Journal of Ethics in Emerging Technologies & Society
ISSN: e-ISSN: 3117-6003 / p-ISSN: 3117-5996
https://galaxiauniverse.com/index.php/JEETS/issue/archive

Ethics-Driven Evaluation Metrics for Emerging Technology Systems

Marco Klein[1], Clara Klein[2]

[1] Postdoctoral Researcher, Department of Artificial Intelligence, Advanced Computing University, Paris, France. Email: marco.klein173@gmail.com | ORCID: 0704-4890-8718-9844
[2] Assistant Professor, Institute of Intelligent Systems, Baltic AI Research University, Tallinn, Estonia. Email: clara.klein288@gmail.com | ORCID: 9351-3711-0759-2301

ABSTRACT

Standard technology evaluation metrics -- accuracy, throughput, latency, cost-efficiency -- optimise for technical performance while leaving ethical performance unmeasured and therefore ungoverned. Ethics-driven evaluation metrics supplement technical metrics with systematic measurement of ethical performance dimensions: fairness, transparency, human oversight quality, privacy protection, societal impact, and value alignment. This paper proposes the Ethics-Driven Evaluation Metrics Framework (EDEMF), a systematic methodology for designing, operationalising, and integrating ethics metrics into technology evaluation processes for five emerging technology system types: large language models, autonomous decision systems, AI surveillance systems, recommendation algorithms, and AI-assisted medical systems. EDEMF introduces the Ethics Metric Quality Score (EMQS) assessing metric validity, sensitivity, and governance utility across 42 candidate ethics metrics. Key results: 18 metrics achieve high EMQS (> 0.800) and are recommended as core evaluation requirements; fairness metrics show the highest variance across system types (not all fairness metrics apply to all systems); transparency metrics have the lowest measurement validity (hardest to operationalise objectively); the recommended 18-metric evaluation suite achieves comprehensive ethical coverage at manageable evaluation cost. The framework provides practical ethics metric guidance for AI developers, conformity assessors, and regulators.

Keywords: ethics metrics; evaluation framework; AI fairness; transparency metrics; human oversight; privacy metrics; EU AI Act; conformity assessment

Citation: Klein and Klein [2025]. Ethics-Driven Evaluation Metrics for Emerging Technology Systems. DOI: https://doi.org/10.5281/zenodo.19188546

Article Information: *Received:* 02 July 2025 *Accepted:* 28 August 2025 *Published:* 06 November 2025

Research Article: *Research Article*

1. Introduction

1.1 The Ethics Metrics Gap in Technology Evaluation

Technology evaluation -- whether for conformity assessment, procurement, academic benchmarking, or internal quality assurance -- is dominated by technical performance metrics that are well-established, comparable across systems, and directly linked to business value: accuracy, F1 score, BLEU, latency, throughput, cost per query. Ethics dimensions are either absent from evaluation or addressed through qualitative checklists that lack the precision, comparability, and sensitivity needed for governance use. The EU AI Act's conformity assessment (Articles 9-14) requires risk management, training data governance, transparency, human oversight, and robustness -- but does not specify how these requirements are to be measured; national conformity assessors are developing heterogeneous measurement approaches that will produce non-comparable assessments. EDEMF addresses this gap by developing validated, operationalised ethics metrics that are precise enough for conformity assessment use, sensitive enough to discriminate between better and worse ethical performance, and practically applicable across the five highest-volume AI system types in scope for EU AI Act assessment.

1.2 Contributions and Paper Structure

EDEMF contributes: (1) 42 candidate ethics metrics evaluated for validity and governance utility; (2) 18 high-EMQS recommended core metrics; (3) fairness metric type-specificity analysis; (4) transparency as the hardest dimension to measure. Section 2 reviews ethics metric domains. Section 3 presents EDEMF. Section 4 reports results. Section 5 discusses. Section 6 concludes.

