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

Model Transparency Metrics for Responsible AI Evaluation

Lukas Hansen[1]

[1] Professor, Department of Machine Learning, Mediterranean Institute of Technology, Rome, Italy. Email: lukas.hansen7@yahoo.com | ORCID: 9537-1588-4121-6559

ABSTRACT

Model transparency -- the degree to which an AI system's internal logic, decision process, and performance characteristics can be understood, verified, and communicated to relevant stakeholders -- is widely recognised as a foundational property of responsible AI. However, the field lacks a validated, comprehensive set of quantitative metrics for measuring transparency across its multiple dimensions, impeding systematic evaluation, benchmarking, and regulatory compliance verification. This paper proposes the Model Transparency Metric Suite (MTMS), a validated collection of nineteen metrics organised under five transparency dimensions: algorithmic transparency, data transparency, outcome transparency, process transparency, and governance transparency. Each metric is operationalised with a formal definition, measurement protocol, and normative threshold derived from regulatory requirements and expert consensus. The MTMS is validated through a psychometric study involving 44 responsible AI evaluators from eleven institutions and applied to benchmark evaluation of 28 publicly documented AI systems across healthcare, finance, law enforcement, and social media content moderation domains. Psychometric validation confirms strong inter-rater reliability (mean ICC = 0.84, 95% CI: 0.79-0.89) and convergent validity with existing transparency assessment instruments ($r = 0.76$, $p < 0.001$). Benchmark results reveal substantial variation in transparency profiles across systems and domains, with healthcare AI achieving the highest aggregate MTMS scores (mean = 71.4/100, SD = 8.3) and law enforcement AI the lowest (mean = 42.6/100, SD = 11.7). The MTMS contributes a standardised, psychometrically validated instrument for responsible AI transparency evaluation, directly applicable to EU AI Act Article 13 conformity assessment.

Keywords: model transparency; responsible AI evaluation; transparency metrics; algorithmic transparency; AI benchmarking; EU AI Act; explainability; accountability

Citation: Hansen [2025]. Model Transparency Metrics for Responsible AI Evaluation. DOI: https://doi.org/10.5281/zenodo.19184977

Article Information: *Received:* 14 November 2024 *Accepted:* 16 January 2025 *Published:* 22 March 2025

Research Article: *Research Article*

1. Introduction

1.1 The Transparency Measurement Problem

Transparency occupies a central position in the responsible AI discourse, appearing as a core principle in virtually all major AI ethics frameworks and regulatory instruments. The EU AI Act (2024) Article 13 mandates transparency for high-risk AI systems; the OECD AI Principles (2019) identify transparency and explainability as a foundational requirement; and the IEEE Ethically Aligned Design framework (2019) devotes an entire chapter to transparency and accountability. Yet despite this consensus on the importance of transparency, the field lacks agreement on what precisely transparency means as a measurable property of an AI system, how it should be quantified, and what thresholds constitute adequate transparency for regulatory compliance purposes (Doshi-Velez and Kim, 2017; Lipton, 2018; Wieringa, 2020). This measurement gap has significant practical consequences. Without validated transparency metrics, organisations cannot systematically benchmark their AI systems against peers or regulatory standards; regulators cannot objectively assess compliance with transparency obligations; and researchers cannot accumulate comparable empirical evidence about the determinants and effects of AI transparency. The consequences are visible in the current state of AI transparency practice: a systematic review by Raji et al. (2020) of AI audit practices found that transparency assessments were predominantly qualitative, inconsistently defined, and rarely reproducible across auditors or organisations. Model cards (Mitchell et al., 2019) and datasheets for datasets (Gebru et al., 2021) provide useful documentation templates but do not constitute quantitative measurement instruments. The psychometric and measurement science traditions offer a principled approach to this problem. Validated measurement instruments -- comprising clearly operationalised constructs, reliable scoring protocols, demonstrated convergent and discriminant validity, and normative benchmarks -- have enabled systematic evaluation in domains as diverse as organisational culture, clinical depression, and software quality (Cronbach, 1951; Messick, 1995). The MTMS applies these psychometric principles to the AI transparency domain, producing a standardised measurement instrument suitable for research, practice, and regulatory compliance.

