

Article

# Risk-Based AI Assurance Framework

Aoun E. Muhammad \* and Kin-Choong Yow 

Faculty of Engineering and Applied Science, University of Regina, Regina, SK S4S 0A2, Canada

\* Correspondence: amo874@uregina.ca

## Abstract

The aim of this research is to present a risk-based AI assurance framework that produces quantifiable metrics for auditors and stakeholders to make deployment decisions with evidence-driven assurance of traceability, explainability, accountability, and reproducibility. Our proposed framework incorporates risk severity core with additional modifiers to accommodate the context, governance obligations, technical and environmental exposure, and residual risk relevant to the AI model. This multi-tiered technique enables stakeholders and governance teams to operationalize the safe deployment assurance. The final Assurance Adequacy Score (AAS) comprises a Governance Readiness Score (GRS) along with two additional indices to quantify the traceability and explainability of the AI model. The Traceability Adequacy Index (TAI) is calculated by evaluating the attributes such as the dataset and model versioning, pipeline logging, model audit completeness, and reproducibility. And an Explainability Adequacy Index (EAI) is calculated by evaluating the attributes such as the fidelity for local and global explanations, stability, faithfulness of the explanation provided, robustness, coverage, and human comprehension. This architecture enables integration of risk assessment and enables continued AI assurance by deploying a bottleneck principle where the readiness of the AI model is confined by the weaker of the indices. Finally, a tiered gate mechanism is applied on the Assurance Adequacy Score to enforce minimum assurance floors for high-risk AI systems. The evaluation conducted on multi-domain AI models demonstrates the Risk-Based AI Assurance Framework's (RBAAF) ability to yield stable and consistent readiness decisions with sensitivity analysis and re-scoring. The use cases demonstrate that even comparable risk levels can lead to significantly different deployment outcomes depending on assurance maturity, and design-specific improvements in traceable or explainable domains have the ability to shift gate outcomes. Combining governance regulations with a standardized and quantifiable traceability and explainability score enables the stakeholders to evaluate the AI system for an accountable and regulation-compliant deployment.

**Keywords:** AI risk assessment; risk scoring; risk tier classification; AI governance frameworks; risk quantification



Academic Editor: Tieling Zhang

Received: 6 January 2026

Revised: 27 February 2026

Accepted: 3 March 2026

Published: 5 March 2026

**Copyright:** © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

## 1. Introduction

The operationalization of Artificial Intelligence (AI) systems across the mission-critical domains has also resulted in human dependence on the decision-making by these systems for high-risk operations such as diagnosis support, credit approval, and threat detection. And this exponential adoption has also resulted in an ever-increasing number of real-world failures. The AI Incident Database (AIID) is a repository that documents failures such as AI systems vulnerable to implicit bias, privacy violations, and adversarial attacks [1].

The pinpoint accuracy of AI systems is not the only criterion for AI deployment. Other factors that contribute heavily to responsible AI deployment are the AI systems ability to be explainable, reproducible, and traceable in their decision-making to justify their deployment in mission-critical environments.

The regulatory and governance frameworks such as the European Union Artificial Intelligence Act (EU AI Act) [2], the National Institute of Standards and Technology (NIST) AI Risk Management Framework (AI RMF) [3], ISO/IEC 42001 (Artificial intelligence management system (AIMS)—Requirements) [4], Organization for Economic Co-operation and Development (OECD) AI Policy Observatory [5] and The United Nations Educational, Scientific and Cultural Organization (UNESCO) issued recommendations on the Ethics of Artificial Intelligence [6] provide high-level direction on the requirements of responsible AI development. All these frameworks and governance regulations collectively emphasize transparency, proper documentation, decision accountability, and human oversight over AI decision-making. While all these extremely important frameworks provide direction and guidance on what responsible deployment should look like, they offer limited directions on how to operationalize consistent risk scoring and severity benchmarking on purpose. On the other hand, cybersecurity has long benefited from quantitative scoring and risk measurement systems such as those presented by The Common Vulnerability Scoring System (CVSS) [7], Factor Analysis of Information Risk (FAIR) [8], and NIST Special Publication (SP) 800-30 Revision 1 (Guide for Conducting Risk Assessments) [9] to assist in risk prioritization, and enable governance reporting.

This study provides a scenario-based demonstration of Risk-Based AI Assurance Framework (RBAAF) to show (i) how the auditable risk and evidence adequacy scores are computed and (ii) how the proportional deployment gating performs across governance contexts, while the use cases are experimental demonstrations and not drawn from a production ready deployment, we have made available all scoring inputs, weights, and intermediate results in the Supplementary File (Supplementary Tables S1–S3) to ensure re-assessment, and future empirical replication in operational settings.

### 1.1. Motivation and Problem Statement

The AI Incident Database (AIID) [1], AI Vulnerability Database (AVID) [10], and MIT Incident Tracker [11] document detailed descriptions of failures and hazards related to AI but do not quantitatively score the AI-related risks and vulnerabilities. The regulatory frameworks, such as the EU AI Act [2] and NIST AI RMF [3], categorize risk by the application of the system but do not quantify technical vulnerabilities to be addressed during engineering design and risk mitigation control selection. The operational gap between governance-driven requirements and technical implementation of AI models results in an ongoing challenge for organizations, i.e., mapping general risk awareness to prioritized mitigation and deployment decisions.

Our proposed framework addresses the aforementioned problems by:

- i. quantifying the AI-related risk by introducing a structured scoring rubric.
- ii. mapping regulatory risk classifications to design specifications controls by scoring technical and operational vulnerabilities.
- iii. Turning high-level governance requirements into measurable operations. Where NIST AI RMF defines what good governance should look like, we provide a way to quantify risk and track it repeatably across teams and over time.
- iv. converting risk awareness into risk prioritization and mapping identified vulnerabilities into a governance risk score.

- v. avoiding static risk treatment by design. The proposed framework is proposed to ensure continuous monitoring and auditing for both pre- and post-deployment assessments, enabling it to detect fragility, drift, and tail risk.

### 1.2. Research Questions

During this study, we formalized the research objectives into the following research questions (RQs):

- RQ1: Can governance-aligned risk severity and evidence maturity be quantitatively linked for traceability and explainability?
- RQ2: Does using assurance as a bottleneck term (via Traceability Adequacy Index: TAI or Explainability Adequacy Index: EAI) prevent the averaging of outcomes for readiness, especially if traceability or explainability evidence is insufficient?
- RQ3: Do tier-driven assurance thresholds yield outcomes that reflect proportional governance expectations?

### 1.3. Related Work

The EU AI Act [2] makes it compulsory for high-risk AI systems to maintain a record of decision-making metrics and to implement controls for continuous monitoring. The NIST AI RMF [3] proposes four pillars of risk mapping, measurement, management, and governance but does not specify how to operationalize risk scoring. ISO/IEC 42001 (Artificial Intelligence management system (AIMS)—Requirements) [4] proposes an AI management system that replicates the logic of its earlier established standards like ISO/IEC 27001 (Information security management systems (ISMS)—Requirements) [12] and ISO 31000 (Risk management—Guidelines) [13], to reinforce accountability as an organizational capability. The OECD principles [5,6] present a more human-centered, rights-aware governance approach but, similar to other frameworks, do not prescribe any scoring methods. The high-level compliance requirements are hard to transform into concrete engineering modules.

A number of research projects related to the effectiveness of the EU AI Act show that compliance becomes easier to achieve when backed by logging, traceability, and auditable controls [14,15]. Researchers [16] have warned about using static risk criteria to evaluate AI systems due to their dynamic nature and ever-changing deployment environment, including drift and emergent behaviors.

The quantitative scoring system is fairly common in the field of cybersecurity. The vulnerability scoring and risk analysis frameworks, such as CVSS [7], FAIR [8], and the NIST 800-30 [9] framework facilitate a common, comprehensible method for severity benchmarking.

On the other hand, current AI risk management frameworks are heavily dependent on taxonomies of vulnerabilities and attacks. And significantly lag in scoring mechanisms to map technical vulnerabilities to severity metrics. Threat modeling frameworks such as STRIDE (Spoofing, Tampering, Repudiation, Information disclosure, Denial of service, Elevation of privilege) [17], MITRE ATLAS (Adversarial Threat Landscape for AI Systems) [18], and OWASP Top 10 for Large Language Model Applications (OWASP Top 10 for LLMs) [19] are helpful in identification of risk sources and implementing respective mitigation techniques. But these threat modeling frameworks inherently lack the capability to enable risk scoring. Other variants of threat modeling frameworks, such as STRIDE-AI (an AI-adapted STRIDE threat modeling approach) [20], transform threat modeling to machine learning (ML) and continuous integration and continuous delivery/deployment (CI/CD) pipelines but remain primarily descriptive. Similarly, blue-team playbooks [21] and reference architectures [22,23] assist in risk mitigation guidance while still leaving risk prioritization an unaddressed problem. The use of static vulnerability assessment

techniques fails to cater to the dynamic nature of complex AI systems [24–26]. This research gap has resulted in AI risk-driven scorecards to explore ways for quantifiable coverage.

Recent research studies related to the explainability and trustworthiness of AI demonstrate the need to tackle the opaque nature of AI models in high-risk environments. The explainability techniques, such as Local Interpretable Model-agnostic Explanations (LIME) [27] and SHapley Additive exPlanations (SHAP) [28], provide local and global attribute relevance and explanations, while Gradient-weighted Class Activation Mapping (Grad-CAM) [29] provides visual explainability for deep networks. However, providing explanations is not equivalent to providing assurance. The explanations can be unstable, unfaithful, or provide unrelated details to the end user [30–32].

