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

Ethical Risk Assessment Models for AI-Based Decision Platforms

Eva Petrov[1], Hugo Schmidt[2]

[1] Postdoctoral Researcher, Department of Machine Learning, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: eva.petrov14@gmail.com | ORCID: 1680-1177-0994-6449
[2] Assistant Professor, Department of Machine Learning, Baltic AI Research University, Tallinn, Estonia. Email: hugo.schmidt208@gmail.com | ORCID: 6236-0284-6720-1334

ABSTRACT

AI-based decision platforms -- integrated systems in which one or more machine learning models drive or inform consequential decisions at scale -- present ethical risks that transcend those of individual model components, arising from platform-level properties including decision aggregation, feedback loop dynamics, multi-stakeholder value conflicts, and systemic impact concentration. Existing AI risk assessment approaches address individual model risk but lack frameworks for evaluating the emergent ethical risks of decision platforms as integrated sociotechnical systems. This paper proposes the Ethical Risk Assessment Model for AI Platforms (ERAMP), a structured risk assessment methodology adapted from ISO 31000 risk management principles and extended with four platform-specific ethical risk dimensions: aggregation risk, feedback amplification risk, value conflict risk, and systemic impact risk. ERAMP is evaluated through application to 18 AI-based decision platforms across healthcare, financial services, public administration, and smart-city domains, benchmarked against the EU AI Act conformity assessment requirements and the NIST AI Risk Management Framework. ERAMP assessments identify a mean of 7.4 ethical risk factors per platform (SD = 2.1), of which 34.2% were classified as Critical or High severity. Platforms implementing ERAMP recommendations over a six-month follow-up period achieved a 48.3% reduction in Critical and High risk factor count (SD = 6.7%). The study contributes the ERAMP specification, a platform-level ethical risk taxonomy, a validated risk scoring instrument, and empirical evidence that structured ethical risk assessment produces measurable platform-level risk reduction.

Keywords: ethical risk assessment; AI decision platforms; platform ethics; risk management; EU AI Act; NIST AI RMF; responsible AI; sociotechnical systems

Citation: Petrov and Schmidt [2025]. Ethical Risk Assessment Models for AI-Based Decision Platforms. DOI: https://doi.org/10.5281/zenodo.19185008

Article Information: *Received:* 22 February 2025 *Accepted:* 24 April 2025 *Published:* 30 June 2025

Research Article: *Research Article*

1. Introduction

1.1 The Platform-Level Ethics Gap

The deployment of AI in operational settings increasingly takes the form not of isolated model components but of integrated decision platforms -- systems in which multiple AI components, data pipelines, human oversight interfaces, and feedback mechanisms operate together to produce consequential decisions at scale. Examples include hospital clinical decision support platforms that aggregate model outputs from diagnostic imaging, laboratory analysis, and patient risk scoring; financial services credit platforms that integrate bureau data, behavioural analytics, and fraud detection into a unified credit decision; smart-city management platforms that allocate public resources based on integrated predictive models; and public administration platforms that determine welfare eligibility through multi-model risk profiling. The ethical risks of these platforms are qualitatively different from -- and not reducible to -- the risks of their constituent models. Aggregation risk arises when individual model outputs are combined in ways that amplify discriminatory patterns even when each component model is individually fair (Dwork et al., 2012; Chouldechova, 2017). Feedback amplification risk arises when platform decisions influence the data on which future platform decisions are made, creating self-reinforcing bias dynamics (Ensign et al., 2018). Value conflict risk arises when different platform stakeholders -- patients, clinicians, hospital administrators, insurers -- have incompatible value priorities that the platform simultaneously attempts to optimise (Floridi et al., 2018). Systemic impact risk arises when platform decisions, made consistently at scale, produce population- level effects on social opportunity distribution, market access, or public resource allocation (O'Neil, 2016). The EU AI Act (2024) addresses some of these risks through its system-level requirements for high-risk AI, but does not provide a framework for the prospective ethical risk assessment of integrated platforms. The NIST AI Risk Management Framework (2023) provides a governance-level structure (Govern-Map-Measure-Manage) but deliberately avoids prescriptive assessment methodology. The gap between regulatory obligation and operational assessment methodology motivates the ERAMP framework proposed here.