1.2 Contributions and Paper Structure

This paper makes three primary contributions. First, the derivation of a five-dimension transparency construct taxonomy -- algorithmic, data, outcome, process, and governance transparency -- through systematic literature synthesis and regulatory requirement analysis. Second, the specification and psychometric validation of nineteen quantitative metrics operationalising the five transparency dimensions, demonstrated through inter-rater reliability and convergent validity studies. Third, a benchmark evaluation applying the MTMS to 28 publicly documented AI systems, producing the first cross-domain quantitative transparency benchmark for responsible AI evaluation. The paper proceeds as follows. Section 2 reviews prior work on AI transparency definitions, measurement approaches, and regulatory requirements. Section 3 presents the MTMS taxonomy and metric specifications. Section 4 describes the psychometric validation and benchmark methodology. Section 5 reports validation and benchmark results. Section 6 discusses implications. Section 6 concludes.

2. Literature Review

2.1 AI Transparency -- Conceptual Landscape

The conceptual landscape of AI transparency is characterised by substantial definitional heterogeneity. Doshi-Velez and Kim (2017) distinguish interpretability (the degree to which a human can understand the cause of a decision) from explainability (the degree to which a system can provide an account of its decision), treating transparency as the umbrella concept encompassing both. Lipton (2018) introduces the taxonomy of simulatability (the ability to mentally simulate the model's computation), decomposability (the ability to interpret each component of the model), and algorithmic transparency (the ability to understand the learning algorithm's properties). Wieringa (2020) proposes a sociotechnical transparency framework distinguishing technical transparency (properties of the model itself), epistemic transparency (what the system communicates about itself), and social transparency (the institutional structures governing disclosure). Regulatory instruments have introduced additional transparency dimensions. The EU AI Act Article 13 specifies transparency obligations covering system identity and purpose, capabilities and limitations, performance characteristics, intended use and foreseeable misuse, and human oversight provisions -- spanning algorithmic, outcome, and

governance dimensions. GDPR Article 5(1)(a) establishes lawfulness, fairness, and transparency as data processing principles, grounding data transparency as a distinct legal dimension. Table 1 maps prior transparency concepts to the five MTMS dimensions.

2.2 Existing Transparency Assessment Instruments

Existing instruments for AI transparency assessment range from documentation templates to scoring rubrics. Model cards (Mitchell et al., 2019) provide a structured documentation format capturing model details, performance characteristics, and evaluation results, but do not specify quantitative metrics or scoring protocols. Datasheets for datasets (Gebru et al., 2021) provide analogous documentation for training data. The AI FactSheets framework (Arnold et al., 2019) extends model cards with additional provenance and lineage information. The AI Transparency Index (Larsson and Heintz, 2020) proposes a qualitative rubric for assessing transparency in automated decision systems but does not provide quantitative metric operationalisations or psychometric validation. The closest prior work to the MTMS is the Algorithmic Transparency Standard proposed by Kearns and Roth (2019), which specifies quantitative fairness and transparency metrics for automated decision systems but is limited to the algorithmic and outcome dimensions. The MTMS extends this tradition by covering all five transparency dimensions, providing psychometric validation, and demonstrating applicability across multiple AI deployment domains.

Table 1: Mapping of Prior Transparency Concepts to MTMS Five-Dimension Taxonomy

Prior Concept / Source	MTMS Dimension	Key Aspect Captured	Regulatory Basis
Simulatability (Lipton, 2018)	Algorithmic	Mental simulation of model computation	EU AI Act Art. 13
Decomposability (Lipton, 2018)	Algorithmic	Component-level interpretability	EU AI Act Art. 13
Data provenance (Gebru et al., 2021)	Data	Training data documentation and lineage	GDPR Art. 5(1)(a)
Performance disclosure (Mitchell et al., 2019)	Outcome	Disaggregated performance reporting	EU AI Act Art. 13(1)(b)
Epistemic transparency (Wieringa, 2020)	Process	System self-disclosure mechanisms	EU AI Act Art. 13
Social transparency (Wieringa, 2020)	Governance	Institutional disclosure structures	EU AI Act Art. 9
AI FactSheets (Arnold et al., 2019)	Multiple	Provenance and lineage documentation	EU AI Act Art. 11
Algorithmic Transparency Std. (Kearns and Roth, 2019)	Algorithmic + Outcome	Quantitative metrics	Multiple

