Towards Trustworthy AI in Critical Systems: From Evaluation Criteria to Metric-Based Risk Assessment

Marco De Santis[1,*,†], Christian Esposito[1,†]

[1]Università degli Studi di Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano SA, Italy

Abstract

The growing adoption of Artificial Intelligence (AI) in high-stakes domains such as healthcare, energy, and mobility demands a shift from traditional accuracy-centered evaluation towards a broader paradigm of trustworthiness. Although several frameworks and standards have been introduced to address ethical, legal, and organizational risks, a consistent methodology for aligning them at the system level is still missing. This paper presents an integrated evaluation framework that bridges the Fundamental Rights Impact Assessment (FRIA), mandated by the European AI Act, with the ISO/IEC 42005 standard for AI risk governance. By decomposing their normative requirements into five evaluation dimensions—fundamental rights, governance, robustness, transparency, and dependability—we develop a metric-alignment matrix that links each obligation to measurable technical indicators. The methodology is validated through a healthcare case study, where a clinical decision- support system is assessed under simulated stress conditions and risk mapping. Results demonstrate that integrating FRIA and ISO/IEC 42005 enables a traceable, auditable, and performance-aware evaluation process. The framework not only enhances accountability and regulatory compliance but also establishes a scalable foundation for trustworthiness assurance in other critical AI domains.

Keywords

Trustworthy AI, High-Risk AI, AI Act, FRIA, ISO/IEC 42005, Metric Alignment, Fairness Metrics, Robustness, Explainability, Data Quality, AI Governance

1. Introduction

Artificial Intelligence systems are increasingly deployed in safety and mission critical domains, from clinical monitoring to power grid management, autonomous mobility, and high risk financial services [1]. While high predictive accuracy is necessary, it is not sufficient to ensure operational resilience, fairness, and transparency under adverse conditions such as operational stress or cyberattacks [2]. For applications that fall under the high risk categories of the EU AI Act [3], accuracy centric development shows structural limits [4] and is inadequate to protect safety, fundamental rights, and stakeholder trust across the lifecycle [5].

To address these limits, the trustworthiness paradigm integrates technical, ethical, and regulatory requirements [5, 6]. It is reflected in reference frameworks such as the EU AI Act [7], the NIST AI Risk Management Framework [8], and the OECD AI Principles [9]. The AI Act introduces binding obligations for high risk systems and requires a Fundamental Rights Impact Assessment that extends the traditional DPIA beyond privacy to rights such as non discrimination, freedom of expression, and security [7, 10]. The NIST RMF provides voluntary guidance for operational risk management and properties such as reliability, robustness, accountability, fairness, transparency, and privacy [8]. The OECD principles offer non binding yet widely adopted recommendations [9]. Despite their complementarity, the field still lacks an operational framework that connects these principles to measurable and verifiable practices [11].

We address this gap by studying the operational convergence between the FRIA required by the AI Act and ISO/IEC 42005, the international standard that structures auditable impact assessment and

QualITA 2025: The Fourth Conference on System and Service Quality, June 25 and 27, 2025, Catania, Italy

*Corresponding author.

†These authors contributed equally.

mdesantis@unisa.it (M. D. Santis); esposito@unisa.it (C. Esposito)

0009-0004-6514-4168 (M. D. Santis); 0000-0002-0085-0748 (C. Esposito)

lifecycle governance for AI systems [12]. We clarify where the two instruments converge and where they diverge, and we introduce a Metric Alignment Matrix that translates FRIA and ISO/IEC 42005 requirements into measurable indicators, including dependability and governance metrics such as MTBF, Uptime, latency under stress, drift reaction time, auditability and traceability, as well as fairness and explainability [13, 14]. The cross reference that binds FRIA fields to ISO/IEC 42005 clauses and to concrete indicators is reported in Table 6 and is used throughout the validation.

For empirical validation we present a healthcare case study. The system and its evaluation are described in Section 4, the FRIA to metrics instantiation is detailed in Section 4.1, and the full FRIA compilation is provided in Appendix 8.1. Operational stress conditions used to exercise the system are described within the case analysis in Section 4, and the observed failure conditions and their implications are discussed in Section 5. This preserves coherence with the present structure of the paper and enables readers to locate each component of the evidence with precision.

Section 2 surveys the regulatory and standards landscape and clarifies the roles of FRIA, ISO/IEC 42005, the NIST AI RMF, and the OECD principles. Section 3 formalizes the integrated framework and introduces the *Metric Alignment Matrix*, including the cross reference that links FRIA fields, ISO/IEC 42005 clauses, and operational indicators reported in Table 6. Section 4 presents the healthcare case study, describes how the FRIA is operationalized, and details the FRIA to metrics instantiation in Section 4.1; within Section 4 we also describe the operational stress conditions considered. Section 5 discusses the observed failure conditions, their implications for rights and service dependability, and the main governance trade offs. The full FRIA compilation is provided in Appendix.

2. Regulatory & Standard Landscape

In recent years, the introduction of technologies based on artificial intelligence (AI) has generated growing social and economic concerns, driving the development of more structured and proactive regulatory frameworks. In response to the complexity of the challenges posed by AI, numerous regulatory frameworks and international standards have been established to ensure that technological applications are safe, trustworthy, and respectful of human rights and fundamental freedoms [15, 16]. Among these, the FRIA, introduced by the European Union AI Act, and the risk assessment in the ISO/IEC 42005 standard represent key complementary tools for integrated impact assessment and AI governance. Apart from the fact that the AI Act is a binding legislative framework while ISO 42005 is a voluntary international standard, they also differ significantly in their approach to risk assessment, especially for high-risk AI applications. The AI Act introduces a legal classification system and the relative risk assessment is top-down and predefined [17]: if a given system falls into a listed category, it is, by law, considered high-risk, and compliance with strict requirements becomes mandatory. On the other hand, ISO 42005 promotes a contextual, bottom-up risk-based approach: there is no predefined classification of AI systems by use case, but the standard provides guidelines for continuously assessing and mitigating risks throughout the AI lifecycle. Risk is evaluated based on its likelihood and severity within a specific operational context, utilizing principles of quality and information security management.

