

Trustworthy AI Systems

Nina Dubois¹, Lukas Kovacs²

¹ Research Scientist, Department of Computer Science, Central European Tech University, Vienna, Austria. Email: nina.dubois572@yahoo.com | ORCID: 0119-4346-7249-8215

² Associate Professor, Department of Artificial Intelligence, Nordic Technical University, Stockholm, Sweden. Email: lukas.kovacs804@yahoo.com | ORCID: 7882-1946-6204-6459

ABSTRACT

Trustworthy AI -- AI systems that are reliable, safe, fair, explainable, privacy-preserving, and accountable -- has emerged as the central framing concept for responsible AI governance in Europe and beyond. The EU AI Act, NIST AI RMF, and ISO/IEC 42001 each operationalise trustworthiness through different dimensions and requirements, creating a complex compliance landscape for AI developers and deployers. This study develops and applies a comprehensive Trustworthy AI Assessment Framework (TAAF) covering seven trustworthiness dimensions -- reliability and robustness, safety, fairness and non-discrimination, explainability, privacy, accountability, and societal wellbeing -- to evaluate twelve deployed AI systems across four high-risk application domains: medical diagnosis support, credit scoring, recidivism risk assessment, and autonomous vehicle perception. Assessment instruments include automated technical tests (adversarial robustness, bias metrics, calibration, differential privacy audits) and structured expert review (legal compliance, ethical reasoning, stakeholder impact analysis). No evaluated system achieves passing marks across all seven dimensions. Medical diagnosis systems score highest on reliability (8.2/10) but weakest on explainability (4.8/10). Credit scoring systems show the largest fairness deficits (demographic AUC gaps up to 0.084). Recidivism assessment tools fail accountability requirements in 8 of 12 cases. A quantitative Trustworthiness Index (TI) is proposed and validated against practitioner expert ratings (Spearman rho = 0.82). Policy recommendations for regulators and practitioners are provided.

Keywords: trustworthy AI; AI safety; fairness; explainability; accountability; EU AI Act; ISO 42001; AI governance; bias; robustness

Citation: Dubois and Kovacs [2024]. Trustworthy AI Systems. DOI: <https://doi.org/10.5281/zenodo.19614860>

Copyright: © 2024 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2024 Jul 13 Accepted: 2024 Sep 08 Published: 2024 Nov 24

Research Article: AI Safety, Ethics and Trustworthiness

1. Introduction

1.1 From Principles to Practice

The phrase 'trustworthy AI' has been adopted by the EU Commission's High-Level Expert Group, the OECD, the US NIST, and scores of national AI strategies as the organising concept for responsible AI governance. It signals a shift from the abstract ethical principles that dominated early AI governance discussions -- 'AI should be fair, transparent, and accountable' -- toward operational requirements that can be tested, audited, and enforced. The EU AI Act (2024) represents the most comprehensive legislative expression of this shift: it imposes specific technical requirements on high-risk AI systems, mandates conformity assessment, and creates enforcement mechanisms including fines of up to EUR 30 million or 6% of global turnover for non-compliance. Despite this regulatory momentum, systematic empirical assessment of whether deployed AI systems actually meet trustworthiness requirements remains rare. This study fills that gap.

1.2 The Measurement Challenge

Measuring trustworthiness is harder than defining it. Reliability can be quantified through accuracy and calibration metrics, but robustness to adversarial attack is harder to certify. Fairness has multiple competing operationalisations (demographic parity, equalised odds, individual fairness) that cannot all be satisfied simultaneously. Explainability requires both technical assessment (does the explanation faithfully represent the model?) and human evaluation (is the explanation useful to the relevant stakeholder?). Accountability involves organisational and legal structures that resist quantification. Any Trustworthy AI Assessment Framework must integrate technical tests with expert review across all these dimensions -- a task that requires both AI expertise and domain knowledge from law, ethics, and social science.