Human-centric methods like evaluation criteria, causability in clinical reasoning, and standardized metrics to measure fidelity, faithfulness, and robustness have been proposed to provide related and adequate explanations to the human-in-the-loop (HITL) [33,34]. Following the enforcement of the GDPR, the “right to explanation” has been discussed as a significant legal and governance expectation that encourages the provision of explanations that are understandable and support accountability [35–37]. The level and quality of explainability are evaluated differently in high-risk operations such as credit risk modeling and clinical decision support, where decision-makers require transparency and consistent behavior to do their job [38].

Traceability and reproducibility are the other two attributes that, along with explainability, form the concept of trustworthy AI. The errors in reproducibility occur mainly due to having undocumented hyperparameters, inaccessible data cards, or the absence of any other critical data that may result in irreproducible decision-making under nominally similar conditions [39]. Techniques like model versioning, and data versioning, model lineage, data lineage enabled data versioning, lineage, and documented audit trails can be used to rectify these traceability issues [40]. The researchers have proposed standardizing reporting, data, and model cards to ensure structured provenance in high-risk environments [41].

The traceability methods like dataset version control and standardized documentation templates enable operational logging data and model provenance, but these are tool-specific measures and do not offer cross-domain compatibility [40]. The majority of governance frameworks, explainability methods, and traceability and reproducibility practices lack a scored and auditable approach that can link (i) governance-aligned risk severity, (ii) traceability and reproducibility maturity, and (iii) quality of explanations’ provided into deployability decisions [42–44]. Additionally, the privacy and security governance is addressed via risk, control, and maturity frameworks such as NIST SP 800-53, NIST SP 800-30, the NIST Privacy Framework, ISO/IEC 27701, Center for Internet Security (CIS) Controls v8, and OWASP Software Assurance Maturity Model (SAMM) [45–50]. These frameworks provide guidance for threat modeling, control selection, privacy risk management, and secure development processes [45–50]. The Risk-Based AI Assurance Framework (RBAAF) proposed in this paper uses quantitative risk scoring as the foundation with additional integration of traceability and explainability indices as assurance modules to ensure responsible deployment, and is designed to be used alongside these privacy and security frameworks rather than be used as a utility to replace them.

Prior quantitative approaches in cybersecurity, such as CVSS, FAIR, and NIST risk assessment guidance, provide severity benchmarking and prioritization [7–9], while the AI governance frameworks, such as the EU AI Act, NIST AI RMF, ISO/IEC 42001, define process and documentation expectations but do not limit their guidance with any unified auditable score mapping between risk severity and evidence maturity [3,4,51]. RBAAF adapts a novel approach by integrating deployability computation, i.e., it combines a governance-aligned risk score (GRS) with explicit evidence adequacy indices (TAI and

EAI), and enforces tier-dependent thresholds through an auditable gate (Algorithm 1). This approach enables RBAAF to produce reviewable and auditable deployment outcomes rather than descriptive readiness assumptions.

---

**Algorithm 1** RBAAF computation and deployment gate

---

```

1: Inputs: risk inputs  $L, I$  (and curvature  $k$ ); modifier overlays  $C, G, T, E, R$ ; GRS weights
   ( $\alpha, \beta_C, \beta_G, \beta_T, \beta_E, \beta_R$ ); evidence component scores  $\mathbf{z}^{(T)} \in [0, 1]^m, \mathbf{z}^{(E)} \in [0, 1]^n$ ; evi-
   dence weights  $\mathbf{w}^{(T)}, \mathbf{w}^{(E)}$ ; tier thresholds  $\tau_{TAI}(GRS), \tau_{EAI}(GRS)$ .
2: Outputs:  $GRS, TAI, EAI, AAS \in [0, 1]$  and decision  $\in \{\text{Deploy, Conditional deploy,}$ 
   Limited/internal use, Sandbox, Block, Research sandbox only $\}$ .
Require:  $GRS \in [0, 1]$ ; TAI scores; EAI scores; tier thresholds  $\tau_{TAI, EAI}(GRS)$ 
Ensure:  $AAS \in [0, 1]$ 
Ensure: decision  $\in \{\text{Deploy, Conditional deploy, Limited/internal use, Sandbox,}$ 
   Block, Research sandbox only $\}$ 
3: Compute  $TAI$  from weighted traceability evidence.
4: Compute  $EAI$  from weighted explainability evidence.
5:  $t \leftarrow \min(TAI, EAI)$ 
6:  $AAS \leftarrow \sqrt{GRS} \cdot t$ 
7: if  $GRS \geq 0.85$  then ▷ Critical tier
8:    $\tau_{TAI} \leftarrow 0.80, \tau_{EAI} \leftarrow 0.75$ 
9:   if  $TAI \geq \tau_{TAI}$  and  $EAI \geq \tau_{EAI}$  then
10:    return ( $AAS$ , Deploy: continuous monitoring/audits)
11:   else
12:    return ( $AAS$ , Block)
13:   end if
14: else if  $0.70 \leq GRS < 0.85$  then ▷ High tier
15:    $\tau_{TAI} \leftarrow 0.70, \tau_{EAI} \leftarrow 0.70$ 
16:   if  $TAI \geq \tau_{TAI}$  and  $EAI \geq \tau_{EAI}$  then
17:    return ( $AAS$ , Deploy with controls)
18:   else
19:    return ( $AAS$ , Block (fix TAI or EAI))
20:   end if
21: else if  $0.50 \leq GRS < 0.70$  then ▷ Moderate tier
22:    $\tau_{TAI} \leftarrow 0.60, \tau_{EAI} \leftarrow 0.60$ 
23:   if  $TAI \geq \tau_{TAI}$  and  $EAI \geq \tau_{EAI}$  then
24:    return ( $AAS$ , Conditional deploy: pilot/controlled rollout)
25:   else
26:    return ( $AAS$ , Sandbox)
27:   end if
28: else if  $0.30 \leq GRS < 0.50$  then ▷ Low tier
29:    $\tau_{TAI} \leftarrow 0.50, \tau_{EAI} \leftarrow 0.50$ 
30:   if  $TAI \geq \tau_{TAI}$  and  $EAI \geq \tau_{EAI}$  then
31:    return ( $AAS$ , Limited/internal use)
32:   else
33:    return ( $AAS$ , Sandbox)
34:   end if
35: else ▷ Minimal tier
36:   return ( $AAS$ , Research sandbox only)
37: end if

```

---

Evaluation and contextualization overview. We evaluated RBAAF through reproducible end-to-end demonstrations to compute GRS, TAI, EAI, AAS, and the respective gate decisions. All these scores and findings are contextualized with respect to relevant governance standards, frameworks, and technical assurance practices discussed in Section 1.3.

#### 1.4. Contributions

The contributions of this paper are as follows:

- We calculate a Governance Risk Score (GRS) via a utility-transformed likelihood x impact score that accommodates context sensitivity, governance obligations, technical and environmental exposure, and residual risk.
- We propose an Assurance Adequacy Score (AAS) that addresses the traceability, reproducibility and explainability research problems by incorporating a Traceability Adequacy Index (TAI) and an Explainability Index (EAI). Both of the aforementioned indices are calculated by analyzing the quality and documented provenance of verifiable evidence, versioning of data, model and code enabled, replication tests, and explanations provided.
- We propose a final deployability gate that integrates risk scoring with assurance readiness and enforces minimum tier-specific TAI or EAI thresholds that produce auditable outcomes (Deploy, Conditional Deploy, Sandbox/Pilot, Block) whichever is suitable to ensure adherence to governance obligations.

RBAAF is primarily an AI assurance and governance readiness framework that focuses on operationalizing evidence maturity for traceability and explainability, while privacy and security are critical aspects for trustworthy AI deployment, they are treated in this research as adjacent governance domains rather than being the only scoring criterion. RBAAF is intended to be used alongside established privacy engineering and security risk frameworks, and any future extensions may incorporate dedicated privacy and security indices as additional assurance pillars.

Paper organization: The rest of this paper is organized as follows. Section 2 presents the methodology for the Risk-Based AI Assurance Framework (RBAAF) that includes defining and materializing the Governance Risk Score (GRS) the Traceability and Explainability Adequacy Indices. Section 3 presents the evaluation results across representative use cases and analyzes risk-tier behavior under uncertainty together with the resulting gate decisions. Section 4 discusses practical implications, interpretability of the proposed metrics, and how the framework supports governance workflows. Finally, Section 5 concludes the paper and outlines directions for future work. Additionally, the Discussion and Conclusion sections explicitly discuss the findings with respect to the stated RQs along with summarizing limitations and outlining future research directions.