In addition to these two frameworks, we also find that the NIST AI Risk Management Framework (RMF) [18], developed by the US National Institute of Standards and Technology, and the OECD AI Principles [19], adopted in 2019 by OECD member countries and endorsed by the G20, represent two influential, non-binding frameworks designed to promote the responsible development and deployment of artificial intelligence. Still, they differ in scope, structure, and intended use. The NIST RMF is organized around four core functions: Map, Measure, Manage, and Govern, which help organizations identify AI risks in context (Map), assess them qualitatively and quantitatively (Measure), implement controls and mitigations (Manage), and maintain oversight, accountability, and alignment with organizational values and legal requirements (Govern). The RMF is highly operational and technical, with a strong focus on trustworthiness characteristics, including explainability, robustness, fairness, and privacy. It is designed for organizations that design, develop, or utilize AI systems and aim to integrate risk management into

their AI lifecycle. The OECD AI Principles provide high-level, policy-oriented guidance for governments and stakeholders. They define five key value-based principles: 1) AI should benefit people and the planet; 2) AI systems should be fair and inclusive; 3) AI should be transparent and explainable; 4) AI systems should be robust, secure, and safe; and 5) AI actors should be accountable. In addition, the OECD outlines recommendations for national policies to promote trustworthy AI, such as investment in research, fostering a digital ecosystem, and promoting international cooperation.

In essence, the NIST RMF is a practical implementation framework for managing AI risks within organizations, whereas the OECD principles provide normative guidance for shaping ethical and inclusive AI policies at the national and international levels. Together, they represent complementary layers of AI governance: the OECD sets the vision, and NIST offers a path to operationalize it. By comparing them with the EU AI Act, which has introducted mandatory compliance procedures and specific ex ante measures such as the FRIA. In contrast, the NIST Risk Management Framework (RMF) and the OECD guidelines propose a model based on general principles of accountability, fairness, and transparency, but lacking the binding force of European regulations. Although influential, the latter operate primarily within soft law and voluntary guidelines, thus limiting their operational effectiveness in ensuring complete trustworthiness.

The AI Act comprehensively addresses the regulatory and technical aspects of trustworthiness, focusing particularly on safety, human oversight, and the protection of fundamental rights. However, it may overlook broader dimensions such as environmental and social sustainability, identified as critical in the literature. Conversely, the approaches from NIST and OECD encompass broader ethical and social dimensions, but remain less defined regarding operational metrics and specific requirements needed for effective governance.

A focused evaluation at the single AI system level is crucial because it allows the identification of specific risks and impacts, avoiding ambiguity typical of abstract general principles. This targeted approach enables precise measurement and tailored interventions, essential for ensuring effective regulatory compliance and responsible risk management throughout the system's life cycle.

The fragmentation of approaches among the EU, USA, and OECD creates significant operational challenges for economic operators, forced to navigate heterogeneous regulatory frameworks. This increases compliance costs, reduces the effectiveness of governance controls[20], and creates potential conflicts in transnational operations, ultimately weakening overall trust in AI systems and the ability to promote common global standards.

In summary, FRIA anchors the process in fundamental rights and legal obligations, ISO/IEC 42005 provides organizational structure and lifecycle management, NIST AI RMF embeds a systematic, iterative risk-based workflow, and OECD Principles set high-level ethical and policy values. Table 1 shows an integration among them.

Governance Element	**FRIA (Legal/Rights)**	**ISO/IEC 42005 (Organizational)**	**NIST AI RMF (Lifecycle Risk)**	**OECD AI Principles (Ethical/Policy)**
1. Context and Scoping	Define system purpose, legal basis, and rights at stake	Clause 4 - Define scope, context, and stakeholders	**Map** - Understand intended use and environment	Human-centered values, inclusive growth
2. Stakeholder Involvement	Include affected users, legal experts, civil society	Clause 4.2 - Address needs of interested parties	**Map/Manage** - Incorporate stakeholder input	Democratic participation, fairness

3. Risk Identification	Identify impacts on privacy, equality, autonomy	Clause 6 - Risk planning	**Map/Measure** - Identify risks to rights, safety, fairness	Safety, rule of law
4. Risk Assessment	Analyze severity, reversibility, likelihood, scope	Clause 6.1 - Evaluate and prioritize risks	**Measure** - Quantify and assess risks and uncertainty	Accountability, proportionality
5. Data and Model Governance	Assess data representativeness, fairness, legality	Clause 8 - Control over data, training, validation, performance	**Manage** - Implement data quality and integrity controls	Robustness, transparency
6. Mitigation and Oversight	Design safeguards (e.g., human-in-the-loop, bias audits)	Clause 8 - Establish technical and organizational controls	**Manage** - Apply risk mitigations and controls	Human oversight, contestability
7. Documentation and Traceability	Record rights risks, justifications, decisions, mitigations	Clause 7.5 - Maintain documented information	**Govern** - Ensure traceability, auditability	Transparency, explainability
8. Monitoring and Improvement	Periodically review rights impact and system evolution	Clauses 9–10 - Monitor, audit, improve AIMS	**Govern** - Lifecycle-based adaptation to change	Sustainable and adaptive AI
9. Redress and Accountability	Provide complaint mechanisms, appeal rights, legal remedies	Clause 10.2 - Handle nonconformities, feedback loops	**Govern** - Enable redress and incident handling	Access to remedy, fairness
10. Legal and Ethical Alignment	Align with EU Charter, GDPR, anti-discrimination law	Cross-cutting compliance obligations	Cross-cutting governance integration	Respect for law, ethical use, cooperation

Table 1: Integration of the main frameworks for AI governance

Table 1 highlights key areas of alignment between FRIA and ISO/IEC 42005, such as stakeholder inclusion, traceability, and oversight. Yet, their focus diverges: FRIA prioritizes the protection of fundamental rights, while ISO 42005 emphasizes procedural rigor and auditability. Notably, lifecycle monitoring and ISO integration—core to ISO 42005—are absent in FRIA, whereas legal redress mechanisms in FRIA are only marginally addressed in the standard. This asymmetry underscores the need to combine both perspectives to ensure trustworthiness that is both operational and normatively robust.

2.1. FRIA: Objectives, Structure, and Criticalities

The Fundamental Rights Impact Assessment (FRIA), mandated by Article 27 of the AI Act, aims explicitly to proactively evaluate the impacts of high-risk AI systems on fundamental rights, promoting the adoption of solutions that comply with the European regulatory framework and ensuring the protection

of individual and collective freedoms. This process includes identifying affected stakeholders, assessing specific risks, and implementing human oversight and internal governance measures.

FRIA extends evaluation to all fundamental rights, unlike the Data Protection Impact Assessment (DPIA), which primarily focuses on personal data protection. FRIA therefore addresses broader rights, such as non-discrimination, freedom of expression, and equitable access to public services, differentiating itself in the breadth and complexity of topics covered.

