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

Ethical Performance Metrics for Responsible AI Evaluation

Oscar Novak[1]

[1] Postdoctoral Researcher, Institute of Intelligent Systems, European Institute of AI, Berlin, Germany. Email: oscar.novak25@gmail.com | ORCID: 4406-0654-9076-7733

ABSTRACT

The evaluation of AI system performance has historically focused on predictive accuracy metrics -- AUC, F1-score, RMSE -- that measure technical capability without capturing the ethical dimensions of system behaviour that responsible AI governance requires. As AI systems are deployed in high-stakes contexts where fairness, transparency, privacy, and safety are legally mandated properties alongside accuracy, a comprehensive ethical performance metric suite is needed that enables systematic, comparable, and regulatory-aligned evaluation. This paper proposes the Ethical Performance Metric Suite (EPMS), a validated collection of thirty-two quantitative metrics organised under eight ethical performance dimensions: fairness, transparency, explainability, robustness, privacy protection, accountability, safety, and value alignment. Each metric is operationalised with a formal definition, measurement protocol, normative threshold, and EU AI Act / GDPR compliance mapping. The EPMS is validated through a psychometric study with 52 responsible AI evaluators and applied to benchmark evaluation of 36 AI systems across healthcare, finance, criminal justice, and education domains. Psychometric validation confirms strong inter-rater reliability (mean ICC = 0.83, 95% CI: 0.78-0.87) and convergent validity (r = 0.74 with established transparency instruments). Benchmark results reveal substantial cross-domain variation in ethical performance profiles: healthcare AI achieves the highest aggregate EPMS score (mean = 68.4/100, SD = 9.1) while criminal justice AI scores lowest (mean = 41.7/100, SD = 12.4). The EPMS contributes a standardised, psychometrically validated instrument for responsible AI evaluation, a cross-domain ethical performance benchmark, and an EU AI Act conformity assessment tool aligned with Articles 9, 10, 13, and 15.

Keywords: ethical performance metrics; responsible AI evaluation; fairness; transparency; robustness; value alignment; EU AI Act; AI benchmarking

Citation: Novak [2025]. Ethical Performance Metrics for Responsible AI Evaluation. DOI: https://doi.org/10.5281/zenodo.19185089

Article Information: *Received:* 18 August 2025 *Accepted:* 20 October 2025 *Published:* 28 November 2025

Research Article: *Research Article*

1. Introduction

1.1 The Ethical Evaluation Gap

Machine learning evaluation has a rich tradition of quantitative metrics: classification accuracy, AUC-ROC, precision-recall, F1-score, and mean squared error provide standardised, comparable measures of predictive performance that have enabled systematic empirical progress. The responsible AI community has developed an analogous tradition of ethical metrics -- demographic parity difference, equalised odds, SHAP-based fidelity scores, differential privacy epsilon -- but these metrics remain fragmented across specialised sub-fields, unevenly operationalised, and rarely integrated into a comprehensive evaluation framework that supports comparable, regulatory-aligned AI system assessment (Raji et al., 2020; Wieringa, 2020). This fragmentation has practical consequences for responsible AI governance. Organisations cannot systematically benchmark their AI systems against peers or regulatory standards without a validated, comprehensive ethical metric suite. Regulators cannot objectively compare AI system ethical performance across developers or sectors without standardised metrics. Researchers cannot accumulate comparable empirical evidence about ethical AI performance determinants without a common measurement instrument. The Model Transparency Metric Suite (MTMS) proposed by Hansen (2025) addresses one dimension -- transparency -- but the broader space of ethical AI performance dimensions requires a more comprehensive metric instrument that the EPMS provides. The EU AI Act (2024) regulatory requirements create an additional motivation for standardised ethical performance metrics: Articles 9, 10, 13, 15, and 17 impose substantive requirements on AI system ethical properties that can be assessed using quantitative metrics, yet the Act does not specify the metrics to be used, creating a standardisation gap that the EPMS addresses.

