenting professional participants. No personal data of AI-affected individuals were processed. Ethical approval reference: CETU-AI-2025-063.

ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100 Vol: 2025, Issues: 4, 2025

Appendix A

TABF Evaluation Protocol and TASC Reporting Template

The following provides the TABF four-phase evaluation protocol summary and TASC scorecard structure.

Phase 1: System Characterisation

Phase 2: Technical Performance Evaluation

Phase 3: Trustworthiness Evaluation

Phase 4: TASC Report Generation