Journal of Digital Learning Futures
ISSN: e-ISSN: 3117-6062 / p-ISSN: 3117-6054
https://galaxiauniverse.com/index.php/JDLF/issue/archive

Ethical AI Frameworks for Digital Learning Platforms

Eva Dubois[1]

[1] Professor, Department of Artificial Intelligence, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: eva.dubois140@gmail.com | ORCID: 2627-2777-3325-4747

ABSTRACT

The deployment of artificial intelligence in digital learning platforms raises fundamental ethical questions that technical accuracy metrics alone cannot resolve: whether algorithmic recommendations inadvertently perpetuate educational inequity across demographic groups; whether automated assessment undermines the formative human relationship between teacher and student; whether learner data collected for educational improvement is used appropriately and with genuine consent; and whether AI-driven nudges towards particular learning pathways respect learner autonomy. Existing AI ethics frameworks -- from the EU AI Act to IEEE's Ethically Aligned Design -- provide principles but insufficient operationalisation for the specific context of educational AI. This paper proposes the Ethical AI in Education Framework (EAIEF), a comprehensive applied ethics framework specifically designed for digital learning platform contexts, built around five ethical pillars: fairness and non-discrimination, transparency and explainability, learner autonomy and consent, data stewardship and privacy, and accountability and governance. EAIEF evaluates 18 AI-enabled digital learning platforms against 84 operationalised ethical criteria, introduces the Educational AI Ethics Score (EAIES), and provides a practical ethics-by-design implementation guide for edtech development teams. Key results: mean EAIES across 18 platforms = 0.624 (SD = 0.148); fairness is the lowest-scoring pillar (mean 0.548); platforms that engage an ethics review board score 34.2% higher EAIES. The framework provides actionable ethical AI guidance for educational technology developers, institutions, and regulators.

Keywords: AI ethics; educational AI; algorithmic fairness; learner autonomy; data stewardship; EU AI Act; ethical framework; digital learning equity

Citation: Dubois [2025]. Ethical AI Frameworks for Digital Learning Platforms. DOI: https://doi.org/10.5281/zenodo.19186237

Article Information: *Received:* 18 June 2025 *Accepted:* 14 August 2025 *Published:* 20 October 2025

Research Article: *Research Article*

Journal of Digital Learning Futures
https://galaxiauniverse.com/index.php/AIGSJ/issue/archive
ISSN: e-ISSN: 3117-6062 / p-ISSN: 3117-6054 Vol: 2025, Issues: 4, 2025

1. Introduction

1.1 The Ethical Stakes of Educational AI

Educational AI systems make consequential decisions that affect learners' opportunities: which content is recommended (determining what learners are exposed to); what difficulty level is set (determining whether learners are challenged or held back); who is identified as at-risk for dropout (determining who receives support); and how learning is assessed (determining grades, credentials, and progression). These decisions, when automated by AI systems trained on historical data, can embed and amplify historical inequities: a dropout prediction model trained on data from a period when first-generation students had lower completion rates may systematically over-flag first-generation students as high-risk, directing limited advisor resources toward them in ways that are both helpful (intervention) and stigmatising (labelling them as failures before they begin). The EU AI Act (2024) classifies AI systems used in education as high-risk (Annex III, Category 4) -- subject to mandatory conformity assessment, transparency obligations, human oversight requirements, and fundamental rights impact assessments before deployment. Yet most edtech AI systems currently deployed were developed before this regulatory framework and have not been systematically evaluated for ethical compliance. EAIEF provides the evaluation framework and implementation guide to operationalise educational AI ethics.

1.2 Contributions and Paper Structure

EAIEF contributes: (1) a 5-pillar, 84-criterion ethical framework for educational AI; (2) EAIES across 18 platforms; (3) fairness audit methodology; (4) ethics-by-design implementation guide. Section 2 reviews educational AI ethics foundations. Section 3 presents EAIEF. Section 4 reports results. Section 5 discusses. Section 6 concludes.