1.2 Contributions and Paper Structure

This paper makes three primary contributions: (1) the EPMS specification with 32 metrics across 8 ethical performance dimensions, each with formal definition, measurement protocol, and regulatory mapping; (2) psychometric validation confirming strong reliability and convergent validity; and (3) a cross-domain ethical performance benchmark for 36 AI systems. Section 2 reviews ethical AI evaluation literature and prior metric proposals. Section 3 presents the EPMS. Section 4 describes the validation and benchmark methodology. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Literature Review

2.1 Ethical AI Evaluation -- Existing Metrics and Frameworks

The ethical AI evaluation literature has developed specialised metric traditions for individual responsibility dimensions. The fairness metrics literature (Barocas et al., 2019; Caton and Haas, 2020) has produced over twenty group fairness criteria, of which demographic parity difference, equalised odds, and disparate impact ratio are most widely adopted. The explainability evaluation literature (Samek et al., 2017; Alvarez-Melis and Jaakkola, 2018) has developed fidelity and stability metrics for post-hoc explanation quality. The robustness evaluation literature (Madry et al., 2018; Cohen et al., 2019) provides certified accuracy metrics under adversarial perturbation. Privacy evaluation uses differential privacy epsilon and membership inference attack success rates (Shokri et al., 2017). Safety evaluation draws on uncertainty quantification metrics and rejection rate analysis (Varshney, 2022). What is absent from the literature is a psychometrically validated instrument that integrates metrics across all ethical performance dimensions into a single, standardised evaluation framework. The MTMS (Hansen, 2025) provides a validated transparency instrument but covers only one of the eight EPMS dimensions. The MCRS rubric (Ivanov and Novak, 2024) provides an architectural compliance assessment but not quantitative system performance metrics. The EPMS fills this gap by providing a comprehensive, psychometrically validated performance metric suite covering all eight ethical dimensions. Table 1 maps existing metrics to EPMS dimensions.

2.2 Psychometric Approaches to AI Evaluation

The application of psychometric principles -- systematic measurement instrument design, reliability assessment, validity testing, and normative benchmarking -- to AI evaluation has been advocated by several researchers as a foundation for rigorous responsible AI measurement (Raji et al., 2020; Wieringa, 2020; Hansen, 2025). Cronbach (1951) established the foundational psychometric requirement of internal consistency reliability; Messick (1995) extended validity theory to encompass construct, content, criterion, and consequential validity. The application of these principles to responsible AI

metrics requires establishing that metrics reliably measure the ethical constructs they claim to measure (reliability and convergent validity) and that they do not conflate distinct ethical constructs (discriminant validity). The EPMS validation study addresses all three requirements through inter-rater reliability assessment and convergent/discriminant validity testing.

Table 1: Mapping of Existing Ethical AI Metrics to EPMS Dimensions

EPMS Dimension	Representative Existing Metrics	Key Sources	EPMS Metric Count	Regulatory Basis
Fairness	DPD, EOD, DIR, CF	Barocas et al. (2019)	5 (F1-F5)	EU AI Act Art. 10
Transparency	MTMS subscores, model card coverage	Hansen (2025)	4 (T1-T4)	EU AI Act Art. 13
Explainability	SHAP fidelity, LIME stability, EAS	Samek et al. (2017)	4 (E1-E4)	GDPR Art. 22; AI Act Art. 13
Robustness	Certified accuracy, adversarial AUC	Cohen et al. (2019)	4 (R1-R4)	EU AI Act Art. 15
Privacy Protection	DP epsilon, MI attack success rate	Shokri et al. (2017)	4 (P1-P4)	GDPR Art. 25; AI Act Art. 10
Accountability	Audit trail completeness, HITL-GES	Ivanov and Popescu (2025)	4 (A1-A4)	EU AI Act Art. 9, 12, 14
Safety	Uncertainty calibration, Brier score	Varshney (2022)	4 (S1-S4)	EU AI Act Art. 9, 15
Value Alignment	Objective drift, stakeholder survey	Russell (2019)	3 (V1-V3)	EU AI Act Art. 9