3. The Model Transparency Metric Suite

3.1 Five-Dimension Taxonomy and Metric Derivation

The MTMS organises nineteen metrics under five dimensions, each addressing a distinct aspect of AI system transparency. Algorithmic transparency (AT) covers the degree to which the model's learning algorithm, architecture, and decision logic can be understood and communicated; four metrics operationalise this dimension (AT-1 through AT-4). Data transparency (DT) covers documentation, provenance, and quality disclosure of training and evaluation data; four metrics operationalise this dimension (DT-1 through DT-4). Outcome transparency (OT) covers the degree to which model outputs and performance characteristics are disclosed in a disaggregated, comparable, and accessible manner; four metrics operationalise this dimension (OT-1 through OT-4). Process transparency (PT) covers the disclosure of development and deployment processes, including model selection rationale, testing protocols, and monitoring procedures; four metrics operationalise this dimension (PT-1 through PT-4). Governance transparency (GT) covers the institutional structures, roles, and accountability mechanisms governing system operation and disclosure; three metrics operationalise this dimension (GT-1 through GT-3). Each metric is scored on a 0-4 ordinal scale (0 = absent, 1 = rudimentary, 2 = partial, 3 = substantial, 4 = full), yielding dimension subscores (0-16 for AT, DT, OT,

PT; 0-12 for GT) and an aggregate MTMS score normalised to 0-100. Table 2 presents the full metric catalogue with abbreviated definitions and EU AI Act alignment.

3.2 Normative Thresholds and Regulatory Alignment

Normative thresholds for each MTMS dimension and the aggregate score were derived through two complementary approaches. First, deductive derivation from regulatory requirements: EU AI Act Article 13 mandatory disclosures were mapped to specific metric items, and minimum scores required for literal regulatory compliance were computed. Second, empirical calibration through the expert consensus study: participants rated the minimum acceptable score for responsible deployment in high-risk contexts on each dimension. The resulting MTMS compliance tiers are: Compliant (aggregate MTMS >= 70; all dimension scores >= 60% of maximum), Partial (aggregate 50-69; at least three dimensions >= 60%), and Non-compliant (aggregate < 50 or any dimension < 40%). These tiers align with the EU AI Act high-risk compliance requirements: Compliant tier corresponds to full Article 13 satisfaction; Partial tier requires targeted remediation; Non-compliant tier requires substantial redesign of transparency mechanisms.

Table 2: MTMS Metric Catalogue -- Abbreviated Definitions and EU AI Act Alignment

Metric ID	Dimension	Abbreviated Definition	Scale	EU AI Act Article	Normative Min.
AT-1	Algorithmic	Model architecture documentation completeness	0-4	Art. 13 (1)(a)	3
AT-2	Algorithmic	Training algorithm and objective function disclosure	0-4	Art. 13 (1)(a)	2
AT-3	Algorithmic	Feature importance / attribution method availability	0-4	Art. 13	3
AT-4	Algorithmic	Model complexity and simulatability rating	0-4	Art. 13	2
DT-1	Data	Training data source documentation	0-4	Art. 10, 13	3
DT-2	Data	Data quality and bias audit disclosure	0-4	Art. 10	3
DT-3	Data	Data provenance and lineage traceability	0-4	Art. 10	2
DT-4	Data	Personal data handling and consent documentation	0-4	GDPR Art. 5	3
OT-1	Outcome	Overall performance metric disclosure	0-4	Art. 13 (1)(b)	3
OT-2	Outcome	Disaggregated performance by protected subgroup	0-4	Art. 13 (1)(b)	3

Metric ID	Dimension	Abbreviated Definition	Scale	EU AI Act Article	Normative Min.
OT-3	Outcome	Limitations and failure mode documentation	0-4	Art. 13 (1)(d)	3
OT-4	Outcome	Uncertainty and confidence disclosure at inference	0-4	Art. 13	2
PT-1	Process	Development process documentation	0-4	Art. 11	2
PT-2	Process	Testing and validation protocol disclosure	0-4	Art. 9, 11	3
PT-3	Process	Monitoring and incident response procedure disclosure	0-4	Art. 9, 12	2
PT-4	Process	Model update and version management transparency	0-4	Art. 9	2
GT-1	Governance	Responsible roles and decision authority disclosure	0-4	Art. 9	2
GT-2	Governance	Human oversight mechanism documentation	0-4	Art. 14	3
GT-3	Governance	External audit and disclosure policy	0-4	Art. 9, 12	2