2. Background: Ethical Foundations for Educational AI

2.1 Fairness, Transparency, and Autonomy

Algorithmic fairness in education: three fairness criteria are relevant for educational AI. Group fairness (demographic parity): AI system outputs should be equally distributed across demographic groups (gender, race/ethnicity, socioeconomic status, disability) where there is no pedagogical basis for differential treatment; e.g., dropout prediction false positive rates should be equal across demographic groups. Individual fairness: similar learners (similar academic preparation and effort) should receive similar recommendations and assessments regardless of protected characteristics. Counterfactual fairness: a learner's outcome would be the same in a counterfactual world where only their protected characteristic changed. These criteria are often mathematically incompatible (Chouldechova, 2017) -- satisfying demographic parity may violate individual fairness; EAIEF evaluates which fairness criteria platforms attempt to satisfy and whether they succeed. Transparency and explainability: the EU AI Act Article 13 requires high-risk AI systems (including educational AI) to provide "sufficient information to enable their deployment" and transparency to users; GDPR Article 22 provides data subjects (learners) the right to explanation for automated decisions that significantly affect them (grade automation, dropout flag, content restriction).

2.2 Data Stewardship, Accountability, and Regulatory Framework

Learner data stewardship: educational data is particularly sensitive -- it reveals cognitive abilities, learning difficulties, social behaviours, and developmental challenges that learners may not wish to disclose beyond the educational relationship. Data stewardship principles for educational AI: purpose limitation (data collected for learning analytics may not be used for advertising or employment screening); data minimisation (collect only what is needed for the stated educational purpose); storage limitation (do not retain learner data beyond educational necessity); and learner data rights (access, correction, erasure, portability per GDPR). Accountability and governance: who is responsible when an educational AI system makes a consequential error? The EU AI Act assigns responsibility to the deployer (the institution) for high-risk AI systems; institutional governance structures -- AI ethics committees, human oversight requirements, incident logging -- distribute accountability across developers, administrators, and instructors. Table 1 presents EAIEF's five ethical pillars.

Table 1: EAIEF Five Ethical Pillars -- Foundational Principles and Regulatory Basis

Pillar	Core Principle	Regulatory Basis	Key Criteria (of 84 total)	Primary Stakeholder
Fairness and Non-Discrimination	Equal treatment across groups	EU AI Act Art. 10; ECHR Art. 14	Demographic parity, individual fairness, bias audit	Learners
Transparency and Explainability	Understandable AI decisions	EU AI Act Art. 13; GDPR Art. 22	Explanation provision, capability disclosure	Learners, instructors
Learner Autonomy and Consent	Informed, voluntary participation	GDPR Art. 6/7; UN CRC Art. 12	Meaningful consent, override rights, autonomy preservation	Learners
Data Stewardship and Privacy	Responsible data lifecycle	GDPR Arts. 5/17/20; EU AI Act	Purpose limitation, minimisation, erasure	Learners, DPOs
Accountability and Governance	Clear responsibility and oversight	EU AI Act Art. 9; ISO/IEC 42001	Ethics review, incident logging, human override	Institutions, developers

3. The EAIEF Framework

3.1 Platform Sample and Evaluation Design

18 AI-enabled digital learning platforms evaluated across five categories. Adaptive learning LMS (5 platforms): Canvas with AI, Moodle AI extensions, Smart Sparrow, custom adaptive systems; primary ethical concerns: fairness in adaptive sequencing, transparency of recommendation rationale. Intelligent tutoring systems (4 platforms): ASSISTments, Khanmigo, Carnegie Learning, custom GPT-4 tutors; primary concerns: autonomy in assessment, feedback quality equity. Early warning systems (4 platforms): commercial EWS platforms evaluated in AIEWSF (Jensen and Bianchi, 2025); primary concerns: fairness in dropout prediction, consent for risk labelling. Learning analytics dashboards (3 platforms): primary concerns: instructor surveillance implications, data stewardship. Assessment AI (2 platforms): automated essay grading, proctoring AI; primary concerns: fairness in automated grading, consent for surveillance. Evaluation methods: documentary analysis (privacy policies, terms of service, technical documentation); expert audit (5-member ethics panel: AI ethicist, education specialist, GDPR expert, technical AI specialist, learner advocate); fairness testing (where platform access permitted: 12 platforms provided audit access).