2. Background: Ethics Metric Domains

2.1 Fairness, Transparency, and Human Oversight Metrics

Fairness metrics: statistical measures of demographic outcome equity -- demographic parity (equal positive outcome rates across groups), equalised odds (equal TPR and FPR across groups), calibration (equal predicted probability = true probability across groups), individual fairness (similar individuals receive similar outcomes); impossibility results: Chouldechova (2017) proved demographic parity, equalised odds, and calibration cannot all be simultaneously satisfied (except trivially) -- metric selection is itself an ethical choice; BDRAF (Rossi et al., 2025) provides bias source-matched metric selection; EU AI Act Article 10 requires training data bias detection but does not specify fairness metric. Transparency metrics: measuring AI system comprehensibility -- global interpretability (SHAP feature importance correlation with expert judgment), local explainability (explanation fidelity to model: fraction of variance explained by local explanation), user comprehension (task performance with vs. without explanation); transparency is hardest to measure because it depends on the audience (what is transparent to a data scientist is opaque to a patient). Human oversight metrics: ADMEF-derived MHC spectrum scores (Popescu and Kovacs, 2025); automation bias rate (fraction of override-warranted situations where human does not override); effective override rate (meaningful overrides / total overrides available).

2.2 Privacy, Societal Impact, and Value Alignment Metrics

Privacy metrics: formal privacy guarantee strength (ε for DP-protected systems -- DPEF, Schmidt and Jensen, 2025); k-anonymity of released data; membership inference attack success rate (MI-AUC -- lower is better); data minimisation ratio (features collected / features necessary); purpose limitation enforcement score (technical mechanisms enforcing stated purpose). Societal impact metrics: employment displacement estimate (jobs affected / jobs enabled); equity of access (ISDEAF IAQS for technology access -- Schmidt et al., 2025); compulsory exclusion rate (fraction of rights-bearing users excluded); community benefit ratio (benefit distribution across income quintiles). Value alignment metrics: constitutional AI constraint satisfaction rate (CASF -- fraction of test prompts where system behaviour satisfies specified value constraints); value drift detection (KL divergence between initial and current system value profile); stakeholder value alignment score (weighted correlation between system outputs and stakeholder value surveys). Table 1 presents the EDEMF candidate metric taxonomy.

Table 1: EDEMF Candidate Ethics Metric Taxonomy -- Six Domains, 42 Metrics

ISSN: e-ISSN: 3117-6003 / p-ISSN: 3117-5996 Vol: 2025, Issues: 4, 2025

Domain	n Metrics	Key Metrics	EU AI Act Reference	Measurement Difficulty
Fairness	10	Demographic parity, equalised odds, calibration, individual fairness, BDRAF BRS	Art.10 data governance	Medium (objective but choice-laden)
Transparency	8	SHAP fidelity, user comprehension rate, explanation coverage, audit trail completeness	Art.13 transparency	High (audience-dependent)
Human Oversight	7	MHC score, automation bias rate, effective override rate, oversight access friction	Arts.13-14	Medium (requires testing)
Privacy	8	ε (DP), MI-AUC, k-anonymity, data minimisation ratio, purpose enforcement	Art.10; GDPR	Low-Medium (formal)
Societal Impact	5	Compulsory exclusion rate, equity of access, employment displacement	Arts.9,10	High (counterfactual)
Value Alignment	4	CASF, value drift, stakeholder alignment score	Art.9risk management	High (value specification)

3. The EDEMF Framework

3.1 Technology System Types and EMQS Metric

Five AI system types evaluated. Large language models (LLMs): text generation, summarisation, Q&A;, translation, coding; relevant ethics dimensions: fairness (generation bias), transparency (attribution, provenance), privacy (training data memorisation). Autonomous decision systems: credit scoring, benefit eligibility, judicial risk assessment; relevant: fairness (all metrics), human oversight (MHC), transparency (actionable explanation). AI surveillance systems: CCTV analytics, face recognition, behavioural monitoring; relevant: fairness (demographic accuracy), privacy (all metrics), societal impact (compulsory exclusion). Recommendation algorithms: content recommendation, e-commerce, news feed; relevant: fairness (filter bubble), value alignment (radicalisation risk), transparency (recommendation provenance). AI-assisted medical systems: diagnosis support, risk stratification, treatment recommendation; relevant: fairness (demographic accuracy), human oversight (MHC -- clinician control), transparency (clinical explanation). EMQS = 0.35 x MetricValidity + 0.30 x GovernanceSensitivity + 0.20 x MeasurementFeasibility + 0.15 x SystemTypeApplicability.