A possible template of the FRIA according to the AI Act is available as follows:

1. Project/System Identification

Field	**Details**
Name of the AI system/project	
Description of functionality	
Development/Deployment phase	☐ Design ☐ Development ☐ Testing ☐ Deployment ☐ Post-market
Purpose and goals	
Responsible organization(s)	
Contact point	

2. Legal and Operational Context

Legal basis (e.g. GDPR Art. 6)	
Sector of application	
Target users	
Affected individuals/groups	
Geographic area	
Use of personal data?	☐ Yes ☐ No If yes, specify:
Use of biometric/sensitive data?	☐ Yes ☐ No If yes, specify:
High-risk AI under EU AI Act?	☐ Yes ☐ No ☐ Not sure

3. Fundamental Rights at Stake

Fundamental Right	**Affected?**	**Description of Potential Impact**
Human dignity (Art. 1 EU Charter)	☐ Yes ☐ No	
Privacy and data protection (Art. 7–8)	☐ Yes ☐ No	
Non-discrimination (Art. 21)	☐ Yes ☐ No	
Freedom of expression/information (Art. 11)	☐ Yes ☐ No	
Right to good administration (Art. 41)	☐ Yes ☐ No	
Access to justice/fair trial (Art. 47–48)	☐ Yes ☐ No	
Other (specify)	☐ Yes ☐ No	

4. Risk Assessment

Potential severity	[Low / Medium / High]
Likelihood of impact	[Unlikely / Possible / Likely]
Affected population	
Reversibility	
Cumulative/systemic risks	

5. Mitigations and Safeguards

Mitigation Measures	**Description**
Human oversight	
Transparency and explainability	
Data minimization, pseudonymization	
Bias detection and correction	
User access and appeal mechanisms	
Independent auditing	
Other safeguards	

6. Stakeholder Engagement

Stakeholder Group	**Method of Engagement**	**Feedback and Concerns**
End users		
Civil society		
Vulnerable communities		
Data Protection Officer		
Legal/Human Rights Experts		

7. Final Evaluation and Recommendations

Residual risks (after safeguards)	
Acceptability of risks	☐ Acceptable ☐ Acceptable with conditions ☐ Unacceptable
Recommendations	
Responsible authority's decision	
Monitoring plan	

The EU AI Act does not explicitly define a fixed list of quantitative quality metrics for AI systems. Still, it outlines essential qualitative characteristics and requirements that AI systems—especially those classified as high-risk—must satisfy. However, it is possible to identify quality measures from Titles III and IV, especially Articles 9–15, which are listed in Table 2. Moreover, the law does not prescribe numerical thresholds, but risk-based adequacy is expected: the higher the potential impact on fundamental rights or safety, the stricter the performance and governance obligations.

Quality Dimension	AI Act Requirement	Example Metrics
Accuracy	Article 15: Systems must achieve appropriate accuracy levels	Classification accuracy, precision, recall, F1-score, operational error rate
Robustness and Resilience	Article 15: ensure resilience to errors and manipulation	Accuracy under noise, adversarial robustness, uncertainty intervals, stress testing
Cybersecurity	Article 15: protect against adversarial attacks and tampering	Attack detection rate, patch frequency, model inversion resistance
Explainability and Transparency	Articles 13–14: ensure system outputs are interpretable and documented	SHAP/LIME coverage, model card completeness, percentage of interpretable outputs
Fairness and Non-discrimination	Article 10: ensure training data and outcomes are free from bias	Demographic parity, disparate impact ratio, equalized odds, bias audit reports
Data Quality and Governance	Article 10: data must be relevant, complete, and representative	Missing data rate, label noise, dataset representativeness, preprocessing integrity
Human Oversight	Article 14: human intervention must be possible and meaningful	Human-in-the-loop rate, intervention latency, override availability
Accountability and Auditability	Articles 11–12: ensure traceability and logging of decisions	Audit trail completeness, log coverage, documented decision events, retention period

Table 2: AI quality measures from AI Act

2.2. ISO/IEC 42005: Purpose, Scope, and Process

The ISO/IEC 42005 standard aims to provide a structured and systematic framework for conducting AI System Impact Assessments, promoting transparency, traceability, and detailed reporting of the evaluation process throughout all phases of the AI system life cycle. This standard supports organizations in clearly defining roles, responsibilities, risk management processes, and documentation practices.

ISO/IEC 42005 outlines four main operational phases: identification, analysis, risk assessment, and mitigation and monitoring, with particular attention to integrating existing organizational governance activities and proactively involving relevant stakeholders, thus ensuring a participatory and transparent process.

ISO/IEC 42005 addresses concrete and quantitative operational requirements of trustworthiness, such as documented risk management, traceability, and auditability. This complements FRIA, which primarily focuses on qualitative and normative aspects of trustworthiness based on human rights protection.

Table 3 contains the description of the steps of AI assessment and governance concerning the indications of the ISO 42005 standard.

Phase	Description
1. Establish AI Management System (AIMS)	Define the scope, roles, responsibilities, and policies for AI governance. Ensure top-level commitment, align with organizational values, and adopt ethical principles.
2. Contextual and Risk-Based Analysis	Identify internal/external issues and stakeholder requirements. Conduct risk assessments for AI systems, evaluating severity, likelihood, and systemic impact. Align with other impact assessments (FRIA, DPIA, AIA).
3. Operational Control and Lifecycle Management	Implement governance and quality controls throughout the AI lifecycle, including design, testing, deployment, and updates. Manage risks from third-party tools or datasets.
4. Monitoring and Continuous Improvement	Establish performance indicators, conduct internal audits, monitor for failures or bias, and review management system effectiveness periodically.
5. Documentation and Transparency	Maintain detailed records of all processes and decisions. Ensure traceability of models and datasets. Provide documentation such as model cards or data sheets for transparency.
6. Communication and Stakeholder Engagement	Enable both internal and external communication about AI risks and governance. Engage users, affected communities, regulators, and partners.
7. Integration with Other ISO Systems	Align AI governance with existing systems such as ISO/IEC 27001 (security), ISO 9001 (quality), and ISO 23894 (AI risk management), promoting consistency and efficiency.

Table 3: ISO 42005 steps and AI governance model

By integrating FRIA and ISO/IEC 42005 into a single evaluation framework, it becomes possible to bridge the gap between abstract ethical principles and operational measurability, thus ensuring an optimal balance between respecting fundamental rights and the concrete technical governance of AI systems.

3. Methodological Framework

The proposed framework aims to support the evaluation of high-risk AI systems by integrating regulatory requirements with observable and measurable technical properties. It is structured around five key dimensions: fundamental rights, governance and auditability, statistical robustness, explainable transparency, and operational performance. These dimensions serve as analytical lenses to interpret and assess the trustworthiness of AI systems operating in safety-critical domains.

The methodology follows a three-step structure inspired by the design science paradigm: decomposition and mapping of normative requirements, definition and selection of metrics, and integration of evaluative evidence. This process transforms legal and procedural obligations into criteria that can be assessed using concrete indicators [21].