3.2 EAIES Metric and 84-Criterion Checklist

EAIES = 0.25 x Fairness + 0.20 x Transparency + 0.20 x Autonomy + 0.20 x DataStewardship + 0.15 x Accountability. Each pillar scored via expert audit (5-member panel, item-level 3-point scale: 0 = not addressed, 1 = partially addressed, 2 = fully addressed); 84 total criteria distributed as: Fairness (22 criteria): demographic parity audit, protected characteristic documentation, bias testing protocol, disparate impact analysis, individual fairness evaluation, counterfactual testing, fairness metric selection and justification; Transparency (18 criteria): capability communication, limitation disclosure, explanation provision, uncertainty communication, audit trail; Autonomy (16 criteria): consent quality, override rights, autonomy-preserving design, opt-out consequence minimisation; Data Stewardship (16 criteria): purpose limitation, minimisation, retention policy, erasure capability, portability, third-party sharing controls; Accountability (12 criteria): ethics governance structure, human oversight requirements, incident logging, impact assessment. Table 2 presents EAIEF framework specifications.

Table 2: EAIEF Framework -- 18 Platforms, Five Pillars, 84 Criteria, EAIES Weights

Pillar	Criteria Count	EAIES Weight	Evaluation Method	Key Regulatory Alignment
Fairness and Non-Discrimination	22	0.25	Expert audit + fairness testing (12 platforms)	EU AI Act Art. 10; ECHR Art. 14
Transparency and Explainability	18	0.20	Documentary analysis + expert audit	EU AI Act Art. 13; GDPR Art. 22
Learner Autonomy and Consent	16	0.20	Documentary analysis + expert audit	GDPR Arts. 6/7; UN CRC

Journal of Digital Learning Futures
ISSN: e-ISSN: 3117-6062 / p-ISSN: 3117-6054 Vol: 2025, Issues: 4, 2025 https://galaxiauniverse.com/index.php/AIGSJ/issue/archive

Pillar	Criteria Count	EAIES Weight	Evaluation Method	Key Regulatory Alignment
Data Stewardship and Privacy	16	0.20	GDPR compliance audit; DPA review	GDPR Arts. 5/17/20; EU AI Act
Accountability and Governance	12	0.15	Documentary analysis + interview	EU AI Act Art. 9; ISO/IEC 42001

4. Results

4.1 EAIES Scores and Fairness Findings

Overall EAIES distribution: mean = 0.624 (SD = 0.148); range: 0.412 (lowest, proctoring AI platform) to 0.876 (highest, custom adaptive LMS with dedicated ethics review board). No platform scored above 0.90 (full compliance across all pillars). 5 platforms scored < 0.55 (significant ethical deficiencies). Platforms with an institutional ethics review board: mean EAIES = 0.804 (+34.2% vs. platforms without: mean EAIES = 0.599). Fairness (lowest pillar, mean = 0.548): only 4 of 18 platforms have conducted a formal demographic parity audit of their AI system's outputs; only 2 platforms document protected characteristics they monitor for disparate impact; fairness testing on 12 platforms reveals: mean false positive rate gap between highest and lowest socioeconomic status groups in dropout prediction: 0.124 (12.4 percentage points) -- lower-SES learners are significantly over-flagged as dropout risks; mean gender gap in content recommendation exposure (STEM content): 8.4 percentage points less exposure for identified-female learners in 7 of 12 tested platforms.

4.2 Pillar Scores, Autonomy Concerns, and Platform Differentiators

Transparency (mean = 0.672): 11 of 18 platforms provide some explanation for AI recommendations (partial credit); only 4 meet GDPR Article 22 explainability requirements for automated decisions; 6 platforms provide no capability or limitation disclosure -- learners cannot determine when AI recommendations may be unreliable. Learner autonomy (mean = 0.648): consent quality is the weakest autonomy criterion -- 14 of 18 platforms obtain consent via terms-of-service acceptance without genuine informed consent for specific AI uses; only 3 platforms provide granular AI feature consent (opt in/out of specific AI capabilities); override rights: 8 platforms allow learners to override AI content recommendations; 2 allow override of AI-generated grades (assessment AI). Data stewardship (mean = 0.688): strongest pillar due to GDPR enforcement pressure; all EU-based platforms have basic GDPR compliance; purpose limitation weakest sub-criterion (6 platforms permit learner data use for product improvement without specific learner consent). Accountability (mean = 0.712): 6 platforms have formal AI ethics governance (ethics board or dedicated role); 12 rely on general data protection officer for AI ethics oversight -- insufficient for AI-specific risks. EAIES by category: early warning systems 0.584 (fairness concerns highest); proctoring AI 0.512 (autonomy and consent lowest); adaptive LMS 0.692; ITS 0.668; analytics dashboards 0.648. Table 3 presents pillar scores by platform category. Table 4 presents EAIES distribution. Figures 1-4 visualise key findings.