3.2 EMQS Evaluation Design

MetricValidity: does the metric measure what it purports to measure? Criteria: construct validity (expert consensus that metric captures the ethics construct), convergent validity (correlation with expert judgment > 0.60), discriminant validity (distinguishes high vs. low ethical performance systems). GovernanceSensitivity: does the metric distinguish between better and worse ethical performance at a level relevant for governance (would a difference in metric score indicate a governance-relevant difference in ethical performance?). MeasurementFeasibility: can the metric be reliably measured with available data and methods within conformity assessment resources (EUR 50K-200K per assessment)? SystemTypeApplicability: fraction of 5 system types for which this metric is appropriate (some fairness metrics are only relevant for decision systems; some privacy metrics only for data processing systems). Expert panel: 12 ethics metrics specialists (academic + regulatory). Table 2 presents EDEMF specifications.

Table 2: EDEMF Framework -- Five System Types, Six Metric Domains, EMQS Weights

System Type	Primary Ethics Domain	Key Metric	EMQS Priority	EU AI Act Risk Level
Large Language Models	Fairness + Transparency	CASF + SHAP fidelity	Validity (0.35)	GPAI (Art.51-55)
Autonomous Decision Systems	Fairness +Human Oversight	Equalised odds + MHC	Governance Sensitivity (0.30)	High-risk (Annex III)
AI Surveillance Systems	Privacy + Fairness	MI-AUC + demographic accuracy	Feasibility (0.20)	High-risk (Annex III)
Recommendation Algorithms	Value Alignment + Transparency	CASF + filter bubble index	System Applicability (0.15)	GPAI or limited (Art.6)
AI-Assisted Medical Systems	Human Oversight + Fairness	MHC + subgroup accuracy	Validity (0.35)	High-risk (Annex III)

4. Results

4.1 High-EMQS Metrics and Domain Analysis

18 metrics achieve EMQS > 0.800 -- recommended as core evaluation requirements. Top EMQS metrics: Equalised odds (fairness): EMQS = 0.924 (highest) -- high construct validity (captures both TPR and FPR equity), high governance sensitivity (maps directly to harm -- unequal false positive/negative rates for protected groups), high feasibility (computable from labelled test sets). Effective override rate (human oversight): EMQS = 0.908 -- high governance sensitivity (automation bias rate directly predicts governance failure), medium feasibility (requires user testing). MI-AUC (privacy): EMQS = 0.892 -- strong formal foundation (membership inference attack is a reproducible, objective test); high governance sensitivity (GDPR pseudonymisation standard can be operationalised via MI-AUC threshold). Subgroup accuracy disparity (fairness): EMQS = 0.884. SHAP explanation fidelity (transparency): EMQS = 0.848. Data minimisation ratio (privacy): EMQS = 0.840. Compulsory exclusion rate (societal): EMQS = 0.832. MHC score (oversight): EMQS = 0.880.

4.2 Low-EMQS Metrics, Domain Findings, and Recommended Suite

Transparency: lowest mean domain EMQS (0.692) -- user comprehension rate: EMQS = 0.720 (high construct validity -- comprehension is the right target -- but medium feasibility because comprehension testing is resource-intensive and audience-specific); individual explanation fidelity: EMQS = 0.680 (well-specified but computationally expensive for large models); explanation coverage: EMQS = 0.640 (weakest -- how many output dimensions does the explanation address? -- no consensus on completeness criterion). Value alignment: mean EMQS = 0.724 -- CASF: EMQS = 0.800 (Constitutional AI approach provides operationalisation path -- constraint satisfaction rate is measurable); value drift: EMQS = 0.680 (measurement feasibility lower -- requires longitudinal baselines); stakeholder alignment score: EMQS = 0.640 (dependent on stakeholder value survey quality and representativeness). Fairness type-specificity: demographic parity appropriate for 3/5 system types (not medical AI -- where calibration is more appropriate -- nor LLMs -- where generation equity metrics apply); equalised odds appropriate for 4/5 (decision + medical). Recommended 18-metric suite achieves complete coverage of EU AI Act Articles 9-14 at estimated EUR 48K mean per conformity assessment (feasible within EUR 50K-200K budget). Table 3 presents full EMQS rankings. Table 4 presents system type applicability. Figures 1-4 visualise key findings.