## 2. Methodology

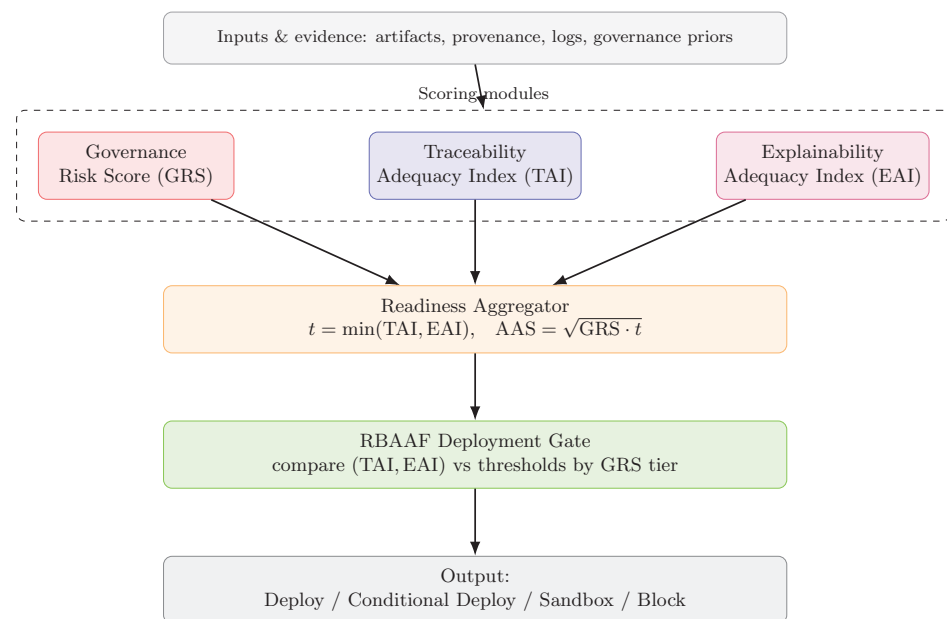
This section presents the end-to-end computation of the Risk-Based AI Assurance Framework (RBAAF) including: (i) the Governance Risk Score (GRS), (ii) the Traceability and Explainability Adequacy Indices (TAI/EAI) that are then used to compute the Assurance Adequacy Score (AAS), and (iii) the tiered deployment gate that maps these scores to actionable outcomes. Figure 1 presents the workflow, and Algorithm 1 shows the decision procedure used throughout the evaluation.

### 2.1. Methodology Overview

RBAAF operationalizes governance-aligned deployment readiness via an auditable pipeline:

- Risk scoring is calculated via a governance-aligned risk score (GRS) from severity and modifier overlays.
- Evidence adequacy is calculated via TAI and EAI from scored, verifiable evidence artifacts.
- Assurance synthesis is determined by combining risk and evidence maturity via  $t = \min(\text{TAI}, \text{EAI})$ , and AAS is computed.

- (iv) Tiered deployment gate is applied to determine the decision outcome, i.e., Deploy/Conditional deploy/Sandbox/Block.



**Figure 1.** RBAAF architecture consisting of three scoring modules (GRS, TAI, and EAI).

The Risk-Based AI Assurance Framework (RBAAF) produces a readiness recommendation that comprises a Governance Risk Score (GRS) and Assurance Adequacy Scores (AAS). This approach enables RBAAF to evaluate and quantify the maturity of traceability and explainability evidence. RBAAF is designed to inherently support and operationalize the governance practices like internal reviews and conformity assessments by providing continuous and discrete outcomes of Deploy, Conditional Deploy, and Block. These outcomes are directly mapped to risk classification concepts as laid out in governance regulations and frameworks [2–4].

RBAAF has three main components:

1. Governance Risk Score (GRS): a composite score bounded by  $[0, 1]$  and is derived from a utility-transformed severity score ( $L \times I$ ) and five governance modifiers.
2. Assurance Adequacy Scores (AAS): that comprises a Traceability Adequacy Index (TAI) and an Explainability Adequacy Index (EAI), each bounded by  $[0, 1]$ .
3. Integration and gating: is the final bottleneck rule where the model readiness is bounded by the weaker of TAI and EAI. At this stage we compare the gate results with the tier-specific thresholds to produce a deployability outcome.

## 2.2. Materials

During this research, five use-case demonstrations are used for evaluation, spanning from cybersecurity intrusion detection and biometric access control to credit risk scoring. The use cases A1 and A2 are L-XAIDS implementation in a lab scenario and Security Operations Center (SOC)-ready deployment, respectively. The use cases B1 and B2 are evaluations of biometric access control with weak and strong assurance evidence, respectively. Lastly, C1 is the demonstration and evaluation of a credit scoring AI model. For each case, we specify (i) GRS inputs (severity and modifier overlays with weights) and (ii) evidence vectors for traceability and explainability used to compute TAI and EAI. All parameter demonstrations, intermediate metrics, and computed outputs are provided in Supplementary Tables S1–S3 to enable reassessment.

### 2.3. RBAAF Framework Overview

Figure 1 illustrates the architecture of RBAAF by showing how the Governance Risk Score (GRS) and the two Assurance Adequacy Scores (TAI, EAI) are combined via a bottleneck gate to produce an auditable deployability recommendation.

### 2.4. Governance Risk Score (GRS)

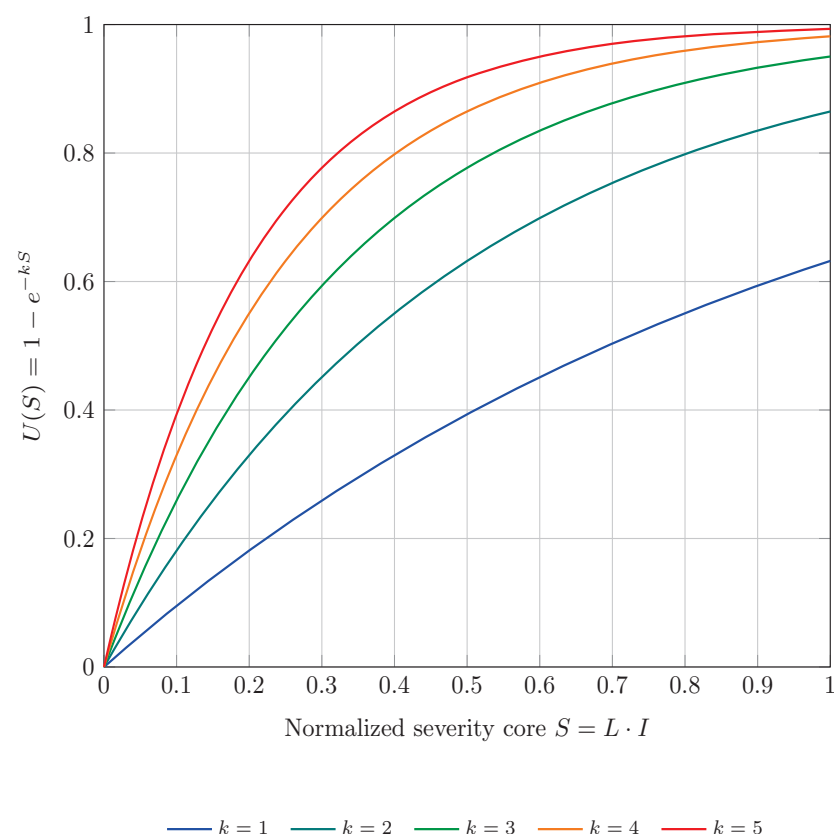
The main sub-components of GRS are the (i) Risk Severity Core and (ii) governance-relevant modifier overlays. Both of these components are discussed below.

#### 2.4.1. Risk Severity Core

Risk severity is calculated as the product of Likelihood ( $L$ ) and Impact ( $I$ ) as per the standard risk analysis practices [9]. Likelihood addresses the probability of a given vulnerability materializing under realistic exposure. And Impact deals with the expected consequence of the vulnerability, if exploited. To reduce the saturation effects and preserve discriminability for higher risks, we apply a utility transform to the  $L \cdot I$ :

$$U(L, I) = 1 - e^{-k(L \cdot I)}, \quad (1)$$

where  $k$  is a calibration constant that controls the curvature and sensitivity at higher  $L \cdot I$  values. Figure 2 shows the behavior of the curvature constant at different values for the utility transformation as defined in Equation (1).



**Figure 2.** Utility transform  $U(S) = 1 - e^{-kS}$  for normalized severity core  $S = L \cdot I \in [0, 1]$  for different values of curvature constants.

#### 2.4.2. Governance-Relevant Modifier Overlays

The dynamic nature of risks in AI model deployment is unique, as it is not only dependent on the baseline risk severity but also on the context in which the AI model is

being used and the governance obligations with respect to the jurisdiction and domain of its application. We therefore incorporate five normalized modifier overlays in GRS:

- Context sensitivity (*C*) modifier incorporates the criticality of the sector AI is being used in, groups affected by the usage, and the safety and rights sensitivity.
- Governance (*G*) modifier deals with the regulatory and organizational controls required to be implemented before the AI model deployment, such as continuous logging, human oversight, conformity assessment, and documentation provenance.
- Technical exposure (*T*) modifier deals with the attack surface related to the AI model under evaluation, accessibility of the system, integration complexity with already deployed systems, and technical exposure if a model is exploited.
- Environmental exposure (*E*) modifier deals with the operational environment uncertainty, the adversarial pressure, and the performance and volatility of the model being deployed.
- Residual risk (*R*) modifier consists of the risk remaining after controls and safeguards have been implemented. This shows the quantifiable risk appetite of the organization deploying or using the AI model, along with controls it has already implemented and risk that is still present.

The Governance Risk Score is computed as a weighted combination of the risk severity core and governance-relevant modifier overlays:

$$GRS = \alpha \cdot U(L, I) + \beta_C C + \beta_G G + \beta_T T + \beta_E E + \beta_R R, \quad (2)$$

where

$$\alpha + \beta_C + \beta_G + \beta_T + \beta_E + \beta_R = 1, \quad GRS \in [0, 1]. \quad (3)$$

The cumulative weights approach is adapted to ensure the organizations can fine-tune their governance posture as per their organizational requirements. For example, a more conservative deployment may emphasize more on *G* and *R*, whereas for a more security-specific deployment, organizations may emphasize more on *T* and *E* [8].

#### 2.4.3. Risk Tiering and Uncertainty Characterization

The GRS values are mapped to risk tiers to support the operational use of threshold-based decisions and to ensure consistent reporting. This risk tiering aligns with the practical requirement to transform continuous risk into actionable governance by implementing additional compensating controls, strict human oversight requirements, or domain restrictions [2,3].