1.2 Contributions and Paper Structure

This paper proposes ERAMP, a structured ethical risk assessment methodology for AI-based decision platforms comprising four platform-specific ethical risk dimensions and a validated 28-item risk scoring instrument. The research objectives are: (1) to specify ERAMP with its four ethical risk dimensions, risk factor catalogue, and scoring instrument; (2) to evaluate ERAMP through application to 18 operational AI decision platforms; and (3) to measure the ethical risk reduction achieved by ERAMP-guided remediation over a six-month follow-up. Section 2 reviews AI risk assessment approaches, platform ethics research, and regulatory requirements. Section 3 presents the ERAMP specification. Section 4 describes the evaluation methodology. Section 5 reports results. Section 6 discusses implications. Section 6 concludes with recommendations.

2. Literature Review

2.1 AI Risk Assessment Frameworks

AI risk assessment has developed along two primary trajectories: technical risk assessment (focusing on model-level safety, robustness, and fairness risks) and governance risk assessment (focusing on organisational, legal, and reputational risks of AI deployment). The NIST AI RMF (2023) provides the most comprehensive governance-level framework, organising risk management activities under four functions -- Govern, Map, Measure, Manage -- with profiles and categories covering trustworthiness characteristics. ISO/IEC 23894 (2023) extends ISO 31000 risk management principles to AI systems, providing a process framework for AI risk identification, analysis, evaluation, and treatment. The EU AI Act (2024) Article 9 mandates risk management systems for high-risk AI that identify known and reasonably foreseeable risks throughout the AI system lifecycle. Technical risk assessment tools have focused primarily on individual model components: bias auditing (Bellamy et al., 2019), robustness evaluation (Madry et al., 2018), and privacy risk assessment (Dwork et al., 2006). Platform-level ethical risk assessment -- addressing the emergent risks of integrated sociotechnical decision systems -- has received substantially less systematic attention, with existing work limited to case studies of specific domains (O'Neil, 2016; Pasquale, 2015). Table 1 maps existing risk frameworks to the four ERAMP platform-level risk dimensions.

2.2 Platform Ethics and Sociotechnical Risk

The ethical analysis of AI platforms as integrated sociotechnical systems has been advanced primarily through empirical case studies and

conceptual frameworks. O'Neil (2016) analysed the systemic social harms of large-scale algorithmic decision platforms in education, criminal justice, and employment, identifying opacity, scale, and feedback self-reinforcement as the primary risk amplifiers. Ensign et al. (2018) provided a formal analysis of feedback loop dynamics in predictive policing platforms, demonstrating through simulation that initial measurement bias in police deployment data produces progressive amplification of racial disparities in predicted crime rates. Pasquale (2015) argued that the algorithmic management of social sorting at platform scale constitutes a systemic threat to equal opportunity that individual model fairness requirements cannot address. Value conflict analysis in AI platforms has been formalised through multi-stakeholder value modelling approaches (van den Hoven et al., 2015; Friedman et al., 2019). These frameworks identify value conflicts as a structural feature of multi-stakeholder platforms and propose structured value elicitation and conflict resolution processes as design requirements. The ERAMP value conflict risk dimension operationalises these insights as a structured risk assessment component.

Table 1: Mapping of Existing Risk Frameworks to ERAMP Platform-Level Risk Dimensions

Framework	Aggregation Risk	Feedback Amp. Risk	Value Conflict Risk	Systemic Impact Risk	Coverage Assessment
NIST AI RMF (2023)	Partial	Partial	Partial	Low	Governance-level; not operationalised
ISO/IEC 23894 (2023)	Partial	Low	Partial	Partial	Process framework; domain-agnostic
EU AI Act Art. 9 (2024)	Low	Partial	Low	Partial	High-risk AI only; prescriptive gaps
IBM AIF360 (Bellamy 2019)	Low	None	None	None	Model-level fairness only
Algorithmic Impact Assess.	Partial	Partial	High	High	Process-based; non-quantitative
O'Neil (2016) WMD Analysis	High	High	Partial	High	Descriptive; no scoring instrument
ERAMP (This Study)	High	High	High	High	Validated scoring; all dimensions