The first step involves decomposing FRIA and ISO 42005 into their elementary requirements. From FRIA, we extracted five critical elements: contextual risk definition, stakeholder consultation, proportionality assessment, human oversight, and contestability. These requirements were selected because they represent recurring obligations in risk impact assessments and are explicitly reflected in regulatory

proposals such as the AI Act. Each element is assigned to one or more framework dimensions. For example, stakeholder consultation relates to the dimension of fundamental rights, while human oversight aligns with explainable transparency. Similarly, ISO 42005 provides procedural phases that were mapped accordingly: scope definition, impact identification, risk evaluation, mitigation and documentation, and continuous monitoring. These are assigned to governance-related or performance-related dimensions depending on their operational function.

The second step focuses on selecting measurable metrics aligned with each dimension. The choice is based on three criteria: semantic relevance to the associated requirement, objective measurability, and applicability in safety-critical environments. Well-known metrics such as fairness score, mean time between failures (MTBF), uptime ratio, and severity-likelihood matrices are integrated with adapted or newly proposed indicators. These include time-to-unsafe-state, human oversight latency, and traceability index. Each metric is clearly connected to a specific requirement and mapped onto its corresponding evaluative dimension.

Table 4 summarizes this conceptual alignment. Each requirement is linked to a technical metric through an intermediate dimension. This mapping enables an operational interpretation of abstract principles and serves as the backbone for system-level evaluation.

Table 4
Mapping matrix: regulatory requirements, framework dimensions, and associated metrics

Source	Requirement	Framework Dimension	Associated Metric
FRIA	Contextual risk definition	Fundamental rights	Context Sensitivity Index
	Stakeholder consultation	Fundamental rights	Stakeholder Inclusion Index
	Proportionality assessment	Fundamental rights	Risk-Benefit Ratio
	Human oversight	Explainable transparency	Human Oversight Latency
	Contestability	Explainable transparency	Explainability Coverage
ISO 42005	Scope definition	Governance and auditability	Scope Completeness Score
	Impact identification	Governance and auditability	Impact Taxonomy Completeness
	Risk evaluation	Performance and robustness	Severity-Likelihood Risk Matrix
	Mitigation and documentation	Governance and auditability	Mitigation Traceability Index
	Continuous monitoring	Governance and performance	Auditability Score / Drift Reaction Time

This mapping is not only conceptual but also instrumental to implementation. By linking regulatory criteria to evaluative dimensions and technical metrics, the framework provides a repeatable, auditable structure for compliance assessment. The matrix enables its application to real systems, supporting both ex ante evaluation and in-operation monitoring, which will be demonstrated in the following section. In addition, as the ISO 42005 holds a broader focus and model than the AI Act, it is possible to embed the FRIA into the AI Management System (AIMS) required by ISO 42005, as follows in Table 5.

FRIA Element	Mapped ISO/IEC 42005 Requirement	Integration Description
System Scoping and Context Definition	Clause 4: Context of the organization	FRIA defines the AI system's purpose, actors, and affected individuals, aligning with ISO 42005's requirement to define scope, external/internal issues, and stakeholders.
Identification of Affected Rights	Clause 6: Planning & Clause 8: Operation	Rights such as privacy, dignity, and non-discrimination are mapped and assessed, feeding into ISO 42005's planning and operational risk controls.
Risk Assessment (severity, likelihood, reversibility)	Clause 6.1: Actions to address risks and opportunities	FRIA's risk logic integrates with ISO 42005's enterprise risk-based planning, ensuring human rights risks are formally managed.
Safeguards and Mitigation Measures	Clause 8.1: Operational planning and control	Safeguards identified in FRIA (e.g., human oversight, bias audits) become formal control measures in the AI lifecycle under ISO 42005.
Stakeholder Consultation	Clause 4.2 & Clause 9.1: Stakeholder needs and monitoring	FRIA requires dialogue with affected parties; ISO 42005 formalizes this as engagement and feedback mechanisms.
Documentation of FRIA Findings	Clause 7.5: Documented information	FRIA records (impact analysis, decisions, mitigations) become compliance documentation under ISO's required document controls.
Monitoring of Rights Impact	Clause 9.1 and Clause 10.2: Monitoring, Evaluation, Improvement	FRIA's feedback and periodic review fit into ISO's monitoring and continual improvement requirements for AI governance.

Table 5: FRIA embedding into the AI governance framework of ISO 42005

4. Case Study: Application of the Framework to a Clinical AI System

The proposed framework was applied to a high-risk AI system deployed in the healthcare sector, aimed at supporting clinicians in identifying and managing patient deterioration risks associated with chronic conditions. The system integrates into hospital workflows and assists in tasks such as emergency triage, individualized monitoring, and early warning of clinical decompensation. It includes a supervised learning module trained on clinical variables, an interface for displaying recommendations to clinicians, and a component for logging decisions and system outputs.

Risk in this system emerges from its dual function: providing autonomous prioritization scores and issuing clinical warnings that influence time-sensitive decisions. The potential consequences of misclassification, delayed human oversight, or model degradation make it a relevant test case for evaluating both compliance and operational trustworthiness.

The first step consisted of mapping regulatory requirements from the framework to the architectural

and procedural elements of the system. FRIA requirements such as contextual risk definition were partially addressed through documented patient groups and clinical use cases, although little attention was given to vulnerable populations or social factors. Stakeholder consultation occurred during early phases with medical staff, but lacked mechanisms for patient inclusion. Human oversight was supported through manual validation of AI-generated suggestions, though response times varied considerably. The absence of contestability features was notable: clinicians could ignore recommendations but had no explicit interface to justify or log such overrides.

The ISO 42005 requirements revealed complementary strengths and weaknesses. The definition of scope and functional boundaries was well specified, and system impact was tracked with respect to process efficiency and alert frequency. However, no structured risk evaluation tools such as severity–likelihood matrices were implemented, and mitigation strategies were generic, based on staff fallback behaviors rather than formal protocols. Continuous monitoring was limited to periodic retrospective reviews without live alerts for data drift or performance degradation.

To operationalize the evaluation, the second phase applied the metrics defined in the alignment matrix. The stakeholder inclusion index revealed limited participatory diversity, especially from non-technical actors. The explainability coverage metric showed that while common outputs included traceable rationales, atypical recommendations lacked transparency and were harder for clinicians to interpret. Human oversight latency was measured through simulation of high-load emergency conditions, revealing delays exceeding accepted thresholds for risk-sensitive interventions. Risk-benefit assessment was performed using test scenarios with known clinical outcomes: while the system reduced under-triage in typical cases, it introduced significant uncertainty when input data were incomplete or inconsistent. Robustness metrics showed vulnerability to missing data and distributional shifts. Performance indicators such as MTBF and uptime were within operational norms but highlighted issues in error recovery and adaptation.