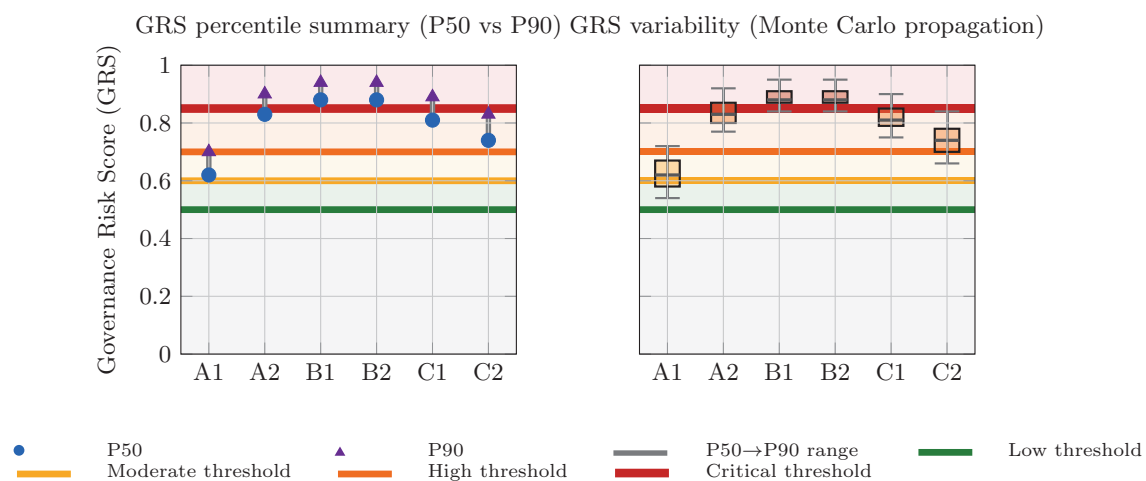
Table 1 shows the tier band definitions used throughout the paper to map risk tiering to assurance thresholds defined in Section 2.6.

The risk scoring for AI is often dependent on incompleteness or absence of evidence, inconsistent documentation, and ever-changing threat landscapes. We employ uncertainty characterization via probabilistic aggregation and Monte Carlo simulation to assess the stability and relevance of risk tier assignments, ranking, and prioritization [52,53]. The uncertainty characterization enables us to score each dimension as a distribution rather than a point estimate. This enables estimation of percentile scores P50 and P90 and volatility. The use of uncertainty characterization provides governance support to the decisions by differentiating between the inherently high-risk vulnerabilities whose exploitative consequences remain robustly under uncertainty and the borderline risks where tier classification is subjective to assumption sensitivity.

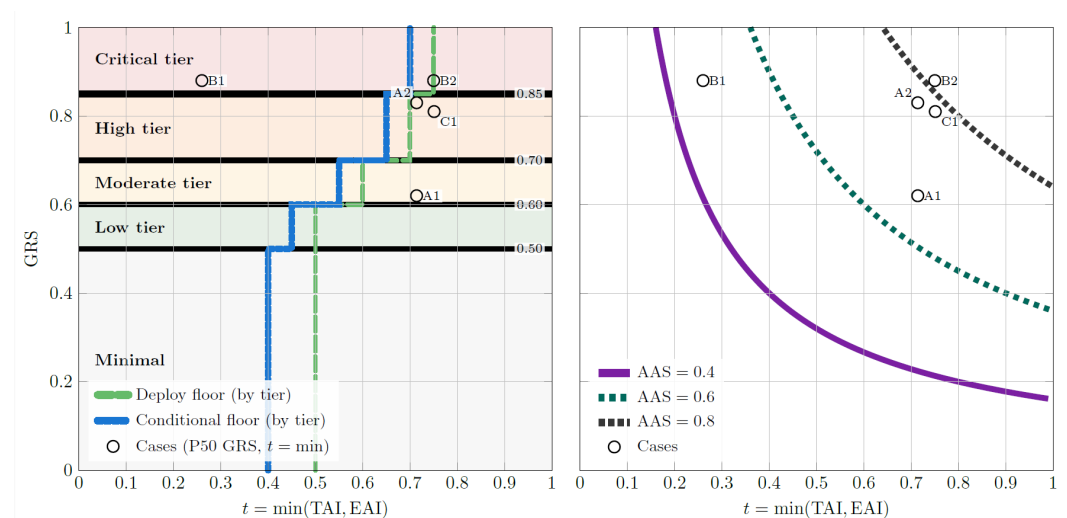
**Table 1.** Governance Risk Score (GRS) tier bands used in RBAAF.

| Tier     | GRS Range   | Governance Posture                                                                                                  |
|----------|-------------|---------------------------------------------------------------------------------------------------------------------|
| Critical | $\geq 0.85$ | Highest oversight and strict evidence requirements with the expectation of continuous monitoring measures in place. |
| High     | 0.70–0.84   | Strong mitigating controls and monitoring should be implemented prior to wide-scale deployment.                     |
| Moderate | 0.50–0.69   | Pilot or controlled roll-out with very explicit remediation plan in place.                                          |
| Low      | 0.30–0.49   | Limited/internal use AI models with baseline controls and risk mitigation defined.                                  |
| Minimal  | $<0.30$     | AI models applicable for research sandbox only.                                                                     |

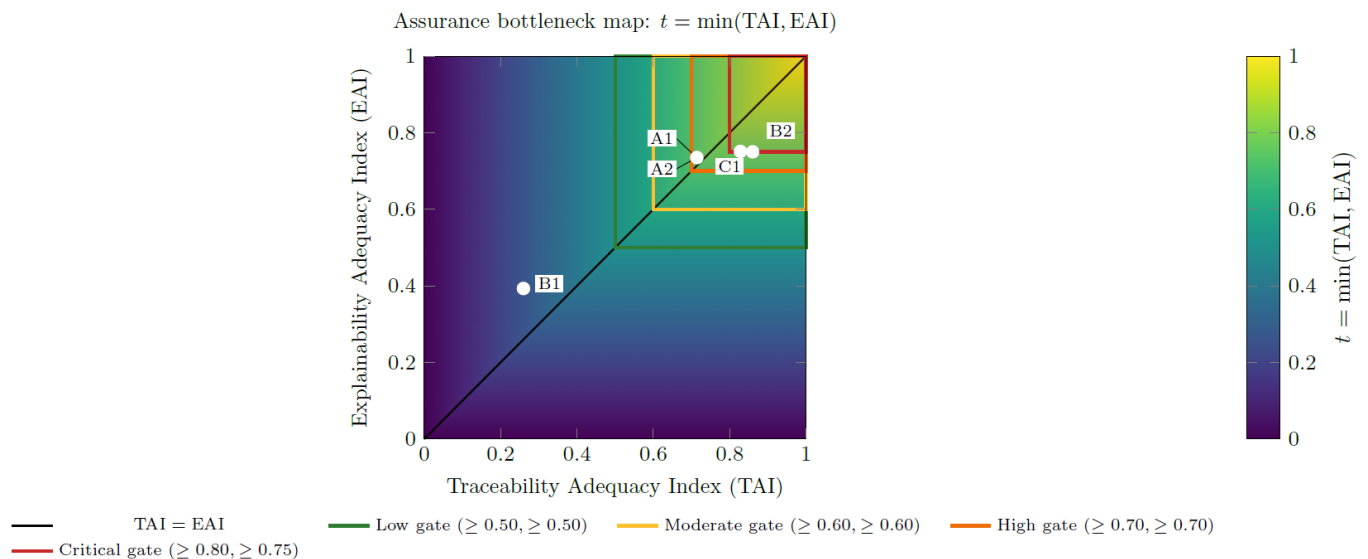
The uncertainty characterization can be summarized as the percentiles for conservative governance, i.e., P50 vs. P90, and a full distributional variability under Monte Carlo simulation. Figure 3 illustrates both these characterizations. The evaluation scenarios depicted in Figures 3–5 follow the five use-case demonstrations defined in Section 2.



**Figure 3.** GRS uncertainty characterization. (Left): percentile summary (P50 vs. P90). (Right): distributional variability from Monte Carlo simulation shown as manual whisker-box summaries (P10/Q1/median/Q3/P90).



**Figure 4.** (Left) GRS tiers and deployment floors (Right) Iso-AAS contours visualizing readiness ranking.



**Figure 5.** Assurance bottleneck map for RBAAF. Color encodes  $t = \min(\text{TAI}, \text{EAI})$ . Tiered deployment gates are shown as nested rectangles.

### 2.5. Assurance Adequacy Scores (AAS): Traceability and Explainability

Assurance Adequacy Scores (AAS) are calculated by evaluating documented artifacts to support auditability, accountability, and independent verification of the evidence [4,39,40]. The two indices AAS relies on are: a Traceability Adequacy Index (TAI) and an Explainability Adequacy Index (EAI). Both of these indices are normalized to  $[0, 1]$  and computed as weighted aggregates of evidence. Default component weights used to compute TAI and EAI are listed in Table 2.