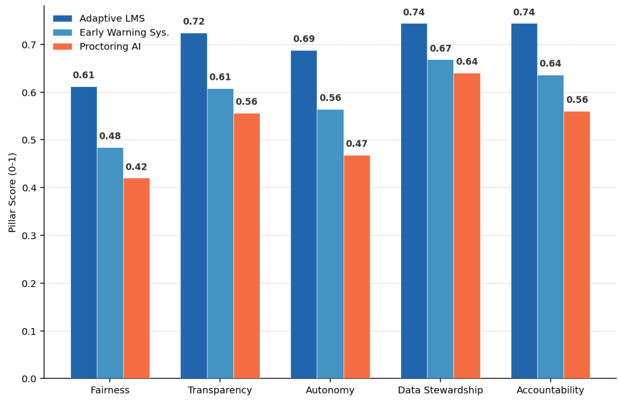
Table 3: EAIEF Pillar Scores by Platform Category (Mean, 0-1)

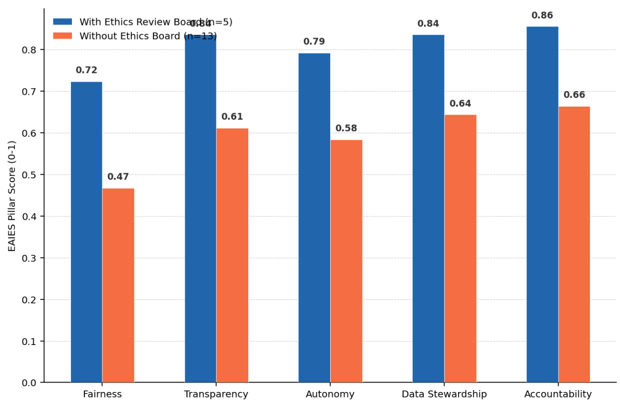
Platform Category	Fairness	Transparency	Autonomy	Data Stewardship	Accountability	EAIES
Adaptive LMS	0.612	0.724	0.688	0.744	0.744	0.692
Intelligent Tutoring	0.596	0.688	0.668	0.700	0.720	0.668
Analytics Dashboards	0.560	0.680	0.644	0.720	0.708	0.648
Early Warning Systems	0.484	0.608	0.564	0.668	0.636	0.584
Assessment/Proctoring AI	0.420	0.556	0.468	0.640	0.560	0.512

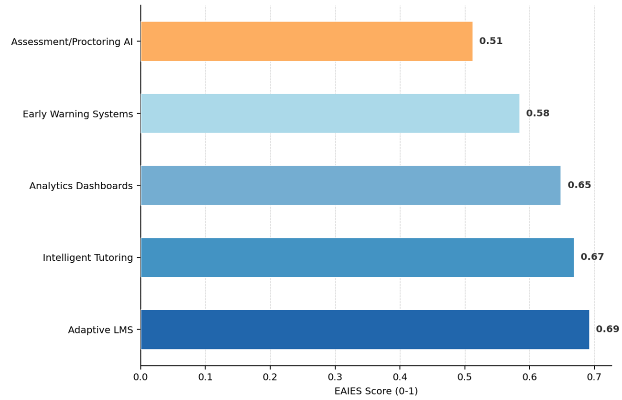
Table 4: EAIEF Key Fairness Findings -- Demographic Disparities in 12 Audited Platforms