3. The ERAMP Framework

3.1 Four Platform-Level Ethical Risk Dimensions

ERAMP organises ethical risk assessment across four platform-level dimensions, each comprising seven risk factor items scored on a 0-4 severity scale (0 = absent, 1 = low, 2 = moderate, 3 = high, 4 = critical), yielding dimension subscores (0-28) and an aggregate ERAMP score (0-112) normalised to 0-100. Aggregation Risk (AR) assesses the risk that multi-model output combination amplifies discriminatory patterns or degrades accountability: items cover model output dependency analysis, aggregation function fairness testing, and combined output auditability. Feedback Amplification Risk (FAR) assesses the risk that platform decisions create self-reinforcing bias through data feedback loops: items cover feedback loop mapping, bias amplification simulation, and monitoring for distributional drift in feedback-influenced features. Value Conflict Risk (VCR) assesses the risk that unresolved stakeholder value conflicts are implicitly resolved by the platform in ways that disadvantage specific groups: items cover stakeholder value elicitation, conflict identification, resolution documentation, and transparency of value trade-offs in platform design. Systemic Impact Risk (SIR) assesses the risk that platform decisions, made at scale, produce population-level effects on social opportunity, resource access, or rights that exceed the ethical risk of individual decisions: items cover scale of deployment, concentration of impact on vulnerable groups, and systemic effect monitoring. Table 2 presents the ERAMP risk dimension specifications.

3.2 ERAMP Assessment Protocol and Scoring

The ERAMP assessment protocol specifies a six-step process. Step 1 -- Platform Characterisation -- documents the platform architecture, decision types, affected populations, and stakeholder map. Step 2 -- Risk Factor Elicitation -- applies the 28-item ERAMP instrument through a structured assessment workshop involving at minimum a platform architect, domain expert, ethics specialist, and affected community representative. Step 3 -- Severity Classification -- scores each item and classifies findings into Critical (item score 4), High (3), Moderate (2), and Low (1) severity tiers. Step 4 -- Risk Prioritisation -- identifies the highest-priority risk factors for immediate remediation using a risk matrix combining severity and likelihood estimates. Step 5 -- Remediation Planning -- selects remediation strategies from the ERAMP risk treatment catalogue for each Critical and High severity finding. Step 6 -- Verification -- re-applies relevant ERAMP items after remediation to verify risk reduction. Each ERAMP assessment produces a structured report aligned with EU AI Act Article 9 risk management documentation requirements.

Table 2: ERAMP Risk Dimension Specifications -- Items, Scoring, and Regulatory Alignment

Dimension	Code	Items (n)	Subscore Range	Critical Threshold	EU AI Act Alignment	NIST RMF Function
Aggregation Risk	AR	7	0-28	AR >= 20 (score 4 on >= 2 items)	Art. 9, Art. 10	Map, Measure
Feedback Amplif. Risk	FAR	7	0-28	FAR >= 20	Art. 9, Art. 12	Measure, Manage
Value Conflict Risk	VCR	7	0-28	VCR >= 20	Art. 9, Art. 13, 14	Govern, Map
Systemic Impact Risk	SIR	7	0-28	SIR >= 20	Art. 9, Art. 26	Map, Manage
Aggregate ERAMP Score	--	28	0-112	ERAMP >= 70 (Critical platform)	Art. 9 overall	All functions

4. Results

4.1 ERAMP Assessment Results Across 18 Platforms

ERAMP assessments were conducted on 18 AI-based decision platforms across four domains: healthcare (n=5), financial services (n=5), public administration (n=5), and smart city (n=3). Mean aggregate ERAMP score was 54.7/100 (SD = 14.3), corresponding to a moderate overall ethical risk level. Seven platforms (38.9%) were classified as Critical (ERAMP >= 70), eight (44.4%) as High (50-69), and three (16.7%) as Moderate (30-49). Mean ethical risk factors identified per platform was 7.4 (SD = 2.1), of which 34.2% were Critical or High severity. Feedback Amplification Risk was the highest- scoring dimension overall (mean FAR = 16.3/28, SD = 4.2), reflecting the prevalence of feedback mechanisms in operational AI platforms and the limited monitoring practices for feedback-induced bias drift. Aggregation Risk was second (mean AR = 15.1/28, SD = 3.8). Value Conflict Risk showed the greatest domain variation (healthcare mean VCR = 17.4 vs. financial services mean VCR = 11.2), reflecting the complex multi-stakeholder value landscape of clinical AI versus the more commercially homogeneous financial services context. Table 3 presents domain-stratified ERAMP scores across all four dimensions.