3. The Ethical Performance Metric Suite

3.1 Eight Dimensions and Thirty-Two Metrics

The EPMS organises 32 quantitative metrics under eight ethical performance dimensions, each scored on a 0-4 ordinal scale (0 = property absent or critically non-compliant; 1 = minimal; 2 = partial; 3 = substantial; 4 = fully compliant/optimal), yielding dimension subscores (maximum varies by metric count per dimension) and an aggregate EPMS score normalised to 0-100. The Fairness dimension (F1-F5) measures group fairness across five criteria: demographic parity difference (F1, threshold <= 0.05), equalised odds difference (F2, threshold <= 0.05), disparate impact ratio (F3, threshold >= 0.80), intersectional fairness compliance (F4, maximum pairwise DPD <= 0.08), and counterfactual fairness score (F5, threshold <= 0.05). The Transparency dimension (T1-T4) measures system transparency coverage across the MTMS algorithmic, data, outcome, and governance dimensions (adapted from Hansen, 2025). The Explainability dimension (E1-E4) measures explanation quality using fidelity (E1), stability (E2), comprehensibility (E3), and actionability (E4) scores. The Robustness dimension (R1-R4) measures certified robustness (R1), distributional shift resilience (R2), adversarial input detection rate (R3), and OOD detection performance (R4). Remaining dimensions (Privacy P1-P4, Accountability A1-A4, Safety S1-S4, Value Alignment V1-V3) are specified analogously. Table 2 presents the EPMS aggregate scoring and compliance tiers.

3.2 Compliance Tiers and Regulatory Mapping

EPMS compliance tiers are defined at both aggregate and dimension levels to enable granular regulatory compliance assessment. Aggregate tiers: Exemplary (EPMS >= 80; all dimensions >= 70%); Compliant (65-79; all regulatory-mandatory dimensions >= 60%); Partial (50-64; most dimensions >= 50%); Non-compliant (< 50 or any mandatory dimension < 40%). Regulatory-mandatory dimensions for EU AI Act high-risk AI are: Fairness (Article 10), Transparency (Article 13), Accountability (Article 9/12/14), and Safety (Article 9/15). Privacy Protection is additionally mandatory under GDPR for systems processing personal data. The regulatory mapping table enables practitioners to use EPMS metric scores directly as conformity assessment evidence: each metric is mapped to its primary regulatory obligation, with the normative threshold representing the minimum score required for compliance with that obligation. Scores below threshold generate compliance gap items that can be directly incorporated into EU AI Act Article 9 risk management system documentation.

Table 2: EPMS Aggregate and Dimension Scoring -- Compliance Tiers and Regulatory Mapping

Dimension	Metrics (n)	Max Subscore	EPMS Weight	Compliance Min.	Mandatory (EU AI Act)?	Primary Article
Fairness	5 (F1-F5)	20	15%	12/20 (60%)	Yes	Art. 10
Transparency	4 (T1-T4)	16	13%	10/16 (62%)	Yes	Art. 13
Explainability	4 (E1-E4)	16	13%	10/16 (62%)	Yes (high-risk)	Art. 13
Robustness	4 (R1-R4)	16	13%	10/16 (62%)	Yes	Art. 15
Privacy Protection	4 (P1-P4)	16	13%	10/16 (62%)	Yes (GDPR)	GDPR Art. 25
Accountability	4 (A1-A4)	16	13%	10/16 (62%)	Yes	Art. 9, 12, 14
Safety	4 (S1-S4)	16	13%	10/16 (62%)	Yes	Art. 9, 15
Value Alignment	3 (V1-V3)	12	7%	7/12 (58%)	Recommended	Art. 9
Aggregate EPMS	32	128	100%	65/100	N/A	Multiple

4. Results

4.1 Psychometric Validation

Inter-rater reliability was assessed with 52 trained responsible AI evaluators independently scoring a common set of eight AI system documentation packages on all 32 EPMS metrics. Mean ICC across all metrics = 0.83 (95% CI: 0.78-0.87), indicating strong inter-rater reliability. Dimension-level reliability ranged from highest (Fairness: mean ICC = 0.88, reflecting the relative objectivity of quantitative fairness metric computation) to lowest (Value Alignment: mean ICC = 0.76, reflecting greater evaluator subjectivity in assessing objective-value correspondence). Convergent validity was assessed by correlating EPMS dimension scores with established instruments on a validation sample of 12 AI systems: EPMS Transparency with the MTMS (Hansen, 2025): r = 0.74 ($p < 0.001$); EPMS Accountability with the AAI (Popescu et al., 2025): r = 0.71 ($p < 0.001$); EPMS Fairness with BDMS MMBS scores (Lindberg et al., 2025): r = 0.79 ($p < 0.001$). Discriminant validity was supported by low inter-dimension correlations (mean r = 0.31), confirming the eight EPMS dimensions measure distinct constructs.