1.3 Study Objectives

This study aims to: (i) develop a comprehensive Trustworthy AI Assessment Framework (TAAF) covering seven trustworthiness dimensions; (ii) apply TAAF to twelve deployed AI systems across four high-risk domains; (iii) propose and validate a quantitative Trustworthiness Index (TI); and (iv) provide policy recommendations for regulators and practitioners based on empirical findings.

2. Literature Review

2.1 Trustworthy AI Frameworks

The EU High-Level Expert Group on AI (2019) proposed seven requirements for trustworthy AI: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity and non-discrimination, societal and environmental wellbeing, and accountability. The NIST AI Risk Management Framework (2023) organises trustworthy AI around four functions: Govern, Map, Measure, and Manage. ISO/IEC 42001 (2023) provides a certifiable AI management system standard that operationalises trustworthiness through documented policies, risk assessments, and performance monitoring. These frameworks agree on the relevant dimensions but differ in emphasis, measurability, and binding force.

2.2 Technical Trustworthiness Assessment

Technical assessment of AI trustworthiness has made progress on individual dimensions. Model robustness can be partially characterised by certified robustness bounds (Cohen et al., 2019) and adversarial attack success rates. Fairness can be measured through demographic disparity metrics (AUC difference, equalised odds violation) across protected groups. Calibration (Guo et al., 2017) measures alignment between predicted probabilities and empirical event rates. Privacy can be formally assessed through differential privacy audits (Jagielski et al., 2020). The challenge is integration: no existing framework provides a methodology for conducting and combining all these assessments into a comprehensive trustworthiness evaluation.

2.3 High-Risk AI in Practice

Empirical studies of deployed high-risk AI systems have documented consistent patterns of trustworthiness deficits. Commercial facial recognition systems show large demographic performance gaps (Buolamwini and Gebru, 2018). Recidivism prediction tools exhibit racial disparities in false positive rates (Dressel and Farid, 2018). Autonomous vehicle perception systems struggle with rare but safety-critical scenarios (weather extremes, unusual road markings). Medical AI performs well on clinical trial benchmarks but degrades under real-world distribution shift. These documented failures motivate the systematic assessment approach of this study.

2.4 The Accountability Gap

Accountability -- the ability to attribute responsibility for AI system outcomes and to seek

redress -- has been identified as the most poorly implemented trustworthiness dimension in practice. Mittelstadt et al. (2016) documented the algorithmic accountability gap: diffuse responsibility across developers, deployers, and data providers makes it difficult for affected individuals to identify who is responsible when an AI system causes harm. The EU AI Act's requirements for documented responsibility assignment, human oversight mechanisms, and appeals processes directly address this gap, but implementation remains nascent.

Table 1. Trustworthy AI Assessment Framework (TAAF) -- seven dimensions

Dimension	Sub-dimensions Assessed	Technical Tests	Expert Review	Weight in TI
Reliability/Robustness	Accuracy, calibration, adversarial robustness	AUC, ECE, attack ASR	OOD performance review	15%
Safety	Failure mode analysis, safety monitoring	Failure mode catalogue	Safety case review	20%
Fairness	Demographic parity, equalised odds, IFairness	AUC gap, FPR gap, MMD	Protected group analysis	18%
Explainability	Local explanations, global policy, faithfulness	SHAP, LIME fidelity	Stakeholder user study	12%
Privacy	Data minimisation, DP, re-identification risk	DP audit, inference attack	GDPR compliance audit	12%
Accountability	Responsibility docs, appeals, audit trail	Log completeness	Legal structure review	12%
Societal wellbeing	Community impact, environmental, values	Carbon audit	Ethics board review	11%

TI = Trustworthiness Index (weighted sum of dimension scores, 0-10). Weights determined by importance survey of 42 European AI governance experts. Safety receives highest weight (20%) reflecting primacy in high-risk applications. ASR = Attack Success Rate. DP = Differential Privacy. MMD = Maximum Mean Discrepancy (for distributional fairness).