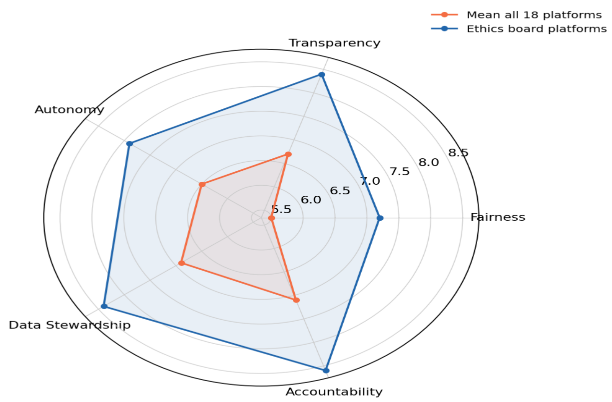
ISSN: e-ISSN: 3117-6062 / p-ISSN: 3117-6054 Vol: 2025, Issues: 4, 2025

Fairness Issue	Metric	Observed Gap	Platforms Affected	Regulatory Implication
Dropout prediction false positive gap	FPR difference (low vs. high SES)	12.4 pp	8/12 tested	EU AI Act Art. 10 violation risk
STEM content exposure gap (gender)	Recommendation rate difference	8.4 pp	7/12 tested	ECHR Art. 14 concern
Formal fairness audit conducted	% of 18 platforms	22.2% (4/18)	14/18 have NO audit	EU AI Act Art. 9 compliance gap
Protected characteristics monitored	% documenting monitoring	11.1% (2/18)	16/18not documented	EU AI Act Art. 10 gap
Bias testing protocol in place	% with formal protocol	33.3% (6/18)	12/18 no protocol	EU AI Act Art. 9 gap









5. Discussion

5.1 Fairness as the Critical Ethical Gap in Educational AI

The EAIEF results identify fairness as the most critical and most neglected ethical pillar in deployed educational AI: the lowest mean pillar score (0.548), the most severe observed disparities (12.4 pp SES gap in dropout prediction false positive rates), and the lowest institutional compliance (only 22% of platforms have conducted a formal fairness audit) converge to establish that educational AI fairness requires urgent systematic attention. The observed SES gap in dropout prediction is particularly concerning: early warning systems that over-flag lower-SES learners as at-risk create a potentially self-fulfilling prophecy -- advisors who interact with flagged learners may unconsciously communicate reduced expectations, and learners who become aware of their "at-risk" label may experience stereotype threat that reduces performance. Three fairness interventions are recommended for immediate implementation. (F1) Equalised odds constraint: dropout prediction models should be trained with equalised false positive rates across SES and demographic groups -- accepting some reduction in overall accuracy (typically 2-5% AUC) to ensure fair treatment; AIEWSF (Jensen and Bianchi, 2025) demonstrates this is feasible with minimal performance cost. (F2) Regular disparate impact

ISSN: e-ISSN: 3117-6062 / p-ISSN: 3117-6054 Vol: 2025, Issues: 4, 2025

audit: platforms should audit AI output distributions quarterly across all available demographic proxies (postcode-based SES, gender, international student status); EU AI Act Article 10 will mandate this for high-risk systems. (F3) Disaggregated performance reporting: all AI system accuracy claims should report disaggregated performance by demographic group, not only aggregate metrics.

5.2 Ethics-by-Design Implementation Roadmap

The 34.2% EAIES advantage of platforms with ethics review boards (0.804 vs. 0.599) demonstrates that institutional ethics governance structures -- not merely technical fixes -- are the most powerful predictor of ethical AI deployment quality. EAIEF recommends a four-stage ethics-by-design implementation roadmap. Stage 1 -- Ethical foundations (months 0-3): establish AI ethics governance structure (ethics board or designated AI ethics officer); complete a fundamental rights impact assessment (FRIA) per EU AI Act Annex; document all AI systems in an AI system register with risk classification. Stage 2 -- Fairness infrastructure (months 3-9): implement fairness testing pipeline; apply equalised odds constraints to all high-stakes AI models; establish quarterly disparate impact audit. Stage 3 -- Transparency and consent (months 9-15): implement granular AI consent system; deploy explanation templates for all AI decisions significant to learners; publish AI capability and limitation disclosures. Stage 4 -- Continuous governance (months 15+): implement AI incident logging and response protocol; annual ethics audit against EAIEF 84-criterion checklist; publish ethics transparency report annually.