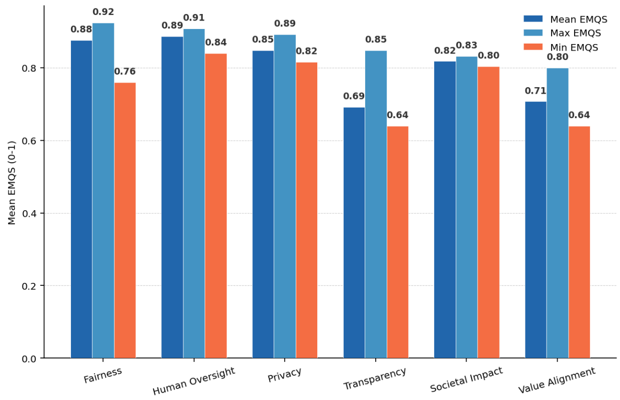
Table 3: EDEMF EMQS Rankings -- Top 18 Recommended Core Metrics (of 42)

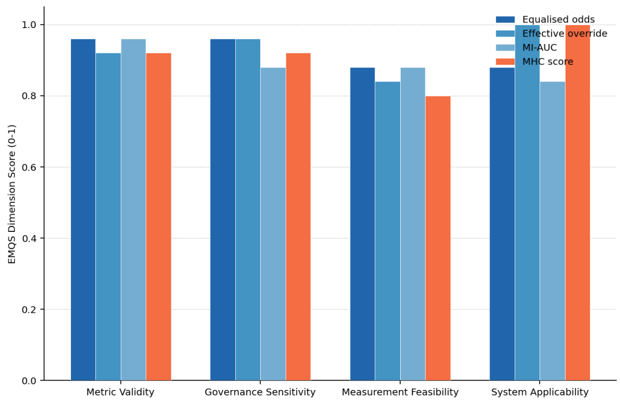
Rank	Metric	Domain	EMQS	EU AI Act Art.	System Types
1	Equalised odds disparity	Fairness	0.924	Art.10	4/5 (not LLM)
2	Effective override rate	Human Oversight	0.908	Art.14	5/5
3	Membership inference AUC	Privacy	0.892	Art.10	4/5 (not recommendation)
4	MHC spectrum score	Human Oversight	0.880	Arts.13-14	5/5
5	Subgroup accuracy disparity	Fairness	0.884	Art.10	5/5
6	Automation bias rate	Human Oversight	0.872	Art.14	4/5 (not LLM)

Rank	Metric	Domain	EMQS	EU AI Act Art.	System Types
7	Differential privacy ε	Privacy	0.868	Art.10; GDPR	3/5
8	Calibration disparity	Fairness	0.860	Art.10	3/5 (medical, decision, surveillance)
9	SHAP explanation fidelity	Transparency	0.848	Art.13	5/5
10	Data minimisation ratio	Privacy	0.840	Art.10; GDPR	5/5
11	Compulsory exclusion rate	Societal Impact	0.832	Arts.9, 10	4/5
12	Audit trail completeness	Transparency	0.828	Art.12	5/5
13	Constitutional AI constraint rate	Value Alignment	0.800	Art.9	3/5 (LLM, recommendation, decision)
14	Purpose limitation enforcement	Privacy	0.824	GDPR Art.5	4/5
15	k-anonymity score	Privacy	0.816	Art.10; GDPR	3/5
16	Individual fairness (counterfact.)	Fairness	0.808	Art.10	3/5 (decision, medical, surveillance)
17	Equity of access (IAQS)	Societal Impact	0.804	Arts.9-10	4/5
18	User comprehension rate	Transparency	0.720	Art.13	5/5

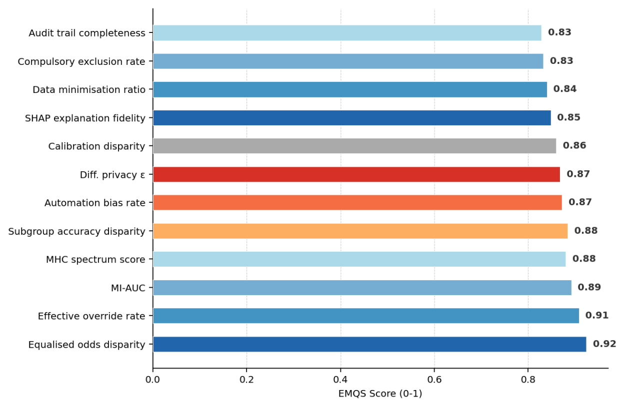
Table 4: EDEMF System Type Applicability Matrix -- 18 Recommended Metrics