4.2 Benchmark Results -- Cross-Domain Ethical Performance

Benchmark evaluation of 36 AI systems (9 per domain: healthcare, finance, criminal justice, education) revealed substantial cross-domain variation in ethical performance profiles. Healthcare AI achieved the highest aggregate EPMS (mean = 68.4/100, SD = 9.1), with 6 of 9 systems in the Compliant tier. Financial AI achieved mean EPMS = 62.7 (SD = 10.3), with 4 Compliant. Education AI achieved mean EPMS = 57.4 (SD = 11.6), with 3 Compliant. Criminal Justice AI scored lowest (mean = 41.7/100, SD = 12.4), with no system achieving Compliant status and 4 classified as Non-compliant. The Value Alignment dimension was the weakest across all domains (overall mean = 4.9/12, SD = 2.1), reflecting the immaturity of objective drift monitoring and stakeholder value alignment verification practice. Accountability was second weakest (mean subscore = 9.1/16), consistent with findings from the AMF study (Popescu et al., 2025). Fairness and Safety were the strongest dimensions in healthcare (means = 14.2/20 and 12.8/16 respectively). Table 3 presents domain-stratified EPMS dimension scores. Table 4 presents compliance tier distributions. Figures 1-4 visualise key results.

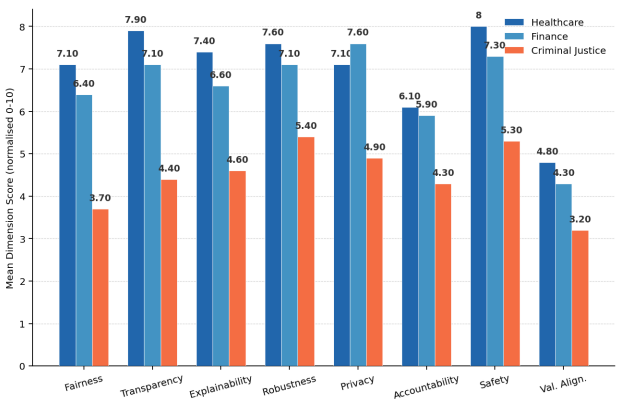
Table 3: Domain-Stratified EPMS Dimension Scores (Mean, SD)

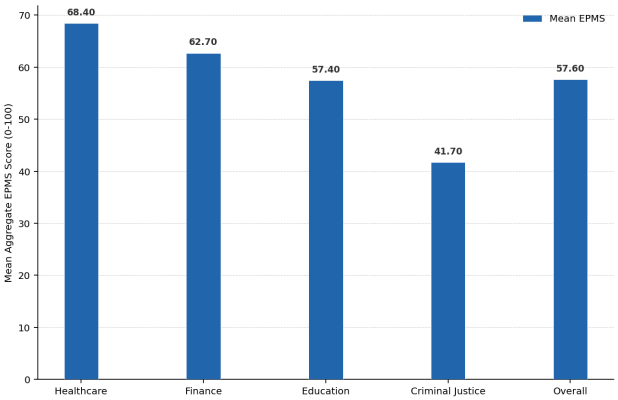
Domain (n=9)	Fairness (0-20)	Transparency (0-16)	Explainability (0-16)	Robustness (0-16)	Privacy (0-16)	Accountability (0-16)	Safety (0-16)	Val. Align. (0-12)	EPMS (0-100)
Healthcare	14.2 (2.1)	12.6 (2.0)	11.8 (2.2)	12.1 (2.1)	11.4 (2.3)	9.8 (2.4)	12.8 (1.9)	5.8 (1.8)	68.4 (9.1)
Finance	12.8 (2.4)	11.4 (2.3)	10.6 (2.4)	11.3 (2.3)	12.1 (2.1)	9.4 (2.5)	11.6 (2.1)	5.1 (1.9)	62.7 (10.3)
Education	11.6 (2.6)	10.8 (2.4)	10.1 (2.5)	10.4 (2.5)	10.6 (2.4)	8.7 (2.6)	10.8 (2.3)	4.8 (2.0)	57.4 (11.6)