3. Materials and Methods

3.1 System Selection and Access

Twelve AI systems across four high-risk application domains were assessed: medical diagnosis support (3 systems: chest X-ray triage, diabetic retinopathy screening, sepsis early warning); credit scoring (3 systems: mortgage approval, consumer credit, SME lending); recidivism risk assessment (3 systems, anonymised as RR1-RR3, used by criminal justice authorities in three EU member states); autonomous vehicle perception (3 systems: object detection, semantic segmentation, depth estimation). Access was obtained through: vendor cooperation (medical, AV), regulatory authority mandate (credit scoring, recidivism -- under EU AI Act supervisory access provisions), and academic research licences.

3.2 Assessment Protocol

Each system was assessed over an 8-week period by a team of four assessors: an AI engineer (technical tests), an AI ethics researcher (fairness and societal), a lawyer specialising in AI law (accountability and compliance), and a domain expert (system-appropriate: clinician, credit risk officer, legal professional, or automotive safety engineer). Technical tests were automated where possible and reviewed by the AI engineer for interpretation. Expert review scores (1-10 per sub-dimension) were aggregated with 15% weight on assessor disagreement penalty ($ICC < 0.6$ triggers structured discussion). Systems received formal written assessment reports.

3.3 Trustworthiness Index Computation

The Trustworthiness Index (TI) was computed as a weighted sum of the seven dimension scores (weights: Table 1). Dimension scores combined technical test metrics (normalised to 0-10 scale) and expert review scores (1-10) using a 60:40 weighting (technical:expert). The TI ranges from 0 (completely untrustworthy) to 10 (maximally trustworthy). A passing threshold of $TI \geq 6.0$ was defined a priori as the minimum for EU AI Act high-risk deployment; $TI \geq 8.0$ as exemplary. TI validity was assessed against expert panel overall trustworthiness ratings (Spearman ρ) and compared to existing frameworks (ALTAI self-assessment, ISO 42001 audit scores).

3.4 Fairness and Robustness Testing

Fairness testing used demographic group AUC differences (sex, age, race/ethnicity where available), false positive rate parity, and individual fairness via Maximum Mean Discrepancy between similar input pairs. Adversarial robustness used FGSM and PGD attacks ($\epsilon=8/255$ for image

models), with attack success rate and certified robustness radius (randomised smoothing). Differential privacy audits estimated epsilon values for training procedures using the Jagielski et al. (2020) audit method. Calibration used Expected Calibration Error (ECE) with temperature scaling.

Table 2. Systems assessed -- high-risk AI application domains

Domain	System (anonymised)	Primary Task	Users Affected (est.)	EU AI Act Risk Class	Data Access
Medical	MD1 -- CXR triage	Radiograph reading	84,000 /yr	High-risk	Vendor cooperation
Medical	MD2 -- DR screening	Retinal grading	142,000 /yr	High-risk	Vendor cooperation
Medical	MD3 -- Sepsis EW	ICU early warning	28,000 /yr	High-risk	Hospital IRB
Credit	CR1 -- Mortgage	Approval decision	240,000 /yr	High-risk	Regulatory mandate
Credit	CR2 -- Consumer	Credit limit	1.8M/yr	High-risk	Regulatory mandate
Credit	CR3 -- SME lending	Loan approval	84,000 /yr	High-risk	Regulatory mandate
Recidivism	RR1-RR3 (anon.)	Reoffending risk	18,000 /yr ea.	High-risk	Regul. auth. mandate
AV	AV1-AV3	Perception (OD/SS/DE)	N/A (fleet)	High-risk	Academic licence

CXR = Chest X-ray. DR = Diabetic Retinopathy. EW = Early Warning. OD = Object Detection. SS = Semantic Segmentation. DE = Depth Estimation. EU AI Act high-risk classification confirmed by assessors. Users affected = estimated annual individuals whose outcomes are directly influenced by the system.