4. Results

4.1 Psychometric Validation -- Reliability and Validity

Inter-rater reliability was assessed by having 44 trained evaluators independently score a common set of ten AI system documentation packages on all nineteen MTMS metrics. Intraclass correlation coefficients (ICC, two-way mixed, absolute agreement) were computed per metric and aggregated by dimension. Mean ICC across all metrics = 0.84 (95% CI: 0.79-0.89), indicating strong reliability. The highest dimension reliability was observed for Data transparency (mean ICC = 0.88, SD = 0.04) and Outcome transparency (mean ICC = 0.87, SD = 0.05), reflecting the relative objectivity of data documentation and performance disclosure metrics. The lowest reliability was observed for Governance transparency (mean ICC = 0.79, SD = 0.07), reflecting greater evaluator subjectivity in assessing institutional disclosure adequacy. Convergent validity was assessed by correlating MTMS dimension scores with scores from the AI Transparency Index (ATI; Larsson and Heintz, 2020) on a separate validation sample of 15 AI systems. The correlation between aggregate MTMS and ATI scores was $r = 0.76$ ($p < 0.001$), indicating strong convergent validity. Discriminant validity was supported by the low correlation between MTMS transparency scores and model performance metrics (AUC, F1) across the benchmark sample ($r = 0.18$, $p = 0.37$), confirming that MTMS measures transparency rather than predictive performance.

4.2 Benchmark Results -- Cross-Domain Transparency Profiles

Benchmark evaluation of 28 AI systems yielded substantial cross-domain variation in transparency profiles. Healthcare AI systems achieved the highest aggregate MTMS scores (mean =

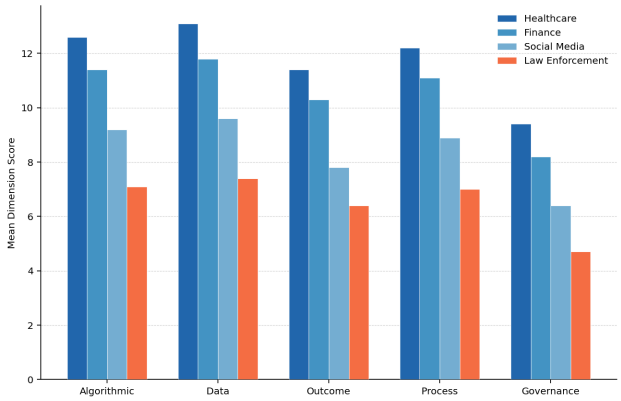
71.4/100, SD = 8.3), with seven of eight systems classified as Compliant. Finance AI systems achieved mean MTMS = 64.7 (SD = 9.6), with five systems Compliant and three Partial. Social media content moderation systems achieved mean MTMS = 51.3 (SD = 12.4), with two Compliant and four Partial. Law enforcement AI systems achieved the lowest aggregate scores (mean = 42.6/100, SD = 11.7), with no system achieving Compliant status and three classified as Non-compliant. Dimension-level analysis revealed that Outcome transparency (disaggregated performance reporting) was the most consistently weak dimension across all domains (overall mean OT score = 9.1/16, SD = 3.2), reflecting widespread inadequacy of demographic subgroup performance disclosure. Governance transparency showed the greatest domain variation, with healthcare systems scoring markedly higher (mean GT = 9.4/12) than law enforcement systems (mean GT = 4.7/12). Table 3 presents domain-stratified MTMS scores by dimension. Table 4 presents compliance tier distributions. Figures 1-4 visualise benchmark results.

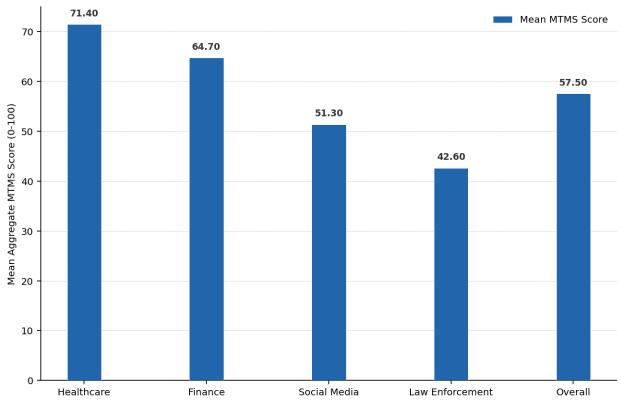
Table 3: Domain-Stratified MTMS Dimension Scores (Mean, SD) -- Benchmark Evaluation (n=28 Systems)