The final stage of the evaluation focused on integrating these findings into an actionable risk profile. The framework enabled the identification of specific gaps in governance (lack of contestability), compliance (absence of formalized proportionality assessments), and operational safety (delayed oversight and weak mitigation logic). The metrics supported a traceable audit trail, making it possible to quantify dimensions of risk that are often treated qualitatively.

By applying the framework, the case study demonstrates how abstract regulatory requirements can be translated into measurable properties and connected to real-world operational challenges. This structured approach allowed for a multi-dimensional risk evaluation and supported the definition of targeted improvements, including integration of explainable dashboards, formalization of risk classification logic, and implementation of adaptive monitoring policies. The methodology thus proves effective in evaluating and strengthening AI system trustworthiness in safety-critical healthcare environments.

4.1. From FRIA to Measurable Evidence in the Clinical Pilot

To operationalize evaluation at system level, we instantiated the FRIA template introduced in Section 2.1 and reported in full in Appendix 8.1. The case refers to an anonymized clinical decision-support system (*Healthcare-DSS*) whose privacy-preserving architecture and scope are summarized in Table 7, while the legal and operational setting (special-category health data under GDPR, user groups, and deployment context) is detailed in Table 8. Together, these two tables fix the evaluative perimeter (purpose, actors, and data) within which impacts and controls are assessed.

The FRIA then identifies the fundamental rights at stake and their qualitative exposure (Table 9), highlighting privacy/data protection and non-discrimination as primary concerns, with additional operational ramifications for the right to good administration in clinical triage. Building on this, the risk analysis (Table 10) specifies severity, likelihood, reversibility, and cumulative/systemic risks, thus establishing the baseline against which the effectiveness of safeguards and the adequacy of monitoring will be judged.

Safeguards are specified and grouped in Table 11 (human oversight with explicit overrides, explainability, bias auditing, data minimization, independent auditing, rollback playbooks), while the participatory

dimension of the process—methods of engagement and salient feedback—is captured in Table 12. The closing governance decision, residual risks, and the cadence of monitoring activities are consolidated in Table 13, which also anchors the escalation and remediation pathways used during the pilot.

Cross-reference: FRIA → ISO/IEC 42005 → Metrics. To turn qualitative fields into auditable evidence, we bind each FRIA block to ISO/IEC 42005 loci and to concrete indicators in the Metric Alignment Matrix (Table 6). In particular, *Human Oversight Latency* (HOL) and *Explainability Coverage* (EC) are the operational counterparts of the oversight and transparency safeguards listed in Table 11; *Auditability/Traceability* reflects the documentation and logging expectations implicit in Table 11 and required for traceable decisions; *Drift Reaction Time* (DRT), *MTBF*, and *Uptime* link the risk posture set in Table 10 to the monitoring and improvement loop formalized in Table 13. The cross-reference in Table 6 thus functions as the operational hinge between FRIA fields and lifecycle governance, ensuring that each normative requirement is paired with a measurable indicator and an instrumentation source.

Operational definitions and measurement. To ensure repeatability, we adopt precise semantics for the key indicators: HOL is the median time between rendering an AI recommendation on the clinician dashboard and a recorded clinician action (accept/override/acknowledge); EC is the fraction of recommendations accompanied by a faithful local rationale meeting a predefined completeness criterion; the Auditability/Traceability score aggregates the linkage between model versions, data snapshots, consent events, and override records; DRT measures the elapsed time from drift detection to the first mitigation (threshold adaptation or rollback); MTBF and Uptime are derived from service health telemetry. These indicators instantiate the safeguards in Table 11 and provide quantitative closure to the risks in Table 10 under the monitoring plan of Table 13.

Findings and governance closure. In the pilot, HOL exceeded acceptable thresholds under high-load conditions, EC was partial for atypical recommendations, and dependability indicators remained within operational ranges while revealing weaknesses in error recovery and adaptation. These outcomes validate the usefulness of the FRIA→ISO/IEC 42005→Metrics pipeline: the rights-focused scoping (Table 9) and the structured risk posture (Table 10) are translated into measurable controls (Table 11 via Table 6) and iterated through the monitoring and improvement cycle (Table 13), thereby closing the loop between normative alignment and technical governance.

Table 6: Cross reference from FRIA fields to ISO/IEC 42005 clauses and associated metrics with pilot instrumentation notes.

FRIA field	**ISO/IEC 42005 clause**	**Metric(s) and instrumentation in the pilot**
System identification and context	Clause 4 context and stakeholders	Scope completeness score and impact taxonomy completeness. Extracted from architecture docs, data flow diagrams, and the stakeholder registry.
Fundamental rights at stake	Clause 6 planning and Clause 8 operation	Context sensitivity index and stakeholder inclusion index. Populated from FRIA fields and engagement traces in Table 12.
Risk assessment severity likelihood reversibility systemic	Clause 6.1 actions to address risks	Severity likelihood matrix and time to unsafe state. Fed by scenario analysis and escalation pathways aligned with Table 10.

FRIA field	ISO/IEC 42005 clause	Metric(s) and instrumentation in the pilot
Mitigations and safeguards oversight explainability bias minimization	Clause 8.1 operational planning and control	Human Oversight Latency and Explainability Coverage and mitigation traceability index. HOL from dashboard telemetry, EC from explanation logs, traceability from links between safeguards and model versions as in Table 11.
Data and model governance	Clause 8 data training validation performance	Missing data rate and representativeness index and data lineage completeness. Computed over HL7/FHIR views and federated learning rounds.
Documentation and traceability	Clause 7.5 documented information	Auditability and traceability score. Completeness of model and data cards, access logs, consent artifacts, and override records.
Stakeholder engagement	Clause 4.2 interested parties and Clause 9.1 monitoring and feedback	Stakeholder inclusion index. Coverage and balance of clinicians, patients, and vulnerable groups derived from activities in Table 12.
Final evaluation and monitoring plan	Clauses 9 to 10 monitoring improvement nonconformities	Drift Reaction Time and MTBF and Uptime. DRT from detection to mitigation timestamps, MTBF and Uptime from service health probes, escalation and rollback playbooks as in Table 13.

5. Discussion

This study shows that aligning a Fundamental Rights Impact Assessment with ISO/IEC 42005 becomes operational once each qualitative requirement is paired with a measurable indicator and a concrete source of evidence. The case under analysis is an anonymized clinical decision support system whose scope and architectural perimeter are established in Table 7, and whose legal and operational setting is detailed in Table 8. This scoping step constrains the rights and risk taxonomy considered in the assessment and anchors the provenance, logging, and contestability assumptions that later inform measurement.