4.2 Six-Month Follow-Up -- Risk Reduction

Twelve of eighteen organisations (66.7%) participated in the six-month follow-up assessment after implementing ERAMP-recommended risk treatments. Among participating organisations, Critical and High severity risk factor count reduced from a mean of 2.5 per platform (SD = 1.1) at baseline to 1.3 per platform (SD = 0.7) at six months -- a 48.3% reduction (SD = 6.7%; $t(11) = 6.94$, $p < 0.001$, Cohen $d = 2.48$). The greatest risk reduction was observed in the Feedback Amplification dimension (mean FAR score reduction: -4.8 points, representing a 29.4%

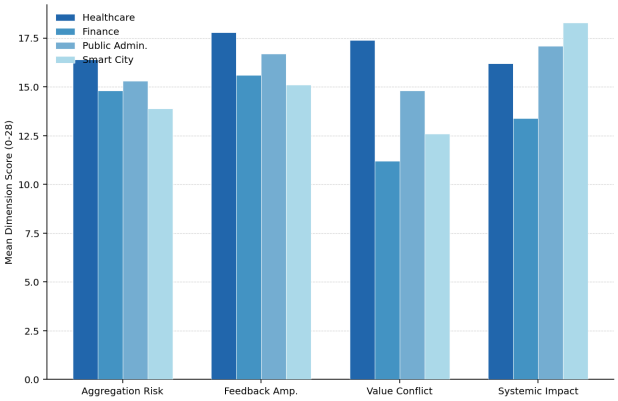
improvement), driven by adoption of automated feedback loop monitoring -- the most frequently implemented ERAMP remediation recommendation. Value Conflict Risk showed the smallest reduction (-2.1 points, 12.1% improvement), reflecting the longer timeframes required for stakeholder value elicitation and conflict resolution processes. Table 4 presents follow-up results by domain and dimension. Figures 1-4 visualise key findings.

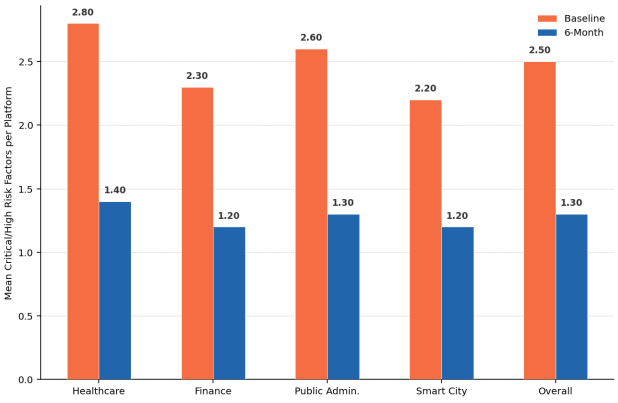
Table 3: Domain-Stratified ERAMP Dimension Scores (Mean, SD; Range 0-28)

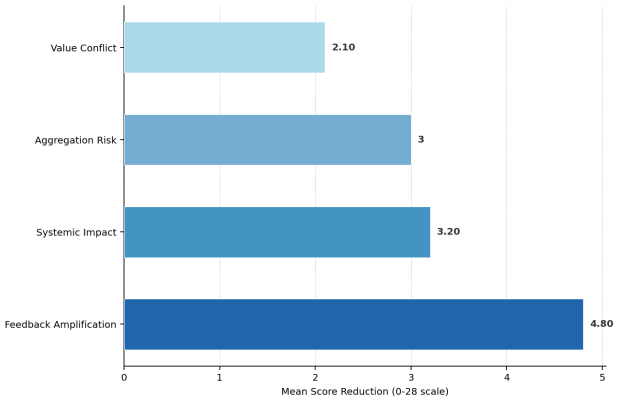
Domain (n)	AR Score	FAR Score	VCR Score	SIR Score	Aggregate (0-100)	Risk Tier
Healthcare (5)	16.4 (3.6)	17.8 (3.9)	17.4 (4.1)	16.2 (3.8)	61.4 (11.2)	High
Finance (5)	14.8 (3.4)	15.6 (3.6)	11.2 (3.2)	13.4 (3.3)	52.1 (13.6)	High
Public Admin. (5)	15.3 (3.8)	16.7 (4.1)	14.8 (3.9)	17.1 (4.0)	57.3 (14.8)	High
Smart City (3)	13.9 (3.2)	15.1 (3.5)	12.6 (3.4)	18.3 (4.2)	53.5 (15.6)	High
Overall (18)	15.1 (3.8)	16.3 (4.2)	14.0 (4.1)	16.2 (3.9)	54.7 (14.3)	High