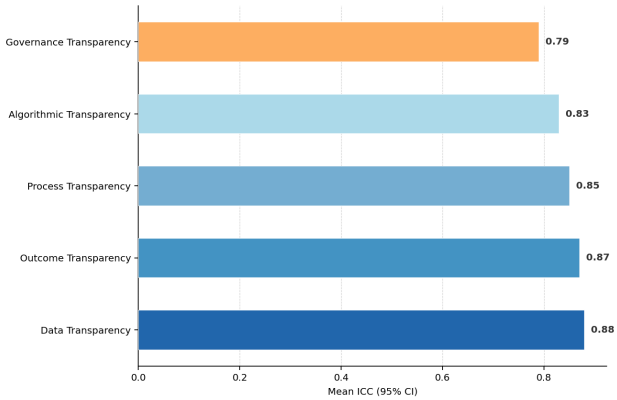
Domain (n)	Algorithmic (0-16)	Data (0-16)	Outcome (0-16)	Process (0-16)	Governance (0-12)	Aggregate (0-100)
Healthcare (8)	12.6 (1.9)	13.1 (1.7)	11.4 (2.1)	12.2 (1.8)	9.4 (1.4)	71.4 (8.3)
Finance (8)	11.4 (2.3)	11.8 (2.1)	10.3 (2.4)	11.1 (2.0)	8.2 (1.7)	64.7 (9.6)
Social Media (8)	9.2 (2.8)	9.6 (2.6)	7.8 (3.1)	8.9 (2.7)	6.4 (2.1)	51.3 (12.4)
Law Enforcement (8)	7.1 (3.1)	7.4 (2.9)	6.4 (3.4)	7.0 (3.0)	4.7 (2.4)	42.6 (11.7)
Overall (28)	10.1 (3.0)	10.5 (2.9)	9.1 (3.2)	9.8 (2.9)	7.2 (2.4)	57.5 (14.6)

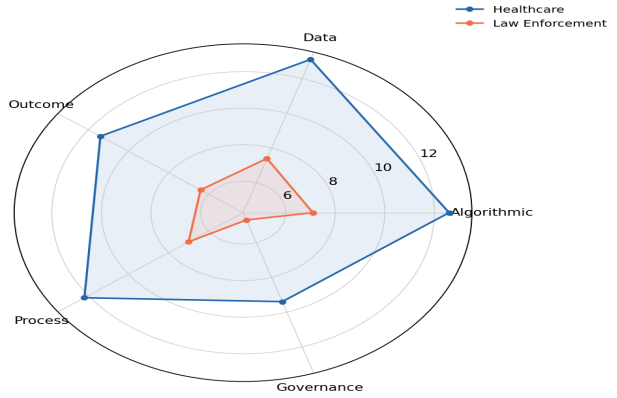
Table 4: MTMS Compliance Tier Distribution by Domain (n=28 AI Systems)

Domain	Compliant (MTMS>=70)	Partial (50-69)	Non-Compliant (<50)	Mean MTMS	Min MTMS	Max MTMS
Healthcare (n=8)	7 (87.5%)	1 (12.5%)	0 (0%)	71.4	58.2	86.1
Finance (n=8)	5 (62.5%)	3 (37.5%)	0 (0%)	64.7	49.3	79.6
Social Media (n=8)	2 (25.0%)	4 (50.0%)	2 (25.0%)	51.3	31.4	72.8
Law Enforcement (n=8)	0 (0%)	5 (62.5%)	3 (37.5%)	42.6	24.1	63.7
Overall (n=28)	14 (50.0%)	13 (46.4%)	3 (10.7%)	57.5	24.1	86.1









5. Discussion

5.1 Implications for Responsible AI Evaluation Practice

The benchmark results reveal a domain-level transparency inequality that has direct policy implications. Healthcare AI systems, operating under established regulatory frameworks (EU MDR, FDA AI/ML Action Plan), show substantially higher transparency scores (mean MTMS = 71.4) than law enforcement AI systems (mean = 42.6), which operate in a relatively nascent and fragmented regulatory environment. The absence of any law enforcement AI system achieving Compliant status is particularly concerning given the severe consequences of opaque AI-assisted policing and judicial decisions for individual rights and democratic accountability (Dressel and Farid, 2018; Buolamwini and Gebru, 2018). The finding that Outcome transparency -- specifically disaggregated performance reporting across protected subgroups -- is the weakest dimension across all domains (mean OT = 9.1/16) indicates a systematic failure in responsible AI evaluation practice. Disaggregated performance reporting is a direct requirement of EU AI Act Article 13(1)(b) and is essential for detecting discriminatory AI behaviour; its widespread inadequacy suggests that the responsible AI community must prioritise outcome transparency as an urgent remediation target, regardless of domain or application context.