A first result concerns the ability of the assessment to surface rights centric exposures that remain invisible with accuracy only testing. As summarized in Table 9, privacy and data protection and non discrimination emerge as primary concerns, with an operational corollary on the right to good administration whenever triage or resource allocation may be affected by degraded or drifting models. The subsequent risk posture in Table 10 refines this picture along severity, likelihood, reversibility, and systemic accumulation, thereby providing the baseline against which safeguards and monitoring must be judged. In this sense, the taxonomy complements the governance decomposition presented in Table 1 and the quality dimensions recognized in Table 2, while ISO/IEC 42005 offers the organizational locus for integration as summarized in Table 3.

A second result pertains to the conversion of safeguards into measurable controls. The set of mitigations in Table 11 consists of human oversight with explicit overrides, explainability, bias auditing, minimization through Solid PODs, independent auditing, and rollback playbooks. These safeguards map to ISO/IEC 42005 clauses and to the indicators of the Metric Alignment Matrix through the cross reference in Table 6. Human Oversight Latency quantifies the time from recommendation display to

clinician action and operationalizes the effectiveness of the oversight workflow. Explainability Coverage captures the availability of faithful local rationales for individual recommendations. The Auditability and Traceability score aggregates documentation, consent events, access logs, model and data cards, and override records. Drift Reaction Time, together with MTBF and Uptime, implements the monitoring and improvement logic required for sustained governance. In this way, Table 6 functions as the hinge that turns qualitative assessment into audit ready evidence within an ISO/IEC 42005 management system.

The pilot yields three observations that link assessment to measurement. Human Oversight Latency exceeded acceptable thresholds under high load conditions, which shows that even when decision quality is preserved, timeliness can fail and thereby impair the practical enjoyment of rights in time sensitive clinical settings. Explainability Coverage was high for common patterns yet partial for atypical recommendations, which limits contestability precisely where transparency is most valuable. Dependability indicators remained within operational ranges while revealing weaknesses in error recovery and adaptation under missing or delayed data. These outcomes link the risk posture in Table 10 to the oversight and documentation safeguards summarized in Table 11 and motivate the escalation and remediation pathways consolidated in Table 13.

The observations reveal two families of trade offs that must be managed as first class design constraints. Auditability and latency interact because enriching logs and documentation introduces overhead that can deteriorate Human Oversight Latency if left unmanaged. Transparency and performance also interact because increasing explanatory fidelity or coverage can affect throughput and, depending on technique, numerical performance, which in turn influences Drift Reaction Time and dependability. The point is not to scale back safeguards, but to budget them with explicit service level objectives that are monitored continuously under the cadence defined in Table 13.

The stakeholder perspective documented in Table 12 clarifies that measurement must be embedded in existing clinical workflows to remain credible. Clinicians requested predictable oversight latency and clearer rationales. Patients and civil society emphasized transparency of data permissions and access logs. Governance actors, including the Data Protection Officer and legal experts, emphasized consent alignment and redress. These inputs shape thresholds and escalation criteria for Human Oversight Latency, Explainability Coverage, and Drift Reaction Time, and they guide the documentation artifacts that feed the Auditability and Traceability score. In this way the participatory dimension advocated by the assessment and the monitoring and improvement loop of ISO/IEC 42005 reinforce each other.

Limitations arise from the scope of the pilot and from the instrumentation strategy, and they delineate the validity envelope of the results. The evidence pertains to an anonymized diabetes onset risk scenario, hence generalization to other pathologies or to decision contexts beyond prevention requires caution. Workload stress and missing data patterns were simulated under controlled conditions, which are representative of operational strain but cannot reproduce all edge cases. The measurement pipeline itself can influence latency and throughput, therefore it must be periodically calibrated to avoid confounding effects. These limits do not undermine the main claim and instead inform the monitoring cadence reported in Table 13.

Two practical implications follow directly. The combined use of the assessment and ISO/IEC 42005 translates ethical and legal requirements into a measurable governance routine where rights at stake in Table 9 become concrete indicators in Table 6 that are continuously verified within a documented management system. The architectural decisions reported in Table 7 and Table 8, namely Solid PODs with granular consent, HL7/FHIR views, Federated Learning, and non persistent dashboards, facilitate measurement by design, lower the cost of auditability, and enable fine grained rollback and redress without central duplication of sensitive data.

In sum, the results indicate that the alignment between the assessment and ISO/IEC 42005, instantiated through Table 6, offers a credible path for transforming high level trustworthiness requirements into verifiable properties of AI systems that operate in safety and mission critical domains.

6. Conclusion

The framework presented in this work offers a structured approach to evaluating AI systems in safety-critical domains by aligning fundamental rights impact assessment with risk governance principles through measurable indicators. While the conceptual structure is domain-agnostic and can be extended to sectors such as energy, transportation, or finance, its practical implementation revealed areas that require further methodological refinement.

A key direction for future work concerns the formalization of metric selection and scoring criteria. Currently, the mapping from regulatory requirements to technical indicators relies on expert interpretation and contextual analysis. Standardizing this process—possibly through templates, decision trees, or machine-readable rule sets—would enhance repeatability and reduce subjectivity, particularly in highly regulated environments.

Another open challenge lies in the automation of evidence collection and audit generation. Many of the proposed metrics (e.g., oversight latency, contestability traces, explainability coverage) can, in principle, be derived from system logs or interaction data. Developing lightweight tools or APIs to extract and aggregate such evidence in real time would significantly lower the cost and friction of compliance assessment, especially in dynamic or adaptive systems.

Furthermore, integrating the framework into organizational assurance ecosystems remains a priority. Alignment with ISO 9001, ISO 27001, and internal quality audit processes could ensure that trustworthiness assessment becomes a continuous practice rather than an isolated certification step. This also requires defining interfaces with risk owners, data protection officers, and technical leads to coordinate operational responsibilities.

Finally, future iterations of the framework should address dimensions that are currently underexplored, such as environmental sustainability, model lifecycle management, and long-term social impact. Incorporating these perspectives will be essential for building AI governance tools that are not only robust and transparent, but also accountable across temporal, organizational, and ecological scales.

By addressing these challenges, the framework can evolve into a comprehensive and operational tool for evaluating and improving AI systems deployed in high-stakes, regulated domains.

Last, the quality of an AI model is not only related to the design and implementation of the model itself, but also the data provided during the training [22]. Our work will also evolve in this direction, as the community is devoting efforts in defining data quality metrics.

7. Acknowledgments

This work was partially supported by the project "DHEAL – COM- Digital Health Solutions in Community Medicine" under the Innovative Health Ecosystem (PNC) - National Recovery and Resilience Plan (NRRP) program funded by the Italian Ministry of Health and DiabeCo.