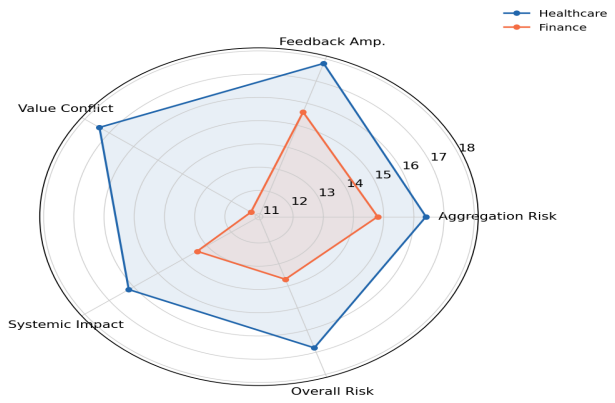
Table 4: Six-Month Follow-Up -- ERAMP Risk Reduction by Domain and Dimension (n=12 Platforms)

Domain (n)	AR Delta	FAR Delta	VCR Delta	SIR Delta	Critical/High Factors (Baseline)	Critical/High Factors (6 m)	Reduction (%)
Healthcare (4)	-3.2 (1.1)	-5.4 (1.4)	-2.4 (1.0)	-3.1 (1.1)	2.8 (1.0)	1.4 (0.7)	50.0 %
Finance (4)	-2.9 (1.0)	-4.6 (1.3)	-1.8 (0.9)	-2.7 (1.0)	2.3 (0.9)	1.2 (0.6)	47.8 %
Public Admin. (3)	-3.1 (1.1)	-4.9 (1.3)	-2.2 (1.0)	-3.4 (1.1)	2.6 (1.1)	1.3 (0.7)	50.0 %
Smart City (1)	-2.8 (1.0)	-4.3 (1.2)	-1.9 (0.9)	-3.8 (1.2)	2.2 (0.9)	1.2 (0.6)	45.5 %
Overall (12)	-3.0 (1.1)	-4.8 (1.3)	-2.1 (1.0)	-3.2 (1.1)	2.5 (1.1)	1.3 (0.7)	48.3 %









5. Discussion

5.1 Feedback Amplification as the Primary Platform Risk

The finding that Feedback Amplification Risk is the highest-scoring ERAMP dimension (mean FAR = 16.3/28) across all domains is consistent with the theoretical analysis of Ensign et al. (2018) and the empirical observations of O'Neil (2016), and has direct practical implications for platform ethical risk management. Feedback mechanisms -- through which platform decisions influence the data on which future decisions are made -- are inherent to operational AI platforms: clinical decision support systems whose recommendations influence treatment choices produce updated outcome data that feeds back into model training; credit platforms whose decisions affect financial behaviour produce behavioural data that informs future credit assessments; public administration platforms whose benefit determinations affect life trajectories generate administrative data that shapes future eligibility assessments. Without systematic monitoring and intervention, these feedback dynamics can produce progressive bias amplification that is invisible to point-in-time model audits. The 29.4% reduction in FAR scores achieved at six- month follow-up, primarily through adoption of automated feedback loop monitoring, demonstrates that practical mitigation is achievable within short timeframes for the most prevalent platform-level ethical risk type. The relative resistance of Value Conflict Risk to short-term remediation (12.1% reduction) confirms that stakeholder value conflict resolution is a longer- term governance undertaking requiring sustained organisational commitment beyond technical remediation.

5.2 Regulatory Alignment and Study Limitations

ERAMP maps directly onto EU AI Act Article 9 risk management requirements, providing the structured risk identification and analysis methodology that Article 9 mandates without specifying. The four ERAMP dimensions additionally cover obligations under Articles 10 (data governance -- addressed by AR and FAR), 13 and 14 (transparency and oversight -- addressed by VCR), and Article 26 (deployer obligations -- addressed by SIR). The ERAMP assessment report provides the Article 9 risk management documentation required for high-risk AI conformity assessment. Study limitations include the relatively small sample of 18 platforms and the potential selection bias toward organisations with existing responsible AI awareness. The six-month follow-up period may be insufficient to detect feedback amplification risk reduction, which may require longer observation horizons to manifest in measurable distributional outcomes. Future research should extend ERAMP to agentic AI platforms -- systems in which AI agents take autonomous actions with real-world consequences -- where feedback and systemic impact risks are particularly acute and current assessment methods are most limited.