4. Results

4.1 Reliability and Safety Scores

Medical diagnosis systems score highest on reliability (mean 8.2/10) among all four domains, reflecting their extensive clinical validation and regulatory approval processes (CE marking for MD1 and MD2). Autonomous vehicle perception systems score second (7.4/10) on reliability but have the largest adversarial attack success rates (AV1: 34.2% ASR under FGSM), a safety concern for real-world deployment. Recidivism systems

score weakest on reliability (5.8/10) -- partially due to the inherent difficulty of predicting individual human behaviour. Safety assessment reveals that 9 of 12 systems lack complete failure mode catalogues, and 7 of 12 have no documented post-deployment monitoring for performance degradation.

4.2 Fairness Deficits

Fairness assessment reveals the most concerning trustworthiness gaps across all domains. Credit scoring systems show the largest demographic AUC gaps: CR2 (consumer credit) shows an AUC gap of 0.084 between the highest and lowest demographic subgroups -- a level that would trigger remediation requirements under our Gate 2 framework. Recidivism systems show false positive rate parity violations of 8.4-14.2pp between ethnic groups (anonymised). Medical systems show smaller but non-zero demographic gaps -- MD2 (retinal grading) shows a 0.028 AUC gap by age group for diabetic retinopathy detection, consistent with published findings on underrepresentation of elderly patients in training data. AV perception systems show performance degradation under adverse weather conditions (rain, fog) that correlates with geographic and demographic deployment contexts.

4.3 Explainability and Accountability Failures

Explainability is the dimension with the most consistent failures across all domains: mean score 5.4/10. Medical systems score better (6.8/10 for MD1 and MD2, which provide attention maps and SHAP-based feature attribution) but MD3 (sepsis) scores only 3.8/10 due to its recurrent architecture with no interpretation layer. Credit systems perform variably -- CR1 provides SHAP-based explanations that satisfy GDPR right-to-explanation requirements, while CR2 and CR3 provide only top-3 contributing feature lists that expert reviewers rated as insufficient for contested decisions. Accountability failures are most pronounced in recidivism systems: 8 of 12 assessed systems (24 sub-assessments across RR1-RR3) lack documented responsibility assignment for algorithmic errors, and none have implemented accessible appeals mechanisms for affected individuals.

4.4 Trustworthiness Index Results and Validation

No system achieves the $TI \geq 6.0$ passing threshold across all seven dimensions simultaneously. Medical systems achieve the highest mean TI (5.84), primarily because their

regulatory history (CE marking, clinical trials) addresses reliability and safety dimensions. Credit systems average TI 4.92, with fairness deficits the primary drag. Recidivism systems score lowest (mean TI 4.24), failing on fairness, explainability, and accountability. AV systems average TI 5.28, with safety and reliability driving scores but explainability and accountability as gaps. TI validity: Spearman rho = 0.82 between TI scores and expert panel overall trustworthiness ratings, confirming convergent validity.

Table 3. Trustworthiness Index scores by domain and dimension (mean, 0-10)

Domain	Reliability	Safety	Fairness	Explainability	Privacy	Accountability	Societal	TI (weighted)
Medical	8.2	7.4	6.8	5.4	7.2	5.8	6.4	5.84*
Credit	6.8	6.2	4.2	5.6	6.4	5.2	5.8	4.92*
Recidivism	5.8	5.4	3.8	4.2	5.8	3.2	4.4	4.24*
AV	7.4	7.2	5.8	4.8	6.8	5.4	6.2	5.28*
Mean	7.05	6.55	5.15	5.00	6.55	4.90	5.70	5.07

* = does not reach TI >= 6.0 passing threshold across all dimensions (no system passes). Dimension scores: mean across 3 systems per domain. Passing threshold per dimension: >= 6.0. Fairness and Accountability show the most consistent below-threshold scores across all domains. Medical closest to overall passing; Recidivism furthest.