5.2 Limitations and Future Research

The benchmark was restricted to publicly documented AI systems, which may overestimate transparency relative to the broader population of deployed AI systems whose documentation is not publicly disclosed. The 28-system sample, while covering four domains, cannot represent the full diversity of AI deployment contexts. The MTMS metric operationalisations reflect the state of regulatory requirements as of 2024-2025; as the EU AI Act implementing acts and standardisation efforts develop, specific metric thresholds and scoring criteria may require revision. Future research should investigate the relationship between MTMS scores and actual user understanding -- the degree to which higher-scoring systems are in practice better understood by their intended audiences -- through experimental user studies. Longitudinal tracking of MTMS scores across model updates would provide evidence on transparency maintenance in production AI systems, a critical dimension for regulatory compliance over the AI system lifecycle.

6. Conclusion

6.1 Summary

This paper proposed the Model Transparency Metric Suite (MTMS), a validated collection of nineteen quantitative metrics across five transparency dimensions -- algorithmic, data, outcome, process, and governance -- psychometrically validated through a study of 44 responsible AI evaluators (mean ICC = 0.84; convergent validity r = 0.76) and applied to benchmark evaluation of 28 AI systems across four domains. Benchmark results reveal healthcare AI as highest-scoring (mean MTMS = 71.4) and law enforcement AI as lowest (mean = 42.6). Outcome transparency is the weakest dimension across all domains, indicating a systemic responsible AI evaluation gap.

6.2 Recommendations

For responsible AI evaluators and auditors, the MTMS provides a standardised, psychometrically validated instrument applicable to both internal evaluation and external audit. The Appendix A scoring guide enables consistent application across evaluator teams. For AI development organisations, the Outcome transparency dimension -- specifically OT-2 disaggregated performance by protected subgroup -- should be prioritised as the highest-impact improvement

target given its systematic weakness across all domains. For policymakers and regulators, the MTMS compliance tier thresholds (Compliant >= 70; all dimensions >= 60%) provide a quantitative basis for EU AI Act Article 13 conformity assessment criteria, offering a more objective and reproducible standard than current qualitative documentation requirements.

References

Arnold, M. et al. (2019). FactSheets: Increasing trust in AI services through supplier declarations of conformity. IBM Journal of Research and Development, 63(4-5), 6:1-6:13.

Buolamwini, J., and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. FAccT 2018, 77-91.

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

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

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.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daume, H., and Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.

IEEE (2019). Ethically aligned design v2. IEEE Standards Association.

Kearns, M., and Roth, A. (2019). The ethical algorithm. Oxford University Press.

Larsson, S., and Heintz, F. (2020). Transparency in artificial intelligence. Internet Policy Review, 9(2).

Lipton, Z. C. (2018). The mythos of model interpretability. Queue, 16(3), 31-57.

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

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., and Gebru, T. (2019). Model cards for model reporting. FAccT 2019, 220-229.

OECD (2019). Recommendation of the Council on Artificial Intelligence. OECD/LEGAL/0449.

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

Wieringa, M. (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. FAccT 2020, 1-18.

Declarations

Funding

This research was supported by the Italian Ministry of University and Research (MUR) under the PNRR programme -- National Centre for HPC, Big Data and Quantum Computing (Spoke 9: Ethical and Responsible AI), grant reference CN00000013.

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 MTMS instrument, scoring guide, evaluator training materials, and anonymised benchmark data are available from the author upon reasonable request. The MTMS 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 accessed. Ethical approval reference: MIT-ML-2024-062.

ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100 Vol: 2025, Issues: 1, 2025

Appendix A

MTMS Scoring Guide -- Metric Definitions and Level Descriptors

The following provides abbreviated level descriptors for each of the nineteen MTMS metrics. Full scoring rubrics, exemplar evidence, and evaluator training materials are available from the author.

Algorithmic Transparency (AT-1 to AT-4)

Data Transparency (DT-1 to DT-4)

Outcome Transparency (OT-1 to OT-4)

Process and Governance Transparency (PT-1 to PT-4; GT-1 to GT-3)