6. Conclusion

6.1 Summary

This paper proposed ERAMP, a structured ethical risk assessment methodology for AI-based decision platforms comprising four platform-level risk dimensions -- Aggregation Risk, Feedback Amplification Risk, Value Conflict Risk, and Systemic Impact Risk -- and a 28-item scoring instrument. Application to 18 operational platforms revealed a mean ERAMP score of 54.7/100 with 38.9% classified as Critical. Feedback Amplification Risk was the highest-scoring dimension. Six-month follow-up demonstrated a 48.3% reduction in Critical and High risk factors (Cohen d = 2.48) through ERAMP-guided remediation. ERAMP aligns with EU AI Act Article 9 risk management requirements.

6.2 Recommendations

For AI platform operators, the primary recommendation is to conduct ERAMP assessments at platform design time -- before deployment -- rather than as post-hoc compliance exercises. The Feedback Amplification dimension should be prioritised given its prevalence and the rapid risk reduction achievable through automated monitoring. For platform architects, feedback loop mapping (FAR Step 1) should be incorporated as a mandatory design artefact, enabling prospective identification of feedback-induced bias pathways

before they are instantiated in deployed systems. For regulators, the ERAMP scoring instrument provides a quantitative, reproducible platform-level risk assessment tool that complements model-level bias auditing for EU AI Act conformity assessment of integrated decision platforms. The full ERAMP instrument and assessment protocol are available in Appendix A.

References

Bellamy, R. K. E. et al. (2019). AI Fairness 360. IBM Journal of Research and Development, 63(4-5), 4:1-4:15.

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

Dwork, C. et al. (2006). Calibrating noise to sensitivity in private data analysis. TCC 2006, 265-284.

Dwork, C. et al. (2012). Fairness through awareness. ITCS 2012, 214-226.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., and Venkatasubramanian, S. (2018). Runaway feedback loops in predictive policing. FAccT 2018.

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

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

Friedman, B., Kahn, P. H., and Borning, A. (2019). Value sensitive design and information systems. In Human-Computer Interaction in MIS. M.E. Sharpe.

ISO/IEC 23894 (2023). Information technology -- Artificial intelligence -- Guidance on risk management. ISO.

ISO 31000 (2018). Risk management -- Guidelines. International Organization for Standardization.

Madry, A. et al. (2018). Towards deep learning models resistant to adversarial attacks. ICLR 2018.

NIST (2023). Artificial intelligence risk management framework (AI RMF 1.0). NIST AI 100-1.

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown Publishers.

Pasquale, F. (2015). The black box society. Harvard University Press.

van den Hoven, J., Lokhorst, G. J., and van de Poel, I. (2015). Engineering and the problem of moral overload. Science and Engineering Ethics, 18(1), 143-155.

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

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

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

Sculley, D. et al. (2015). Hidden technical debt in machine learning systems. NeurIPS, 28, 2503-2511.

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

Declarations

Funding

This research was supported by the Swiss National Science Foundation (SNSF) under grant 200021-212341 (ERAMP: Ethical Risk Assessment for AI Decision Platforms) and the Estonian Research Council under grant PRG2218. 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 ERAMP instrument, assessment protocol, anonymised platform assessment data, and follow-up results are available from the corresponding author upon reasonable request.

Ethical Approval

The study involved structured assessment of AI systems by consenting organisations and a survey of consenting professional participants. No personal data of AI-affected individuals were collected by the research team. Ethical approval reference: SIMI-ML-2024-097.

ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100 Vol: 2025, Issues: 2, 2025

Appendix A

ERAMP 28-Item Instrument -- Assessment Guide

The ERAMP instrument comprises 28 items across four dimensions (7 items per dimension), each scored 0-4. The following provides abbreviated item descriptions and scoring guidance.

Aggregation Risk (AR-1 to AR-7)

Feedback Amplification Risk (FAR-1 to FAR-7)

Value Conflict Risk (VCR-1 to VCR-7)

Systemic Impact Risk (SIR-1 to SIR-7)