**Table 2.** Default weights for Assurance Adequacy Scores: TAI and EAI.

| Component                            | Default Weight | Justification                                                                                        |
|--------------------------------------|----------------|------------------------------------------------------------------------------------------------------|
| Traceability Adequacy Index (TAI)    |                |                                                                                                      |
| Dataset provenance and versioning    | 0.20           | ability to record, maintain, reconstruct exact training/evaluation data lineage.                     |
| Model and config versioning          | 0.20           | ability to recovers exact model binary and training configuration.                                   |
| Pipeline logging and audit trails    | 0.20           | ability to provide audit-ready reproduction of runs and decisions.                                   |
| Lifecycle trace control points       | 0.20           | ability to ensure end-to-end provenance of decisions and access.                                     |
| Reproducibility replication          | 0.20           | ability to demonstrate successful reruns under similar controlled conditions.                        |
| Explainability Adequacy Index (EAI)  |                |                                                                                                      |
| Fidelity and faithfulness            | 0.20           | Explanations provided must reflect actual model behavior.                                            |
| Stability/robustness                 | 0.15           | Explanations provided should not vary unpredictably under small perturbations.                       |
| Coverage                             | 0.15           | Explanations provided must be available across all relevant outputs and regimes.                     |
| Human comprehensibility and task fit | 0.15           | Explanations provided must support stakeholder decision making, and should have operational benefit. |
| Global consistency (runs/splits)     | 0.15           | Supports policy consistency and governance repeatability.                                            |
| Operational logging of explanations  | 0.20           | Ensures explanations are retrievable and auditable at scale.                                         |

### 2.5.1. Traceability Adequacy Index (TAI)

TAI evaluates, quantifies, and records the end-to-end provenance of data, models, code, decisions, and all other system components by reviewing the available documentation for lifecycle elements listed below:

1. Dataset versioning and provenance—evaluates the system’s ability to protect against tampering or unapproved changes [40,41].
2. Model and configuration versioning—evaluates the system’s ability to ensure models and configurations are not prone to tampering or unapproved changes. This can be accomplished by recording, tracking, and maintaining the model checkpoints, hyperparameters, random seeds, and dependency capture [39].
3. Pipeline logging and audit trails—evaluates the system’s ability to record, track, and maintain the variety of logs (execution, access, decision) for audit-ready reproductions and reruns [2,4].
4. Lifecycle trace control points—evaluates the system’s ability to enforce checkpoints across the lifecycle by implementing logging of change-control events, access authorization, deployment approvals, and rollback triggers.
5. Reproducibility replication—evaluates the ability of the AI model to reproduce results under controlled reruns with the same data and configurations. It also assists in documenting the degree and causes of deviations [39].

Each of these lifecycle elements is evaluated based on the completeness and verifiability of the aforementioned artifacts and documents.

Let  $\mathbf{z}^{(T)} \in [0, 1]^m$  denote the normalized evidence component scores used to compute the traceability adequacy index. In our demonstrations,  $m$  and  $n$  are set as follows:  $m = 5$  for traceability components (Table 2) and  $n = 6$  for explainability components (Table 2).

With non-negative weights  $\mathbf{w}^{(T)}$  and satisfying  $\sum_{j=1}^m w_j^{(T)} = 1$ , we compute:

$$\text{TAI} = \sum_{j=1}^m w_j^{(T)} z_j^{(T)}, \quad \sum_{j=1}^m w_j^{(T)} = 1. \quad (4)$$

### 2.5.2. Explainability Adequacy Index (EAI)

Explainability Adequacy Index (EAI) evaluates, quantifies, and records the quality of the explanations provided and the corresponding stability with respect to model behavior and stakeholder usability. EAI is calculated by evaluating the explanations provided and criteria commonly recommended in the explainable AI literature, as mentioned below:

- Fidelity and faithfulness—deals with the quality of explanations provided, which should consistently align with the actual model behavior [30,34].
- Stability and robustness—deals with the invariance of explanations under small perturbations or resampling of the test data [30,31].
- Coverage—deals with the availability and details of explanations across relevant model outputs, subpopulations, and operating conditions [32,33].
- Human comprehensibility—deals with whether explanations provided support real decision-making operations, including operational interpretations [32].
- Global consistency (runs/splits)—evaluates the explanation behavior and checks if the explanations are consistent for retraining runs, dataset splits, and time windows. Furthermore, it checks if explanations provided are reproducible and consistent with organizational policies.
- Operational logging of explanations—evaluates and tracks whether explanations are generated, stored, versioned, and are retrievable. Information such as model version,

input identifiers, explanation method, and timestamps should be included to enable auditors and stakeholders to evaluate the AI model in depth.

Explanations provided for AI decision-making can either be local [27] or global, model-agnostic or model-specific. Our prior work demonstrated the value explainability adds to the workflow in security context [54]. EAI protects against the manipulation of explainability requirements by just presenting the explanations. The mere presence of explanations is considered insufficient, and there should be verifiable evidence present to ensure that explanations are stable, faithful, and usable.

Let  $\mathbf{z}^{(E)} \in [0, 1]^n$  denote the normalized evidence component scores used to compute explainability adequacy (e.g., fidelity/faithfulness, stability/robustness, coverage, human task-fit). With nonnegative weights  $\mathbf{w}^{(E)}$  satisfying  $\sum_{k=1}^n w_k^{(E)} = 1$ , we compute:

$$\text{EAI} = \sum_{k=1}^n w_k^{(E)} z_k^{(E)}, \quad \sum_{k=1}^n w_k^{(E)} = 1. \quad (5)$$

## 2.6. Deployable Gate Outcomes

RBAAF integrates risk (GRS) and assured adequacy score (AAS) using a bottleneck principle. The deployment readiness is constrained by the weaker of traceability or explainability maturity. We define:

$$t = \min(\text{TAI}, \text{EAI}), \quad (6)$$

and Assurance Adequacy Score is defined as as a risk-weighted readiness measure:

$$\text{AAS} = \sqrt{\text{GRS} \cdot t}. \quad (7)$$

This technique ensures that systems with high risk but weak evidence maturity are scored appropriately and prevents a ‘formal compliance’ decision where nominal documentation is present but is not able to support audit functions, reproduce the decision, or provide a faithful and credible explanation [42,43]. Algorithm 1 provides the summary of the computation of TAI, EAI, AAS, and the tiered gating decision.

## Tiered Gating Thresholds

The deployment decisions are provided by comparing the calculated TAI and EAI with the tiered thresholds that proportionally increase with the governance risk. This pattern reflects the regulatory logic regarding higher-risk systems requiring stricter mitigation or compensating controls, extensive documentation, and constant oversight [2–4]. RBAAF provides the following decision gate outcomes:

- **Deploy**—The deploy outcome is produced when the AI system under evaluation meets the minimum requirements of tier-specific evidence. For Critical tiers, the AI models are allowed to be deployed only when there are controls for continuous monitoring and audits in place. And, for High-risk tier AI models, deployment is allowed only if proper controls are in place. Unconditional deployment is not allowed if the AI model is classified in the Critical- or High-risk tiers.
- **Conditional deploy**—The conditional deployment is allowed for Moderate-risk tier AI systems that meet the evidence requirements. The deployment may be restricted to pilot or controlled rollouts with an explicit remediation plan.
- **Limited or internal use**—The limited or internal use is an outcome for Low-risk tier AI systems that meet the evidence requirements.
- **Sandbox**—The sandbox outcome is for AI systems where evidence requirements are not met for Low-risk tier AI systems or Moderate-risk tier systems. Hence, the system

is confined to sandbox deployment until appropriate mitigation or compensating controls are implemented.

- Block—The block outcome is produced when evidence or artifact requirements are not met for High-risk or Critical-risk tier AI systems.
- Research sandbox only—Minimal-risk tier systems with minimal risk but also almost negligible controls are restricted to research sandbox use.

Table 3 defines the tiered TAI/EAI floors that determine deployability for each GRS tier.

**Table 3.** RBAAF deployment gate criteria with respect to *TAI* and *EAI* as required by *GRS*.

| GRS Band | GRS Range   | TAI Min.    | EAI Min.    | Decision                                                                                                               |
|----------|-------------|-------------|-------------|------------------------------------------------------------------------------------------------------------------------|
| Critical | $\geq 0.85$ | $\geq 0.80$ | $\geq 0.75$ | Deploy only with continuous monitoring and audits. BLOCK if monitoring and audits are not configured.                  |
| High     | 0.70–0.84   | $\geq 0.70$ | $\geq 0.70$ | Deploy with controls. BLOCK if controls are not configured.                                                            |
| Moderate | 0.50–0.69   | $\geq 0.60$ | $\geq 0.60$ | Conditional deploy for pilot or controlled roll-outs. Sandbox mode deployment only if remediations are not configured. |
| Low      | 0.30–0.49   | $\geq 0.50$ | $\geq 0.50$ | Limited or internal use deployment is allowed. Sandbox mode deployment only if thresholds are not met.                 |
| Minimal  | $< 0.30$    | –           | –           | Allowed for Research sandbox only.                                                                                     |

Figure 4 illustrates the gate as a risk-versus-assurance map where we have used  $t = \min(\text{TAI}, \text{EAI})$  on the horizontal axis and GRS on the vertical axis. Image on the Left shows the operational deployment logic with horizontal bands indicating the GRS risk tiers, and thick tier boundary lines mark the tier cut-offs. The dashed step functions show the minimum assurance floors required for Conditional and Deploy outcomes as an AI system is eligible for deployment only if its bottleneck assurance  $t$  is better than the tier-specific deploy floor. Otherwise, it is downgraded to the conditional deployment or blocked if it lies below the tier. Image on the Right shows the iso-AAS contours derived from  $\text{AAS} = \sqrt{\text{GRS} \cdot t}$ , which provides a continuous readiness ranking. For a fixed risk level, higher  $t$  increases AAS. And, for a fixed assurance level, higher GRS reduces AAS.

Figure 5 shows the summary of assurance maturity via the (TAI, EAI) plane. The heatmap shows the encoding of the bottleneck principle, such as the limiting pillar, whether it be traceability or explainability, always determining the assurance readiness level.

### 3. Results

We evaluated the performance of RBAAF on a number of simulated AI models for testing purposes. The evaluations comprise five use-case scenarios (A1, A2, B1, B2, and C1), each with fully specified risk inputs and evidence vectors detailed in Supplementary Tables S1–S3. For each case, we computed *GRS*, *TAI*, *EAI*, and *AAS* and applied tiered gates. This helped us in analyzing the stability of GRS rankings under uncertainty as well as the repeatability of TAI or EAI evidence scoring under reassessment. We also compared the integrated gating outcomes to baseline practices that rely on qualitative documentation without standardized evidence scoring.