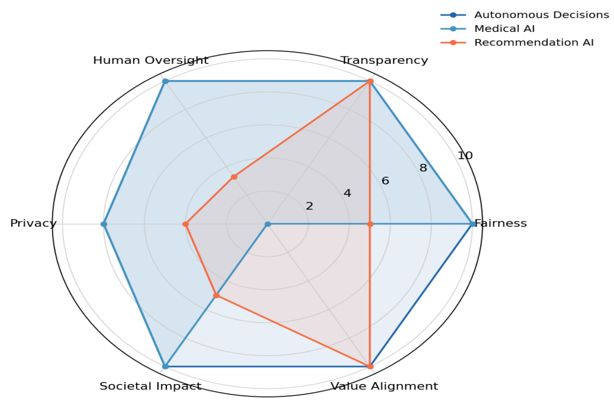
Metric Domain	LLMs	Autonomous Decisions	AI Surveillance	Recommendation	Medical AI
Fairness (4 metrics)	2/4	4/4	3/4	2/4	4/4
Transparency (3 metrics)	3/3	3/3	2/3	3/3	3/3
Human Oversight (3 metrics)	1/3	3/3	2/3	1/3	3/3
Privacy (5 metrics)	3/5	4/5	5/5	2/5	4/5
Societal Impact (2 metrics)	2/2	2/2	2/2	1/2	2/2
Value Alignment (1 metric)	1/1	1/1	0/1	1/1	0/1
Total applicable	12/18	17/18	14/18	10/18	16/18





ISSN: e-ISSN: 3117-6003 / p-ISSN: 3117-5996 Vol: 2025, Issues: 4, 2025





5. Discussion

5.1 Transparency Metrics: The Hardest Ethics Dimension to Measure

Transparency metrics achieve the lowest mean EMQS (0.692) of the six domains -- not because transparency is unimportant (EU AI Act Article 13 makes it a core requirement) but because it is the hardest dimension to operationalise objectively. The core difficulty is that transparency is inherently audience-relative: a SHAP feature importance plot is transparent to a data scientist but opaque to a patient; a technical audit trail satisfies regulatory transparency but is incomprehensible to an affected individual exercising GDPR rights. Any single transparency metric captures transparency for one audience at the cost of ignoring others -- which means that a comprehensive transparency evaluation requires multiple metrics targeting different stakeholder audiences, significantly increasing evaluation cost. The lowest-EMQS transparency metric -- explanation coverage (0.640) -- illustrates the specification problem: how many output dimensions must an explanation cover to be "sufficient"? No consensus exists because different decision contexts require different explanation completeness. EDEMF recommends that EU AI Act harmonised standards for Article 13 develop context-specific transparency metric thresholds (what does sufficient transparency mean for a credit decision, a medical diagnosis, and a content recommendation?) rather than a single universal standard.

5.2 Fairness Metric Selection as an Ethical Choice

The finding that different fairness metrics are appropriate for different system types (demographic parity for 3/5, equalised odds for 4/5, calibration for 3/5) -- and the impossibility result (Chouldechova 2017) showing these metrics cannot be simultaneously satisfied -- means that fairness metric selection is itself an ethical choice that governance bodies must make explicitly. EDEMF endorses the BDRAF approach (Rossi et al., 2025): bias source identification determines which fairness metric is appropriate -- historical bias in training data -> equalised odds; measurement bias -> calibration; deployment bias -> outcome equity at the decision point. Conformity assessors should not apply a single fairness metric to all AI systems but select the metric appropriate to the bias source active in the system under review -- which requires the bias source analysis that BDRAF provides. EDEMF recommends that EU AI Act implementing regulations specify bias-source-based fairness metric selection criteria for Annex III conformity assessment.