8. Appendices

8.1. FRIA — Healthcare Case Study

This appendix reports the compilation of the Fundamental Rights Impact Assessment following the template presented in Section 2.1 and aligned with the integrated framework discussed in the main text. The case study refers to an anonymized clinical decision support system, hereafter *Healthcare-DSS*. The system adopts a patient centric architecture in which raw records remain under individual control in personal data pods, clinical information is exposed through standard interoperable views, learning proceeds in a decentralized manner by exchanging model parameters rather than raw data, and a patient specific digital twin provides risk trajectories and counterfactual simulations at the edge. Prior technical descriptions consistent with this architectural profile motivate these choices and provide additional implementation details [23, 24].

Scope, actors, data classes, and deployment phase are summarized in Table 7, while the legal and operational context is detailed in Table 8. Within this perimeter, the assessment identifies the fundamental rights primarily engaged in Table 9. Privacy and data protection and non discrimination emerge as primary concerns, with operational implications for the right to good administration when triage or allocation decisions may be influenced by model degradation or drift. The risk posture is articulated in Table 10 along severity, likelihood, reversibility, and cumulative or systemic accumulation, and it provides the baseline against which safeguards and monitoring are judged.

Safeguards are grouped in Table 11. Human oversight with explicit overrides is enforced within the clinician workflow. Explainability is provided through local rationales attached to recommendations. Bias detection and correction rely on periodic audits and drift monitoring across federated rounds. Data minimization is supported by the architectural choice to keep raw records in personal pods and to expose only minimal task specific views. Independent auditing and controlled rollback complete the set of measures. Stakeholder engagement is reported in Table 12 and includes clinicians through workshops and dashboard walkthroughs, patients and civil society through in app feedback and consent management, and governance actors through consultations with the data protection officer and legal experts. Residual risks, the acceptability decision, and the monitoring cadence are consolidated in Table 13.

The assessment is connected to measurable evidence through a cross reference between FRIA fields, ISO or IEC 42005 loci, and concrete indicators, as reported in Table 6. Human Oversight Latency and Explainability Coverage instantiate the oversight and transparency safeguards in Table 11. An Auditability and Traceability score aggregates documentation artifacts such as model and data cards, consent and access logs, and override records. Drift Reaction Time links detection to mitigation and reflects the adaptation capacity of the service. Dependability indicators such as MTBF and Uptime are collected through health telemetry. These indicators are sourced from the system itself with minimal friction because the architectural decisions that enable privacy and minimization also enable traceability and measurement by design.

In operation, consented ingestion feeds the personal pods, minimal views are generated for analytics, and federated rounds distribute the current global model, execute local updates, and aggregate parameter deltas without exporting identifiable records. The digital twin consumes the global model and the individual's local state to produce patient specific risk summaries and what if simulations. This workflow has been discussed in the technical literature together with timing characteristics for local epochs, aggregation steps, and data access patterns, and it informs the governance assumptions recorded in the present assessment [23, 24]. The same literature describes the controlled use of synthetic data for augmentation and stress testing. In this case study synthetic profiles are generated and screened for statistical similarity and privacy before limited use, which improves exposure to rare conditions and supports fairness oriented evaluation without altering the fundamental privacy posture of federated training [24].

Evidence collection aligns with the safeguards. Dashboard telemetry supports Human Oversight Latency, explanation logs support Explainability Coverage, consent and access logs and versioning records feed the Auditability and Traceability score, detection to mitigation timestamps determine Drift Reaction Time, and service probes report MTBF and Uptime. Thresholds and escalation paths follow the risk posture in Table 10 and are reviewed under the monitoring plan in Table 13. In this way the qualitative register of the assessment becomes auditable evidence that can be tracked over time within a management system conformant with ISO or IEC 42005, and the architectural choices that protect fundamental rights also lower the cost of sustained governance and verification over the system's lifecycle.

Table 7: FRIA — 1. Project/System Identification (Healthcare-DSS)

Field	**Details**
Name of the AI system/project	*Healthcare-DSS*

Field	Details
Description of functionality	Clinical decision-support system to estimate the risk of diabetes onset in currently healthy individuals. Data are stored in patient-controlled Solid PODs using HL7/FHIR; model training follows a Federated Learning paradigm; a Digital Twin module enables "what-if" simulations for patient-specific scenarios.
Development/Deployment phase	☐ Design ☐ Development ☐ Testing ■ Deployment (pilot) ☐ Post-market
Purpose and goals	Early risk identification to support prevention and clinician triage; strong privacy, traceability, and auditability by design.
Responsible organization(s)	Healthcare provider and research unit
Contact point	Internal AI governance committee

Table 8: FRIA — 2. Legal and Operational Context

Field	Details
Legal basis (e.g. GDPR Art. 6)	GDPR Art. 6(1)(e)/(f) and Art. 9(2)(h) for health data processing in healthcare delivery and public interest.
Sector of application	Healthcare (risk prevention and monitoring)
Target users	Clinicians (web app/dashboard) and patients (mobile app)
Affected individuals/groups	Patients enrolled in the pilot, including potentially vulnerable groups
Geographic area	Pilot sites within the provider's facilities
Use of personal data?	■ Yes ☐ No If yes, specify: health vitals, laboratory parameters (e.g., glucose, HbA1c, HDL/LDL), lifestyle and behavioral data
Use of biometric/sensitive data?	■ Yes ☐ No If yes, specify: special-category health data and biometric signals
High-risk AI under EU AI Act?	■ Yes ☐ No ☐ Not sure

Table 9: FRIA — 3. Fundamental Rights at Stake

Fundamental Right	Affected?	Description of Potential Impact
Human dignity (Art. 1 EU Charter)	Yes	Risk of depersonalization if automated scores override individualized clinical judgment.
Privacy and data protection (Art. 7–8)	Yes	Processing of sensitive health data in patient-controlled PODs; access strictly consented and revocable.
Non-discrimination (Art. 21)	Yes	Potential model bias due to under-represented cohorts or skewed input distributions.
Freedom of expression/information (Art. 11)	No	Not central; relates to the ability of clinicians to record dissent/overrides.
Right to good administration (Art. 41)	Yes	Risk of suboptimal resource allocation if risk scores degrade or drift.

Fundamental Right	Affected?	Description of Potential Impact
Access to justice/fair trial (Art. 47–48)	No	Not directly engaged in the clinical pilot.
Other (Right to health, Art. 35)	Yes	Core interest: quality and timeliness of care depend on reliable recommendations.

Table 10: FRIA — 4. Risk Assessment

Field	Details
Potential severity	High (misestimation may delay preventive interventions or misguide triage)
Likelihood of impact	Possible (data drift, missing or late data, operational overload)
Affected population	Patients in the pilot, including vulnerable subgroups
Reversibility	Limited for time sensitive adverse events
Cumulative or systemic risks	Systemic bias from cohort imbalance; governance risks if logging or traceability are incomplete.