#### 3.1. GRS Behavior and Risk-Tier Stability

The GRS score generates monotonic risk tier assignments for all the AI models evaluated. These tier assignments remain stable under uncertainty characterizations. The P50 scores maintain the rank order with respect to deterministic estimates when scoring

inputs are passed as distributions. Additionally, the P90 values indicate assumption sensitivity [52,53] for borderline use cases. The uncertainty characterization is an important add-on to test for operational governance, as it ensures that deployment outcomes are robust and that smaller perturbations will not bring any significant changes to the audit findings [3,9].

The AI models with comparable risk severity scores are classified differently when the deployment context differs. The difference in context can be and is not limited to rights sensitivity, information being evaluated, or jurisdictional obligations being different for different geographical locations. This observation supports the guidelines provided in governance frameworks to treat risk as not only a model property but also a use-case and operational attribute [2,4–6]. Figure 3 shows the percentile and distributional uncertainty endorsing the tier classification when conducting tier-stability analysis.

### 3.2. Relevance Analysis

TAI and EAI modules ensure that RBAAF provides operational assessments when evaluating the documentation, evidences or artifacts. TAI scores improve when there are measures in place for dataset versioning, model versioning, reproducible run manifests, and audit-ready logs. All these techniques ensure end-to-end provenance and consistency of decisions that is the core spirit of the traceability phenomenon [4,39,40]. The cases of missing documentation or provenance via tools only—such as absence of identifiers in logs or no data or model lineage—result in lower TAI scores. This approach ensures that TAI operates on principle of verifiable artifacts instead of just evaluating narrative descriptions [40].

The mere presence of explanations is not enough to skew EAI scores to acceptable levels. Rather the EAI considers the quality and stability of explanations provided to the users. The AI systems using standard explainable AI methods like LIME or SHAP may improve EAI only if they are accompanied by evidence of fidelity, faithfulness, stability under perturbations, robustness checks, and stakeholder usability assessment [27–32]. This approach resulted in alignment with earlier concerns of untested explanations to be used to induce overtrust or mask failure modes [33].

### 3.3. RBAAF Performance for the Use-Case Applications

We have evaluated RBAAF against three AI models on five different use cases. L-XAIDS [54] is evaluated for two different use cases. A1 being L-XAIDS' use as an academic prototype and then as an operational utility, A2 in a Security Operations Center (SOC) environment. Likewise, biometric access control is assessed twice, B1 with weak TAI and EAI, and less assurance. And B2 being a case of high assurance with stronger scores in TAI and EAI. Both are subject to EU AI Act Annex III [51], as biometric identification systems are explicitly classified as high-risk systems. The third AI model evaluated is for credit scoring that is subject to Basel III guidelines [9] and EU AI Act [51] Annex III regulations. Table 4 summarizes the computed scores and gate outcomes for all use-case applications.

We note that the values reported in Table 4 are calculated outputs obtained by calculating the GRS scoring inputs (severity core and modifier overlays) and the assurance evidence vectors for TAI and EAI for each use case, and then applying Equations (2)–(5) and Algorithm 1. The details of these calculations are provided with the complete parameter definitions and step-by-step calculations for all cases in Supplementary Table S1, including the risk inputs (e.g.,  $L$ ,  $I$ , and modifier overlays  $C$ ,  $G$ ,  $T$ ,  $E$ ,  $R$  with their respective weights), intermediate metrics and terms, and the resulting  $GRS$ ,  $TAI$ ,  $EAI$ , and  $AAS$  values.

The use-case results show the expected governance pattern observed across regulatory and management systems guidance. We observe that as the risk score increases, the burden

of evidence for traceability and explainability must also increase proportionately to ensure the responsible deployment of high-risk AI systems. In RBAAF, this proportionality is operationalized by mapping continuous risk (GRS) with tiers and tier-dependent minimum floors for both TAI and EAI. This approach ensures an auditable and repeatable gate outcome across domains.

**Table 4.** RBAAF Use-case summary.

| Case                        | GRS  | TAI   | EAI   | AAS   | Gate Decision                                                                                                                 |
|-----------------------------|------|-------|-------|-------|-------------------------------------------------------------------------------------------------------------------------------|
| L-XAIDS A1 (lab)            | 0.62 | 0.714 | 0.735 | 0.665 | Conditional deployment only with pilot or controlled rollouts. Classified in Moderate GRS tier                                |
| L-XAIDS A2 (SOC)            | 0.83 | 0.714 | 0.735 | 0.770 | Deploy with controls and continuous monitoring since classified in High GRS tier                                              |
| Biometric access control B1 | 0.88 | 0.26  | 0.39  | 0.48  | BLOCK the deployment since it fails TAI and EAI requirements. Lies in Critical GRS tier.                                      |
| Biometric access control B2 | 0.88 | 0.86  | 0.75  | 0.814 | Deploy with continuous audits since TAI and EAI indices, and AAS score is above the threshold. GRS score is in critical tier. |
| Credit scoring C1           | 0.81 | 0.83  | 0.75  | 0.782 | Deploy with controls since its a high GRS classified system. TAI and EAI match the minimum requirement for the tier           |

### 3.3.1. Cybersecurity Intrusion Detection (L-XAIDS) [54]

In the evaluation of intrusion detection settings, explainability is important because it provides security analysts an ability to understand why an alert is generated in order to respond to threats and avoid false positives. RBAAF evaluated the explainability artifacts, and they contributed positively to EAI, but readiness was confined by the stability and faithfulness of explanation and also on the pipeline artifacts to ensure reproducibility (TAI). The proper versioning and documented audit logs resulted in improved TAI, but if evidence maturity was incomplete, the deployment gate produced a conditional outcome that explicitly identified traceability remediation as the blocking factor [39,40].

### 3.3.2. Biometric Access Control

The use of AI for biometric access control inherently increases the governance sensitivity because of its implications on privacy and potential discriminatory outcomes [2,36]. Due to this inherent high-risk nature of the application, RBAAF rightly assigned higher risk tiers to the biometric AI models. Hence, this tier assignment requires stricter TAI and EAI thresholds. In use case B2, the model lineage, dataset provenance, and decision logging are enabled. Hence, it achieved higher TAI, which was the opposite when compared to use case B1. The outcome of both these use cases demonstrates the effectiveness of conservative gate decisions that are consistent with regulatory expectations for high-risk use cases. RBAAF ensures that high-risk AI systems are not allowed to deploy unless both traceability and explainability evidence maturity meet the tier-specific minimum thresholds, aligning the gate outcome with stricter regulatory expectations for high-risk use cases [2,4].

### 3.3.3. Credit Risk Scoring

The use of AI systems in credit risk scoring requires transparency and sufficient explanations to be provided for human oversight. These requirements are driven by governance regulations and stakeholder needs [36–38]. In the use case C1, the TAI score

is significantly better than the EAI score, which results in the bottleneck term  $t$  confined by EAI. RBAAF, in turn, produces a conditional gate outcome that focuses on the quality improvements of the explanation provided. This outcome illustrates the operational value of integrating evidence maturity into governance decisions [3,4].

Compared to approaches that rely on qualitative documentation alone, the RBAAF demonstrations produce a verifiable link that maps declared risk inputs with relevant evidence artifacts to provide quantifiable scores and an explicit deployability decision. This approach makes it easier to audit why a system is granted a decision outcome of Deploy, Conditional, Sandbox, Block under any given governance posture.

#### 4. Discussion

The results of this study enforce a central pillar of responsible AI deployment, i.e., quantitative risk scoring is necessary to measure risks but is not sufficient unless it is combined with targeted risk treatment and auditable assurance, as demonstrated in Section 3. RBAAF exhibits three intended behaviors across all the evaluated scenarios. Firstly, it shows its inherent property of risk proportionality. The cases that lie in higher GRS tiers require stricter evidence floors to pass the deploy gate. Secondly, it shows the native property of bottleneck conservatism. In cases where even when one pillar is strong, the weaker of traceability or explainability evidence reduces the overall readiness because  $t = \min(\text{TAI}, \text{EAI})$  limits AAS with the weaker assurance axes. Lastly, it demonstrates the inherent property of auditability. RBAAF requires the inputs, weights, and intermediate terms to be explicitly defined. This ensures that respective scores, objective understanding regarding risk magnitudes, risk appetites, and gate decisions can be independently verified and re-assessed.

The Governance Risk Score (GRS) enables organizations to prioritize risks and helps address the recurring operational gaps in AI governance. Many risk assessments remain qualitative, which in turn makes it difficult to benchmark the risks consistently across systems and deployments [2–4]. The TAI score provides coverage for proper documentation of evidence and artifacts. The lack of evidence results in scores that would demonstrate the AI model's lack of deployment readiness even when the technical performance metrics appear to be plausible. The EAI score provides coverage for explanations that are tested for fidelity, faithfulness, robustness, relevance, and stability.

Our research shows that RBAAF should be treated as a governance workflow component rather than using it as a one-time assessment criterion. The following sequence is recommended for implementing RBAAF: (i) compute GRS scores to prioritize risk assessment and treatment; (ii) compute and evaluate TAI scores to ensure provenance, logging, and reproducibility requirements are met; (iii) compute and evaluate EAI scores to ensure explanation quality is appropriate and useful for the stakeholders; and (iv) apply tiered gate outcomes to produce actionable and auditable decisions with proper remediation plans in place. The architecture of RBAAF complements the governance procedures outlined in NIST AI RMF [3] and ISO/IEC 42001 [4].