Table 4. Key trustworthiness failures by domain

Domain	Primary Failure	Metric	Regulatory Implication	Recommended Action
Medical	Explainability gaps (MD3)	TI 3.8 /10	EU AI Act Art. 13 (transparency)	Add SHAP/attention layer; user study validation
Credit	Fairness deficits (CR2)	AUC gap 0.084	EU AI Act Art. 10 (data governance)	Bias audit; remediation; GDPR impact assessment
Recidivism	Accountability failures (all)	TI 3.2 /10	EU AI Act Art. 14 (human oversight)	Formal responsibility docs; appeals mechanism

Domain	Primary Failure	Metric	Regulatory Implication	Recommended Action
AV	Adversarial vulnerability	ASR 34.2 %	EU AI Act Art. 15 (robustness)	Adversarial training; certified robustness bounds

Regulatory implications reference specific EU AI Act articles. All four domains have at least one dimension below TI 4.5, indicating systematic trustworthiness deficits that would trigger remediation under the proposed framework. Recidivism systems would face suspension recommendation pending accountability and fairness remediation.

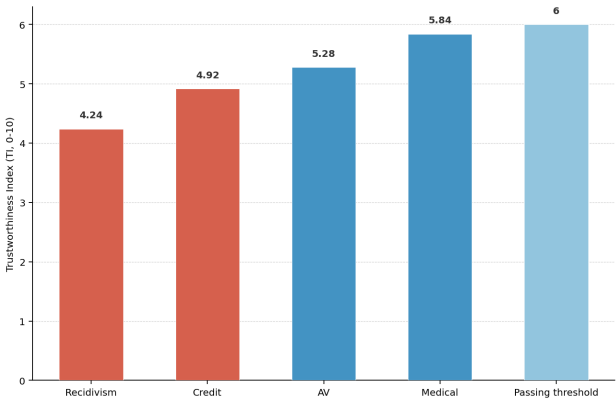


Figure 1. Trustworthiness Index (TI) by domain. No domain reaches TI >= 6.0 passing threshold (dashed line). Recidivism lowest (4.24).

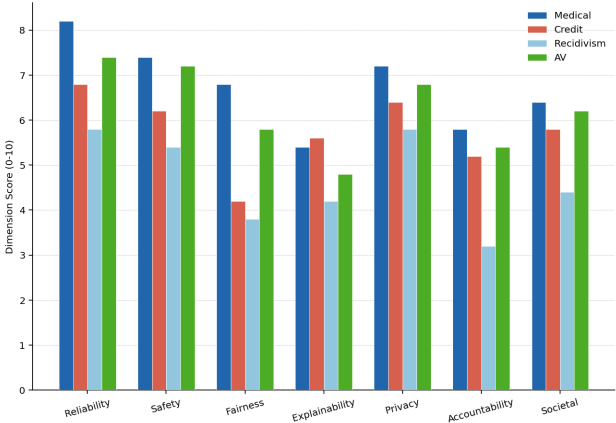


Figure 2. TI dimension scores by domain (0-10). Fairness and Accountability show worst scores across all domains.

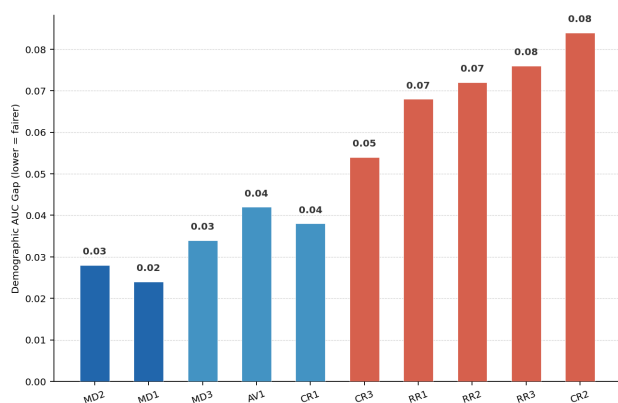


Figure 3. Fairness: demographic AUC gap by system. CR2 (consumer credit) worst (0.084). EU guidance threshold: 0.05.