Table 11: FRIA — 5. Mitigations and Safeguards

Mitigation Measures	Description
Human oversight	Clinician validation before acting on AI suggestions; explicit override with justification to ensure contestability and accountability.
Transparency and explainability	Rationale for scores exposed to end users; extension of explainability coverage to atypical recommendations; use of model or data cards for documentation.
Data minimization, pseudonymization	Storage in Solid PODs with no central duplication; granular and revocable permissions; HL7/FHIR standardization for interoperable minimal data views.
Bias detection and correction	Periodic fairness audits and drift monitoring across the federated lifecycle; targeted dataset enrichment if disparities emerge.
User access and appeal mechanisms	Structured contestation and override log in the clinician workflow; in app feedback channels for patients and staff.
Independent auditing	Internal quarterly reviews and semi annual external audits on metrics, logs, and governance controls.
Other safeguards	Controlled model rollback in case of degradations; stress testing under load and incomplete data.

Table 12: FRIA — 6. Stakeholder Engagement

Stakeholder Group	**Method of Engagement**	**Feedback and Concerns**
End users (clinicians)	Workshops and dashboard walkthroughs	Need for clearer explanations and predictable oversight latency under high workload.
Civil society or patients	In app consent and feedback mechanisms	Desire for transparency on data permissions and access logs.
Vulnerable communities	Targeted focus groups (planned)	Representation and fairness monitoring to be reinforced.
Data Protection Officer	GDPR by design consultation	Alignment between consents, logging and documentation retention.
Legal or Human Rights Experts	Advisory input (planned)	Strengthen contestability and redress pathways.

Table 13: FRIA — 7. Final Evaluation and Recommendations

Field	**Details**
Residual risks (after safeguards)	Oversight latency under stress; partial explainability coverage for atypical recommendations; contestability initially limited; patient and community engagement to be broadened.
Acceptability of risks	☐ Acceptable ■ Acceptable with conditions ☐ Unacceptable
Recommendations	Introduce structured override and appeal logging; deploy adaptive drift and fairness monitoring; define thresholds and playbooks for model rollback; extend stakeholder engagement beyond clinical staff.
Responsible authority's decision	Deployment in pilot settings with conditional safeguards and periodic reviews.
Monitoring plan	Continuous monitoring dashboard; quarterly internal reviews; semi annual external audits; alerts on performance degradation and drift.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, P. Barnes, Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020. doi:10.1145/3351095.3372873.

[2] B. Goodman, S. Flaxman, European union regulations on algorithmic decision-making and a "right to explanation", AI Magazine 38 (2017) 50–57.

[3] E. A. I. Act, The eu artificial intelligence act, European Union (2024).
[4] F. Sovrano, S. Sapienza, M. Palmirani, F. Vitali, Metrics, explainability and the european ai act proposal, J 5 (2022) 126–138. URL: https://www.mdpi.com/2571-8800/5/1/10. doi:10.3390/j5010010.
[5] High-Level Expert Group on Artificial Intelligence, European Commission, Ethics Guidelines for Trustworthy AI, Technical Report, European Commission, 2019. URL: https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1.
[6] A. Awadid, K. Amokrane-Ferka, H. Sohier, J. Mattioli, F. Adjed, M. Gonzalez, S. Khalfaoui, Ai systems trustworthiness assessment: State of the art, in: Proceedings of the 12th International Conference on Model-Based Software and Systems Engineering (MBSE–AI Integration), SciTePress, 2024, pp. 322–333.
[7] European Union, Regulation (EU) 2023/xxxxx of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI Act), Official Journal of the European Union, 2023.
[8] National Institute of Standards and Technology (NIST), Artificial intelligence risk management framework (ai rmf 1.0), NIST Trustworthy and Responsible AI, 2023.
[9] Organisation for Economic Co-operation and Development (OECD), Recommendation of the council on artificial intelligence (oecd ai principles), OECD Legal Instruments, 2019.
[10] European Data Protection Board, Opinion 06/2022 on the interplay between the ai act and gdpr: Fundamental rights impact assessment (fria) vs data protection impact assessment (dpia), EDPB, 2022.
[11] A. Jobin, M. Ienca, E. Vayena, The global landscape of ai ethics guidelines, Nature Machine Intelligence 1 (2019) 389–399. doi:10.1038/s42256-019-0088-2.
[12] International Organization for Standardization (ISO), ISO/IEC 42005:2024—Risk Management for Artificial Intelligence, ISO, 2024.
[13] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, in: Proceedings of the 2017 Workshop on Interpretable Machine Learning, 2017.
[14] A. Soumaya, B. Mohammed, Peeking inside the black-box: A survey on explainable ai, IEEE Access 8 (2020) 112067–112096.
[15] A. Mantelero, The fundamental rights impact assessment (fria) in the ai act: Roots, legal obligations and key elements for a model template, Computer Law & Security Review 54 (2024) 106020. doi:10.1016/j.clsr.2024.106020.
[16] I. O. for Standardization, ISO/IEC 42005:2025 Information Technology — Artificial Intelligence — AI System Impact Assessment, ISO, Geneva, Switzerland, 2025.
[17] L. D. Urquhart, G. McGarry, A. Crabtree, Legal provocations for hci in the design and development of trustworthy autonomous systems, in: Nordic Human-Computer Interaction Conference, 2022, pp. 1–12.
[18] E. Tabassi, Artificial intelligence risk management framework (ai rmf 1.0) (2023).
[19] K. Yeung, Recommendation of the council on artificial intelligence (oecd), International legal materials 59 (2020) 27–34.
[20] V. L. Birchfield, From roadmap to regulation: will there be a transatlantic approach to governing artificial intelligence?, Journal of European Integration 46 (2024) 1053–1071.
[21] G. Malgieri, F. Pasquale, Licensing high-risk artificial intelligence: Toward ex ante justification for a disruptive technology, Computer Law & Security Review 52 (2024) 105899. doi:10.1016/j.clsr.2023.105899.
[22] S. Mohammed, L. Budach, M. Feuerpfeil, N. Ihde, A. Nathansen, N. Noack, H. Patzlaff, F. Naumann, H. Harmouch, The effects of data quality on machine learning performance on tabular data, Information Systems 132 (2025) 102549. URL: https://www.sciencedirect.com/science/article/pii/S0306437925000341. doi:https://doi.org/10.1016/j.is.2025.102549.
[23] L. Fortino, S. D. Vita, M. Esposito, C. Esposito, Diabeco: A decentralized federated learning architecture for diabetes care with personal data pods and digital twin, 2025. Preprint; technical description of the DiaBeCo architecture.
[24] L. Fortino, S. D. Vita, M. Esposito, C. Esposito, Federated digital twin architecture with synthetic data generation for privacy-preserving diabetes management, in: IEEE Workshop on ICT and AI for Digital Twins in Healthcare, IEEE, 2025.