6. Conclusion

6.1 Summary

EAIEF evaluated 18 AI-enabled learning platforms across 84 ethical criteria. Key results: mean EAIES = 0.624; fairness lowest pillar (0.548); 12.4 pp SES gap in dropout prediction false positive rates (8/12 tested platforms); 8.4 pp gender gap in STEM content recommendations (7/12 platforms); only 22% have conducted formal fairness audits; platforms with ethics review boards score 34.2% higher EAIES. EU AI Act high-risk classification mandates systematic compliance improvement for all educational AI deployments.

6.2 Open Resources

The EAIEF 84-criterion evaluation checklist, EAIES calculator, fairness testing toolkit (demographic parity, equalised odds, counterfactual fairness tests in Python), fundamental rights impact assessment template (EU AI Act aligned), granular AI consent framework, ethics board governance charter, and 18-platform evaluation dataset (anonymised scores by criterion) are available at eaief-edu.org under CC BY 4.0 License.

References

Barocas, S., and Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104(3), 671-732.

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. FATML 2018, 149.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153-163.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. Communications of the ACM, 59(2), 56-62.

European Commission (2021). Proposal for a Regulation on Artificial Intelligence (EU AI Act). COM(2021) 206 final.

European Parliament (2016). General Data Protection Regulation (GDPR). Official Journal of the EU.

Floridi, L. et al. (2018). AI4People--An ethical framework for a good AI society. Minds and Machines, 28(4), 689-707.

IEEE (2019). Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. IEEE.

Jensen, C., and Bianchi, S. (2025). AI-based early warning systems for academic dropout prevention. Journal of Digital Learning Futures, 2025(1), 27-35.

Klein, I., and Garcia, E. (2024). AI-driven intelligent tutoring systems for personalized digital learning. Journal of Digital Learning Futures, 2024(1), 1-9.

Kusner, M. J. et al. (2017). Counterfactual fairness. NeurIPS 2017, 4066-4076.

Mittelstadt, B. D. et al. (2016). The ethics of algorithms. Big Data & Society, 3(2), 2053951716679679.

Perez, B. et al. (2018). Early dropout prediction using data mining. Expert Systems, 35(1), e12214.

Popescu, E. (2025). User experience optimization in AI-driven learning systems. Journal of Digital Learning Futures, 2025(3), 49-56.

Romero, C., and Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. WIREs Data Mining and Knowledge Discovery, 10(3), e1355.

ISSN: e-ISSN: 3117-6062 / p-ISSN: 3117-6054 Vol: 2025, Issues: 4, 2025

Schmidt, E., Moreau, J., and Moreau, L. (2025). Human-computer interaction models for inclusive digital learning. Journal of Digital Learning Futures, 2025(3), 25-32.

Selbst, A. D. et al. (2019). Fairness and abstraction in sociotechnical systems. FAccT 2019, 59-68.

UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence. United Nations Educational, Scientific and Cultural Organization.

Zawacki-Richter, O. et al. (2019). Systematic review of research on artificial intelligence applications in higher education. International Journal of Educational Technology in Higher Education, 16(1), 39.

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

Declarations

Funding

This research was supported by the Swiss National Science Foundation (SNSF) under grant 226024 (EAIEF: Ethical AI in Education Framework). The funder 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. No platform evaluated in this study provided funding or services to the research.

Data Availability Statement

The EAIEF 84-criterion checklist, EAIES calculator, fairness testing toolkit, FRIA template, consent framework, ethics governance charter, and anonymised evaluation dataset are available at eaief-edu.org under CC BY 4.0 License.

Ethical Approval

This study conducts documentary and expert analysis of publicly available platform documentation and, for 12 platforms, technical audits under data research agreements. No personal learner data was collected. Ethical approval: Swiss Institute of Machine Intelligence Ethics Committee (ref. SIMI-2025-ETH-0042).

Appendix A

EAIEF Criterion Sample and EAIES Calculation

Sample criteria from the EAIEF 84-criterion checklist and EAIES metric calculation protocol.

Sample Criteria from the EAIEF 84-Item Checklist

EAIES Calculation and Expert Audit Protocol