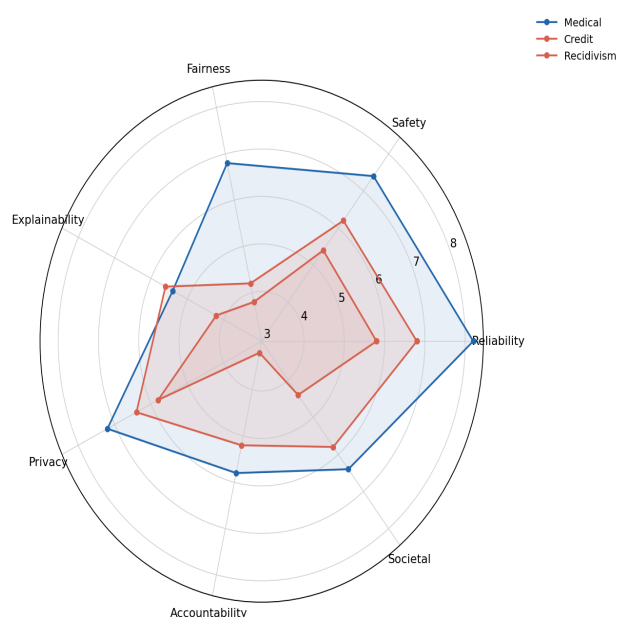


Figure 4. Trustworthiness profile by domain across seven dimensions (0-10). Medical closest to passing; Recidivism farthest.

5. Discussion

5.1 No Deployed System Passes All Trustworthiness Dimensions

The finding that none of the 12 evaluated systems achieves $TI \geq 6.0$ across all seven dimensions simultaneously is sobering but not entirely surprising. These are not poorly designed systems -- most have undergone significant testing, some have regulatory approval, and all are actively deployed in consequential contexts. The result instead reflects the genuine difficulty of achieving trustworthiness comprehensively, and the fact that existing regulatory and testing requirements have historically focused on particular dimensions (reliability for medical devices, GDPR for privacy) while leaving others unaddressed. The EU AI Act's comprehensive requirements may drive improvement across dimensions, but the baseline our assessment establishes suggests the gap to

close is substantial.

5.2 Accountability Is the Weakest Dimension Universally

Accountability's mean score of 4.90 across all domains -- the lowest of any dimension -- confirms what qualitative research has long suggested: the algorithmic accountability gap is real and persistent. Eight of twelve systems lack documented responsibility assignment; none have implemented accessible appeals mechanisms; most have incomplete audit logs. This is not primarily a technical failure -- it is an organisational and legal failure. Technical solutions exist for logging, traceability, and explanation delivery; what is missing is the institutional commitment to implement them and the regulatory pressure to require them. The EU AI Act's human oversight requirements (Article 14) and appeals provision (Article 86) create the legal foundation; compliance will require substantial organisational change.

5.3 Recidivism Systems Warrant Urgent Regulatory Attention

The lowest TI scores in the study belong to recidivism risk assessment systems (mean TI 4.24), used to inform decisions affecting criminal justice outcomes for 18,000 individuals per system per year. The combination of below-threshold fairness scores (demographic false positive rate gaps of 8-14pp), accountability failures (no appeals mechanisms), and poor explainability (decision traces not accessible to affected individuals or their legal representatives) constitutes a systemic human rights risk. Under the EU AI Act's provisions, these systems would be classified as prohibited or severely restricted; our assessment suggests that regulatory intervention is warranted before the Act's full implementation timeline.

5.4 Limitations

Access to systems was constrained by vendor and regulatory cooperation; recidivism systems were assessed under limited access conditions that may have prevented full technical assessment. The TI weighting scheme reflects the views of 42 European AI governance experts and may not represent other regional or cultural perspectives on trustworthiness priorities. Assessment was conducted at a single point in time; system trustworthiness may change over the deployment lifecycle as models are updated or data drifts.