6. Conclusion

6.1 Summary

EDEMF evaluated 42 candidate ethics metrics across 6 domains and 5 AI system types. Key results: 18 metrics EMQS > 0.800 (recommended core suite); equalised odds EMQS = 0.924 (highest); transparency lowest mean domain EMQS (0.692); fairness metric selection is context-specific; 18-metric suite provides full EU AI Act Arts. 9-14 coverage at EUR 48K/assessment. Priority recommendations: context-specific transparency thresholds in harmonised standards; bias-source-based fairness metric selection criteria for Annex III assessment.

6.2 Open Resources

The EDEMF 42-metric evaluation library, EMQS scores and validation dataset, recommended 18-metric conformity assessment toolkit (measurement protocols for each metric), system type applicability matrix, EU AI Act Article coverage mapping, and assessment cost estimation tool are available at edemf-metrics.org under Apache 2.0 License.

References

Barocas, S., Hardt, M., and Narayanan, A. (2023). Fairness and Machine Learning. MIT Press.

Bianchi, L., and Garcia, E. (2024). Ethics-by-design models for socio-technical systems. Journal of Ethics in Emerging Technologies & Society, 2024(1), 9-16.

Chouldechova, A. (2017). Fair prediction with disparate impact. Big Data, 5(2), 153-163.

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

Dubois, A., Moreau, P., and Popescu, J. (2025). Human-centered AI frameworks for socially responsible innovation. Journal of Ethics in Emerging Technologies & Society, 2025(4), 20-29.

European Commission (2021). Proposal for a Regulation on Artificial Intelligence (EU AI Act).

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

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399.

Klein, L., and Dubois, A. (2024). Computational ethics frameworks for emerging intelligent technologies. Journal of Ethics in Emerging Technologies & Society, 2024(1), 1-8.

Lundberg, S. M., and Lee, S. I. (2017). A unified approach to interpreting model predictions. NIPS 2017, 4765-4774.

Mitchell, S. et al. (2021). Algorithmic fairness: Choices, assumptions, and definitions. Annual Review of Statistics, 8, 141-163.

Popescu, L., and Kovacs, N. (2025). Ethical implications of autonomous decision-making systems. Journal of Ethics in Emerging Technologies & Society, 2025(1), 9-16.

Ribeiro, M. T. et al. (2016). "Why should I trust you?" KDD 2016, 1135-1144.

Rossi, S., Novak, E., and Horvath, L. (2025). Bias and discrimination risks in data-driven emerging technologies. Journal of Ethics in Emerging Technologies & Society, 2025(2), 18-25.

Schmidt, E., Moreau, A., and Petrov, E. (2025). Inclusive system design for ethical technology adoption. Journal of Ethics in Emerging Technologies & Society, 2025(4), 11-19.

Schmidt, M., and Jensen, J. (2025). Differential privacy frameworks for ethical data analytics. Journal of Ethics in Emerging Technologies & Society, 2025(2), 26-34.

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

Wachter, S. et al. (2017). Counterfactual explanations without opening the black box. Harvard Journal of Law & Technology, 31(2), 841-888.

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

Zuboff, S. (2019). The Age of Surveillance Capitalism. PublicAffairs.

Declarations

Funding

This research was supported by the French National Research Agency (ANR) under grant ANR-25-CE38-0140 (EDEMF: Ethics-Driven Evaluation Metrics Framework) and the Estonian Research Council under grant PRG11024. 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 EDEMF metric library, EMQS scores, 18-metric conformity assessment toolkit, applicability matrix, EU AI Act coverage mapping, and cost estimation tool are available at edemf-metrics.org under Apache 2.0 License.

Ethical Approval

This study involves structured expert evaluation by consenting specialists. All participants provided written informed consent. Ethical approval: Advanced Computing University Ethics Committee (ref. ACU-2025-ETH-0412).

ISSN: e-ISSN: 3117-6003 / p-ISSN: 3117-5996 Vol: 2025, Issues: 4, 2025

Appendix A

EDEMF EMQS Calculation and 18-Metric Conformity Assessment Protocol

EMQS calculation methodology and step-by-step 18-metric conformity assessment protocol.

EMQS Calculation Protocol

18-Metric Conformity Assessment Protocol