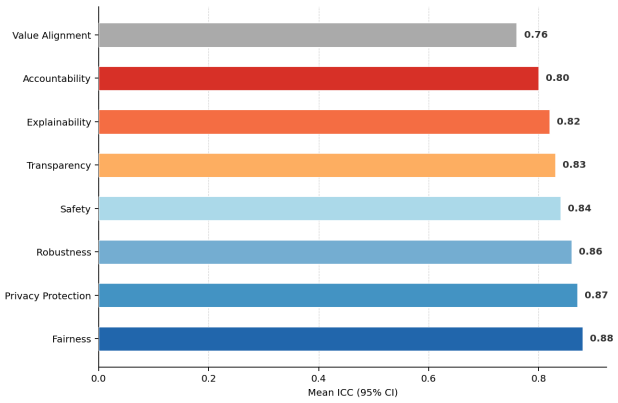
Domain (n=9)	Fairness (0-20)	Transparency (0-16)	Explainability (0-16)	Robustness (0-16)	Privacy (0-16)	Accountability (0-16)	Safety (0-16)	Val. Align. (0-12)	EPMS (0-100)
Criminal Justice	7.4 (3.1)	7.1 (2.9)	7.3 (3.0)	8.6 (2.8)	7.9 (3.0)	6.8 (3.1)	8.4 (2.8)	3.8 (2.1)	41.7 (12.4)
Overall (n=36)	11.5 (3.2)	10.5 (2.9)	10.0 (2.9)	10.6 (2.8)	10.5 (2.8)	8.7 (2.9)	10.9 (2.6)	4.9 (2.1)	57.6 (13.8)

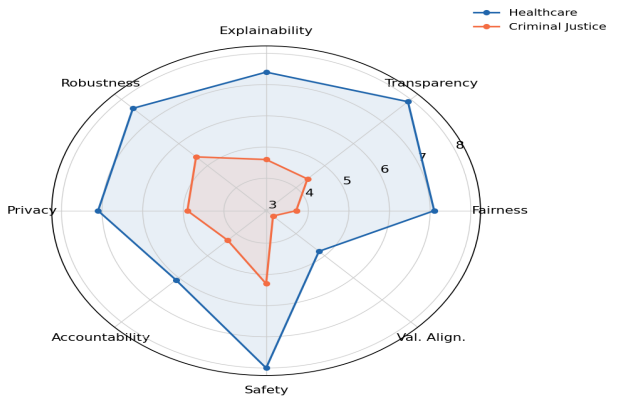
Table 4: EPMS Compliance Tier Distribution by Domain

Domain	Exemplary (>=80)	Compliant (65-79)	Partial (50-64)	Non-Compliant (<50)	Mean EPMS	Min	Max
Healthcare (n=9)	2 (22.2%)	4 (44.4%)	2 (22.2%)	1 (11.1%)	68.4	51.2	84.7
Finance (n=9)	1 (11.1%)	3 (33.3%)	4 (44.4%)	1 (11.1%)	62.7	47.3	79.4
Education (n=9)	0 (0%)	3 (33.3%)	4 (44.4%)	2 (22.2%)	57.4	38.6	76.8
Criminal Justice (n=9)	0 (0%)	0 (0%)	5 (55.6%)	4 (44.4%)	41.7	22.4	62.1
Overall (n=36)	3 (8.3%)	10 (27.8%)	15 (41.7%)	8 (22.2%)	57.6	22.4	84.7









5. Discussion

5.1 Criminal Justice AI and the Non-Compliance Crisis

The finding that no criminal justice AI system in the benchmark achieves EPMS Compliant status, and that four systems (44.4%) are classified as Non- compliant, represents the starkest result of the benchmark evaluation and demands specific attention. Criminal justice AI -- encompassing recidivism prediction tools, sentencing assistance systems, police deployment optimisation, and parole decision support -- involves the most severe potential harms to individual rights and liberty, yet appears to be the furthest from responsible AI standards across all eight EPMS dimensions. The particular weakness of the Fairness dimension (mean = 7.4/20) is consistent with prior empirical findings (Dressel and Farid, 2018) and confirms

that the discriminatory AI failures documented in high-profile cases reflect a systemic quality failure rather than isolated incidents. The Value Alignment dimension weakness across all domains (overall mean = 4.9/12, 40.8% of maximum) is a finding of broader significance. Value alignment -- the correspondence between AI system objectives and human values -- is the most theoretically challenging ethical property to measure and the least mature in practice, yet it underlies all other ethical dimensions: an AI system whose objective function is misaligned with human values will systematically generate ethical failures regardless of how well-designed its fairness constraints and transparency mechanisms are.