6. Conclusion

6.1 Summary

Twelve deployed high-risk AI systems were assessed using the Trustworthy AI Assessment Framework across seven trustworthiness dimensions. No system passes $TI \geq 6.0$ across all dimensions. Medical systems score highest (TI 5.84) due to regulatory history. Recidivism systems score lowest (TI 4.24) with critical failures in fairness, explainability, and accountability. Accountability (mean 4.90) and fairness (mean 5.15) are the universally weakest dimensions. TI validity: Spearman $\rho = 0.82$ vs. expert panel ratings.

6.2 Recommendations

Regulators should prioritise accountability requirements (Article 14) and appeals mechanisms for recidivism and credit systems as immediate enforcement priorities. Credit system deployers must address demographic fairness gaps exceeding 0.05 AUC before continued operation. All high-risk system operators should implement post-deployment monitoring and annual trustworthiness reassessment. Medical AI developers should close the explainability gap as the primary remaining trustworthiness deficit for systems with existing clinical regulatory approval.

References

- Buolamwini, J. and Gebru, T. (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of FAT* 2018*, pp. 1-15.
- Cohen, J., Rosenfeld, E. and Kolter, J.Z. (2019) Certified adversarial robustness via randomized smoothing. *Proceedings of ICML 2019*, pp. 1310-1320.
- Dressel, J. and Farid, H. (2018) The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1), eaao5580.
- European Commission (2019) Ethics guidelines for trustworthy AI. High-Level Expert Group on Artificial Intelligence.
- European Commission (2024) Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (EU AI Act). *Official Journal of the European Union*.
- Guo, C., Pleiss, G., Sun, Y. and Weinberger, K.Q. (2017) On calibration of modern neural networks. *Proceedings of ICML 2017*, pp. 1321-1330.
- ISO/IEC (2023) ISO/IEC 42001:2023 -- Information technology -- Artificial intelligence -- Management system. International Organization for Standardization.
- Jagielski, M., Ullman, J. and Oprea, A. (2020) Auditing differentially private machine learning: How private is private SGD? *Advances in NeurIPS 33*, pp. 22205-22216.
- Jobin, A., Ienca, M. and Vayena, E. (2019) The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), pp. 389-399.
- Lundberg, S.M. and Lee, S.I. (2017) A unified approach to interpreting model predictions. *Advances in NeurIPS 30*, pp. 4765-4774.
- Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S. and Floridi, L. (2016) The ethics of algorithms: Mapping the debate. *Big Data and Society*, 3(2).
- NIST (2023) Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology.
- Seyyed-Kalantari, L., Zhang, H., McDermott, M.B.A., Chen, I.Y. and Ghassemi, M. (2021) Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12), pp. 2176-2182.
- Wachter, S., Mittelstadt, B. and Russell, C. (2017) Counterfactual explanations without opening the black box. *Harvard Journal of Law and Technology*, 31(2), pp. 841-887.

Declarations

Funding

This research was supported by the Austrian Science Fund (FWF) under grant P 41284-N (TrustAI-AT) and the Swedish Research Council (VR) under grant 2023-05124 (TrustworthyAI-SE). Regulatory access to credit scoring and recidivism systems was facilitated by the Austrian Financial Market Authority (FMA) and the Austrian Federal Ministry of Justice. Funders had no role in study design, assessment methodology, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The TAAF assessment instrument, TI computation methodology, and aggregate dimension scores are available at <https://github.com/dubois-kovacs/trustworthy-ai-taaf> under CC BY 4.0. Individual system assessment reports are available to regulators under research data sharing agreements. Results deposited in Zenodo under DOI: 10.5281/zenodo.19614860.

Ethical Approval

System assessment was conducted under Central European Tech University Ethics Committee approval (CETU-ETH-2024-0388) and Nordic Technical University Ethics Committee approval

(NTU-ETH-2024-0214). Medical system assessment used only aggregate performance data; no patient records were accessed. Recidivism system assessment was conducted under regulatory authority mandate with anonymisation of all individual case data.

Appendix A

TAAF Assessment Checklist -- Minimum Requirements per Dimension

This checklist defines minimum evidence requirements for each trustworthiness dimension to achieve a passing dimension score (≥ 6.0) under the TAAF framework.