Summary of use-case scenarios: The outcomes of five use-case scenarios demonstrate that the deployment gate metrics are not just a cosmetic addition to risk scoring. This approach materially changes the deployability decisions when evidence maturity is weak, as evident in Table 4 and Section 3.3. These findings support the hypothesis that risk severity alone is insufficient for responsible deployment, especially if it is without any auditable evidence of maturity. Additionally, uncertainty characterization via P50, P90, and Monte Carlo variability demonstrates a robustness check for borderline tier decisions and aligns with risk-assessment guidance emphasizing sensitivity to assumptions [9,52,53].

Prior governance frameworks and regulations like the EU AI Act, NIST AI RMF, ISO/IEC 42001, and OECD/UNESCO focus primarily on documentation, transparency

principles, accountability of the decisions, and continuous monitoring. But these frameworks and guidelines intentionally do not restrict organizations with operational detail for quantitative risk scoring, linking evidence maturity to deployability outcomes, or any other operational implementations [2–6]. On the contrary, quantitative scoring and risk analysis is an approach used for some time in the cybersecurity domain. CVSS and FAIR are used by organizations to support consistent risk prioritization and reporting [7–9].

RBAAF bridges these two approaches by (i) quantifying governance-aligned severity in a repeatable manner and (ii) adding evidence-backed adequacy indices (TAI and EAI) to translate traceability and explainability readiness into auditable deployment gate outcomes. This design addresses the observed gap where many AI risk taxonomies remain descriptive without a comparable benchmarking mechanism [42–44].

Because weight profiles and thresholds encode organizational posture and risk appetite, practical use of RBAAF should include lightweight sensitivity checks such as varying  $\alpha$  and modifier weights  $\beta$ , or TAI and EAI component weights within plausible ranges, to verify whether tier assignment and gate decisions remain stable. This supports governance robustness and reduces dependence on any single assumed parameter setting, consistent with uncertainty-aware risk assessment practice [9,52,53]. RBAAF can be integrated into existing model governance routines such as pre-deployment reviews, change-control and change management routines, and periodic audits by treating GRS as the prioritization metric and using TAI and EAI as non-exhaustive evidence checklists that are re-calculated when models, data, or deployment contexts change. This approach supports repeatable reporting to internal risk committees and stakeholders and enables traceable justification for deployability outcomes with explicit remediation targets.

## 5. Conclusions

This paper presented the Risk-Based AI Assurance Framework (RBAAF) that integrates quantitative risk scoring with model assurance evidence to ensure responsible AI deployment. RBAAF is designed to map high-level governance requirements into operationally measurable criteria. Overall, our results show the practicality of RBAAF by demonstrating that risk scoring alone cannot guarantee responsible AI deployment. By integrating risk scores with standardized, evidence-driven assurance indices, RBAAF ensures traceability, reproducibility, explainability, and accountability in AI systems with processes in place that can be embedded directly into internal audits and conformity assessments.

Revisiting the research questions. Overall, the findings of this research address the three research questions defined in Section 1.2.

- For RQ1, RBAAF establishes an auditable, quantitative link between governance-aligned risk severity and evidence maturity by calculating GRS, TAI, EAI, and AAS from declared inputs, weights, and evidence artifacts (see Section 3 and Supplementary Tables S1–S3).
- For RQ2, RBAAF addresses this RQ by integrating assurance as a bottleneck term via  $t = \min(\text{TAI}, \text{EAI})$  that prevents optimistic readiness outcomes and does not average out the indices. Instead, it limits the deployment outcome with the weaker index which is reflected in the gate outcomes for cases with weak evidence pillars (Section 3.3).
- For RQ3, RBAAF employs a tier-dependent assurance thresholds that yields deployment outcomes of Deploy, Conditional deploy, Sandbox and Block. This tiered approach reflects the proportional governance expectations across use cases, as summarized in Table 4 and discussed in Section 4.

Limitations and future work. RBAAF presented in this paper is evaluated on the reported experiments and scenarios, should not be treated as an out-of-the-box, tailor-made

solution for all possible scenarios. Practical use of RBAAF may require understanding of the technical and operational landscape of the organization, as well as adhering to the understanding of applicable governance obligations. This study is scenario-based and uses weight profiles. Organizations may need to calibrate the thresholds against their respective risk appetite and audit outcomes. From a framework-extension perspective, future work may include: (i) learning or calibrating weight profiles and tier thresholds from organizational incident history and audit outcomes; (ii) adding dedicated adequacy pillars for privacy and security indices to extend AAS beyond traceability and explainability; and (iii) automating evidence extraction from logs, registries, and CI/CD artifacts to reduce the assessor's or auditor's documentation burden [3,4,9]. As AI governance regimes, such as EU AI Act conformity assessment practices or ISO 42001, mature, the auditable scoring-and-gating framework like RBAAF can support the standardized evidence-based reporting across teams and deployments. The future research direction based on RBAAF may focus on empirical validation in real organizational deployments such as audit cycles or conformity-assessment workflows to evaluate the evidence scoring, threshold calibration against risk appetite for specific high-risk domains, and the predictive utility of gate outcomes for post-deployment incidents and remediation effectiveness.

**Supplementary Materials:** The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/info17030263/s1>, Table S1: GRS inputs and term-by-term calculations (Equations (1)–(5)). Table S2: Scenario-specific traceability evidence scores and weighted computation of TAI using default weights. Table S3: Scenario-specific explainability evidence scores and weighted computation of EAI using default weights. Table S4: End-to-end derivation of bottleneck term  $t = \min(\text{TAI}, \text{EAI})$ , AAS.

**Author Contributions:** A.E.M. conceived the study, developed the methodology, performed the analysis, and wrote the manuscript. K.-C.Y. supervised the research and provided critical feedback on the methodology and manuscript revisions. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The datasets analyzed in this study are publicly available from the AI Incident Database (AIID) and the AVID-ML Repository (AVID-ML). No new datasets were generated during the current study.

**Acknowledgments:** We thank the reviewers and editors for their time and constructive feedback. We also acknowledge the support of colleagues and collaborators who contributed through discussions and early input.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

|           |                                                   |
|-----------|---------------------------------------------------|
| AI        | Artificial Intelligence                           |
| AIID      | AI Incident Database                              |
| EU AI Act | European Union Artificial Intelligence Act        |
| AVID      | AI Vulnerability Database                         |
| ATLAS     | MITRE Adversarial Threat Landscape for AI Systems |
| CIA       | Confidentiality, Integrity, Availability          |
| CVSS      | Common Vulnerability Scoring System               |
| FAIR      | Factor Analysis of Information Risk               |

|               |                                                                                                     |
|---------------|-----------------------------------------------------------------------------------------------------|
| GDPR          | General Data Protection Regulation                                                                  |
| HIPAA         | Health Insurance Portability and Accountability Act                                                 |
| HITL          | Human-in-the-Loop                                                                                   |
| ISO/IEC 27001 | Information security management systems—Requirements                                                |
| ISO 31000     | Risk management—Guidelines                                                                          |
| ISO/IEC 42001 | AI Management System (AIMS)                                                                         |
| NIST AI RMF   | NIST AI Risk Management Framework                                                                   |
| OECD          | Organisation for Economic Co-operation and Development                                              |
| OWASP         | Open Worldwide Application Security Project                                                         |
| UNESCO        | United Nations Educational, Scientific and Cultural Organization                                    |
| RBAAF         | Risk-Based AI Assurance Framework                                                                   |
| GRS           | Governance Risk Score                                                                               |
| TAI           | Traceability Adequacy Index                                                                         |
| EAI           | Explainability Adequacy Index                                                                       |
| AAS           | Assurance Adequacy Score                                                                            |
| NIST          | National Institute of Standards and Technology                                                      |
| SP            | Special Publication                                                                                 |
| CIS           | Center for Internet Security                                                                        |
| SAMM          | Software Assurance Maturity Model                                                                   |
| STRIDE        | Spoofing, Tampering, Repudiation, Information disclosure, Denial of service, Elevation of privilege |
| LLM           | Large Language Model                                                                                |
| ML            | Machine Learning                                                                                    |
| CI/CD         | Continuous Integration and Continuous Delivery/Deployment                                           |
| LIME          | Local Interpretable Model-agnostic Explanations                                                     |
| SHAP          | SHapley Additive exPlanations                                                                       |
| Grad-CAM      | Gradient-weighted Class Activation Mapping                                                          |

## References

1. Partnership on AI. AI Incident Database. 2023. Available online: <https://incidentdatabase.ai/> (accessed on 17 February 2026).
2. European Commission. *Regulation (EU) 2024/1689 of the European Parliament and of the Council on Artificial Intelligence (Artificial Intelligence Act)*; Official Journal of the European Union: Luxembourg, 2024.
3. National Institute of Standards and Technology. *AI Risk Management Framework (AI RMF 1.0)*; Technical Report; U.S. Department of Commerce: Gaithersburg, MD, USA, 2023. [CrossRef]
4. *ISO/IEC 42001:2023; Artificial Intelligence—Management System*. International Organization for Standardization: Geneva, Switzerland, 2023.
5. Organisation for Economic Co-Operation and Development. OECD AI Policy Observatory. 2023. Available online: <https://oecd.ai/en/> (accessed on 17 February 2026).
6. UNESCO. *Recommendation on the Ethics of Artificial Intelligence*; UNESCO: Paris, France, 2021.
7. FIRST.org. Common Vulnerability Scoring System v3.1: Specification Document. 2019. Available online: <https://www.first.org/cvss/> (accessed on 17 February 2026).
8. The Open Group. FAIR: Factor Analysis of Information Risk. 2022. Available online: <https://www.opengroup.org/open-fair> (accessed on 17 February 2026).
9. National Institute of Standards and Technology. *Guide for Conducting Risk Assessments (NIST SP 800-30 Rev. 1)*; Technical Report; U.S. Department of Commerce: Gaithersburg, MD, USA, 2012.
10. Berryhill, J.; Gehrke, N. The AVID Taxonomy for AI Vulnerabilities. 2022. Available online: <https://avidml.org/> (accessed on 10 August 2025).
11. Massachusetts Institute of Technology. MIT AI Incident Tracker. 2023. Available online: <https://airisk.mit.edu/ai-incident-tracker> (accessed on 10 August 2025).
12. *ISO/IEC 27001:2022; Information Security, Cybersecurity and Privacy Protection—Information Security Management Systems—Requirements*. International Organization for Standardization: Geneva, Switzerland, 2022.
13. *ISO 31000:2018; Risk Management—Guidelines*. International Organization for Standardization: Geneva, Switzerland, 2018.
14. Pavlidis, G. Unlocking the black box: Analysing the EU artificial intelligence act’s framework for explainability in AI. *Law Innov. Technol.* **2024**, *16*, 293–308. [CrossRef]