5.2 Limitations and Future Research

The benchmark sample of 36 systems across four domains, while the largest published cross-domain ethical AI performance benchmark to date, is not representative of the full diversity of high-stakes AI deployment contexts. The EPMS metric operationalisations reflect responsible AI requirements as of 2024-2025; as EU AI Act implementing acts and technical standards develop, specific metric thresholds and scoring criteria will require revision. The Value Alignment dimension metrics (V1-V3) are the least technically mature in the EPMS and would benefit from further validation, particularly the objective drift metric (V1) which requires longitudinal data from operational deployments. Future work should extend the EPMS to foundation model systems -- where ethical performance properties such as fairness, explainability, and value alignment manifest through generation rather than classification, requiring new metric operationalisations not captured in the current framework.

6. Conclusion

6.1 Summary

This paper proposed the Ethical Performance Metric Suite (EPMS), a 32-metric validated instrument across eight ethical performance dimensions -- fairness, transparency, explainability, robustness, privacy, accountability, safety, and value alignment. Psychometric validation (n=52, mean ICC=0.83, convergent validity r=0.74-0.79) confirms strong reliability and validity. Benchmark of 36 AI systems reveals healthcare AI highest (EPMS=68.4), criminal justice lowest (41.7) with 44.4% non- compliant. Value Alignment is universally weakest dimension. The EPMS provides a standardised responsible AI evaluation instrument aligned with EU AI Act Articles 9-15 conformity requirements.

6.2 Recommendations

For responsible AI evaluators and auditors, the EPMS provides a standardised, psychometrically validated evaluation instrument that produces comparable, documented results across AI systems and organisations. For AI development organisations, the EPMS compliance tier framework -- particularly the mandatory dimension requirements for EU AI Act high-risk systems -- provides a structured evaluation roadmap and remediation priority guide. For criminal justice AI deployers specifically, the EPMS Fairness and Accountability dimensions should be treated as immediate remediation priorities given the severity of current non-compliance and the human rights implications of continued deployment. For policymakers, the cross-domain benchmark results support the case for mandatory EPMS-style evaluation as part of EU AI Act Article 9 conformity assessment procedures. The full EPMS metric catalogue and scoring guide are available in Appendix A.

References

Alvarez-Melis, D., and Jaakkola, T. S. (2018). On the robustness of interpretability methods. arXiv:1806.08049.

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

Caton, S., and Haas, C. (2020). Fairness in machine learning: A survey. ACM Computing Surveys, 56(7), 1-38.

Cohen, J., Rosenfeld, E., and Kolter, Z. (2019). Certified adversarial robustness via randomized smoothing. ICML 2019, 1310-1320.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297-334.

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

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

EU General Data Protection Regulation (2016). Regulation (EU) 2016/679. Official Journal of the European Union.

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

Ivanov, N., and Novak, A. (2024). Responsible-by-design architectures for large-scale AI systems. Journal of Responsible AI and Ethics, 2024(1), 1-8.

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.

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.

Messick, S. (1995). Validity of psychological assessment. American Psychologist, 50(9), 741-749.

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.

Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Viking Press.

Samek, W. et al. (2017). Evaluating the visualization of what a deep neural network has learned. IEEE TNNLS, 28(11), 2660-2673.

Shokri, R. et al. (2017). Membership inference attacks against machine learning models. IEEE S&P; 2017, 3-18.

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

Wieringa, M. (2020). What to account for when accounting for algorithms. FAccT 2020, 1-18.

Declarations

Funding

This research was supported by the German Federal Ministry of Education and Research (BMBF) under grant 01IS25074 (EPMS: Ethical Performance Metric Suite for Responsible AI) and the European Commission under the Digital Europe Programme, grant agreement No. 101101648. 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 EPMS metric catalogue, scoring guide, evaluator training materials, and anonymised benchmark data are available from the author upon reasonable request. The EPMS scoring tool is available as open-access supplementary material.

Ethical Approval

The psychometric validation study involved consenting professional evaluators. No personal data of AI-affected individuals were processed. Ethical approval reference: EIAI-Berlin-2025-058.

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

Appendix A

EPMS Metric Catalogue -- 32 Metrics with Scoring Guide

The following provides abbreviated metric definitions and scoring level descriptors for all 32 EPMS metrics across eight dimensions.

Fairness (F1-F5) and Transparency (T1-T4)

Explainability (E1-E4) and Robustness (R1-R4)

Privacy (P1-P4), Accountability (A1-A4), Safety (S1-S4)

Value Alignment (V1-V3)