15. Ramos, S.; Ellul, J. Blockchain for artificial intelligence (AI): Enhancing compliance with the EU AI Act through distributed ledger technology. A cybersecurity perspective. *Int. Cybersecur. Law Rev.* **2024**, *5*, 1–20. [\[CrossRef\]](#)
16. Laux, J.; Wachter, S.; Mittelstadt, B. Trustworthy artificial intelligence and the European Union AI Act: On the conflation of trustworthiness and acceptability of risk. *Regul. Gov.* **2023**, *18*, 3–32. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Microsoft. The STRIDE Threat Model. 2009. Available online: <https://learn.microsoft.com/en-us/azure/security/develop/threat-modeling-tool> (accessed on 17 February 2026).
18. MITRE Corporation. Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS). 2023. Available online: <https://atlas.mitre.org/> (accessed on 17 February 2026).
19. OWASP Foundation. OWASP Top 10 for Large Language Model Applications. 2023. Available online: <https://owasp.org/www-project-top-10-for-large-language-model-applications/> (accessed on 17 February 2026).
20. Mauri, L.; Damiani, E. Modeling threats to AI-ML systems using STRIDE. *Sensors* **2022**, *22*, 6662. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Tan, M.; Yamaguchi, K.; Raney, A.; Nockles, V.; Leblanc, M.; Bendelac, S. An AI blue team playbook. In Proceedings of the Assurance and Security for AI-Enabled Systems, National Harbor, MD, USA, 21–25 April 2024; p. 26. [\[CrossRef\]](#)
22. Tyler, M.; McCeney, J. Assured AI reference architecture. In Proceedings of the Assurance and Security for AI-Enabled Systems, National Harbor, MD, USA, 21–25 April 2024; p. 27. [\[CrossRef\]](#)
23. Hamon, R.; Junklewitz, H.; Garrido, J.S.; Sánchez, I. Three challenges to secure AI systems in the context of AI regulations. *IEEE Access* **2024**, *12*, 61022–61035. [\[CrossRef\]](#)
24. Kim, D.; Shin, G.; Han, I.; Oh, H.; Han, M. Attack graph design and target priority based on attacker capabilities and network vulnerabilities. *J. Korean Inst. Intell. Syst.* **2022**, *32*, 332–339. [\[CrossRef\]](#)
25. Petersen, E.; Ganz, M.; Holm, S.H.; Feragen, A. On (assessing) the fairness of risk score models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 817–829. [\[CrossRef\]](#)
26. Anand, P.; Singh, Y.; Selwal, A.; Singh, P.K.; Ghafoor, K.Z. IVQFloT: An intelligent vulnerability quantification framework for scoring Internet of Things vulnerabilities. *Expert Syst.* **2021**, *39*, e12829. [\[CrossRef\]](#)
27. Ribeiro, M.T.; Singh, S.; Guestrin, C. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), New York, NY, USA, 13–17 August 2016; pp. 1135–1144. [\[CrossRef\]](#)
28. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS 30)*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
29. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626. [\[CrossRef\]](#)
30. Zhou, J.; Gandomi, A.; Chen, F.; Holzinger, A. Evaluating the quality of ML explanations: A survey. *Electronics* **2021**, *10*, 593. [\[CrossRef\]](#)
31. Bućinca, Z.; Lin, P.; Gajos, K.; Glassman, E. Proxy tasks and subjective measures can be misleading in evaluating XAI systems. In Proceedings of the 25th International Conference on Intelligent User Interfaces, Cagliari, Italy, 17–20 March 2020; pp. 454–464. [\[CrossRef\]](#)
32. Kim, J.; Maathuis, H.; Sent, D. Human-centered evaluation of XAI applications: A systematic review. *Front. AI* **2024**, *7*, 1456486. [\[CrossRef\]](#)
33. Langer, M.; Baum, K.; Hartmann, K.; Hessel, S.; Speith, T.; Wahl, J. Explainability auditing for intelligent systems: A rationale for multi-disciplinary perspectives. In Proceedings of the 2021 IEEE 29th International Requirements Engineering Conference Workshops (REW), Notre Dame, IN, USA, 20–24 September 2021; pp. 164–168.
34. Doshi-Velez, F.; Kim, P. Towards a rigorous science of interpretable machine learning. *arXiv* **2017**, arXiv:1702.08608. [\[CrossRef\]](#)
35. Papadimitriou, R. The right to explanation in the processing of personal data with the use of AI systems. *Int. J. Law Chang. World* **2023**, *2*, 43–55. [\[CrossRef\]](#)
36. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation). 2016. Available online: <https://eur-lex.europa.eu/eli/reg/2016/679/oj>. (accessed on 24 August 2025).
37. Wachter, S.; Mittelstadt, B.; Russell, C. Why a right to explanation matters in AI. *Harv. J. Law Technol.* **2017**, *31*, 491–505.
38. Bracke, P.; Datta, A.; Jung, C.; Sen, S. Machine learning explainability in finance: An application to default risk analysis. *SSRN* **2019**. [\[CrossRef\]](#)
39. Gundersen, O.; Kjensmo, S. State of the art: Reproducibility in artificial intelligence. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32. [\[CrossRef\]](#)

40. Mora-Cantalops, M.; Sánchez-Alonso, S.; García-Barriocanal, E.; Sicilia, M. Traceability for trustworthy AI: A review of models and tools. *Big Data Cogn. Comput.* **2021**, *5*, 20. [\[CrossRef\]](#)
41. Namli, T.; Sinacı, A.; Gönül, S.; Herguido, C.; García-Canadilla, P.; Muñoz, A.; Ertürkmen, G. A scalable and transparent data pipeline for AI-enabled health data ecosystems. *Front. Med.* **2024**, *11*, 1393123. [\[CrossRef\]](#) [\[PubMed\]](#)
42. Raji, I.; Kumar, I.; Horowitz, A.; Selbst, A. The fallacy of AI functionality. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, 21–24 June 2022; pp. 959–972. [\[CrossRef\]](#)
43. Slattery, P.; Saeri, A.; Grundy, E.; Graham, J.; Noetel, M.; Uuk, R.; Thompson, N. The AI risk repository. *AGI* **2024**, *1*. [\[CrossRef\]](#)
44. Wei, M.; Zhou, Z. AI Ethics Issues in Real World: Evidence from AI Incident Database. Available online: <https://hdl.handle.net/10125/103236> (accessed on 17 February 2026).
45. Joint Task Force. *Security and Privacy Controls for Information Systems and Organizations*; NIST Special Publication 800-53 Rev. 5; National Institute of Standards and Technology (NIST): Gaithersburg, MD, USA, 2020. [\[CrossRef\]](#)
46. Joint Task Force Transformation Initiative. *Guide for Conducting Risk Assessments*; NIST Special Publication 800-30 Rev. 1; National Institute of Standards and Technology (NIST): Gaithersburg, MD, USA, 2012. [\[CrossRef\]](#)
47. Boeckl, K.; Lefkowitz, N. *NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management, Version 1.0*; NIST Cybersecurity White Paper NIST.CSWP.01162020; National Institute of Standards and Technology (NIST): Gaithersburg, MD, USA, 2020. [\[CrossRef\]](#)
48. *ISO/IEC 27701:2025*; Information Security, Cybersecurity and Privacy Protection—Privacy Information Management Systems—Requirements and Guidance. International Organization for Standardization (ISO): Geneva, Switzerland; International Electrotechnical Commission (IEC): Geneva, Switzerland, 2025.
49. Center for Internet Security (CIS). *CIS Critical Security Controls Version 8*; Center for Internet Security (CIS): New York, NY, USA, 2021.
50. OWASP Foundation. *OWASP Software Assurance Maturity Model (SAMM) Version 2*; Release announcement (SAMM v2); OWASP Foundation: Wilmington, DE, USA, 2020.
51. European Commission. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). 2021. Available online: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206> (accessed on 17 February 2026).
52. Dou, M. Principle and applications of Monte-Carlo simulation in forecasting, algorithm and health risk assessment. *Highlights Sci. Eng. Technol.* **2024**, *88*, 406–414. [\[CrossRef\]](#)
53. Pasupuleti, M. Stochastic computation for AI: Bayesian inference, uncertainty, and optimization. *Int. J. Acad. Ind. Res. Innov.* **2025**, *5*, 1–23. [\[CrossRef\]](#)
54. Muhammad, A.E.; Yow, K.C.; Baćanin-Džakula, N.; Khan, M.A. L-XAIDS: A LIME-based eXplainable AI framework for intrusion detection systems. *Clust. Comput.* **2025**, *28*, 654. [\[CrossRef\]](#)

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.