

Enterprise AI Model Risk Scoring and Continuous Assurance Framework

A Comprehensive Enterprise Framework for Quantifying AI Model Risk, Prioritizing Controls, Monitoring Change, and Maintaining Continuous Assurance

Sanjeeve Kumar Gajadi*

*(HCL Technologies, Bengaluru, Karnataka, India)

Email: sanjeevekumar.g@hcltech.com

Abstract:

AI model portfolios now include predictive models, Generative AI, foundation models, Retrieval-Augmented Generation systems, and agentic components with widely different risk profiles. Static approval processes and one-time validation are insufficient because model behavior, data, usage, integrations, and external dependencies can change after deployment. This paper proposes an Enterprise AI Model Risk Scoring and Continuous Assurance Framework that combines risk dimensions, quantitative scoring, control requirements, lifecycle gates, monitoring, change detection, reassessment triggers, and residual-risk reporting. The framework evaluates business impact, autonomy, data sensitivity, model complexity, explainability, external exposure, reversibility, regulatory relevance, security, robustness, provider dependency, and operational resilience. It introduces inherent-risk scoring, control-effectiveness adjustment, residual-risk calculation, confidence scoring, and risk-tier-based assurance frequencies. The proposed approach enables enterprises to prioritize validation and monitoring resources, identify deteriorating controls, and maintain evidence-based assurance throughout the AI lifecycle.

Keywords — AI Model Risk, Model Risk Scoring, Continuous Assurance, AI Governance, Generative AI, Agentic AI, Model Validation, Residual Risk, AI Monitoring, Enterprise Risk Management

I. INTRODUCTION

AI model risk is the possibility that a model or AI-enabled system produces harmful, inaccurate, insecure, non-compliant, or operationally unreliable outcomes.

Modern AI portfolios require a consistent scoring method so that control intensity, validation depth, monitoring, and executive attention are proportional to actual risk.

II. RISK SCORING OBJECTIVES

The framework supports prioritization, risk-based approval, resource allocation, portfolio comparison, recertification, and escalation.

Scores should remain interpretable and should never replace expert judgment; they provide structure for evidence-based decisions.

III. RISK DIMENSIONS

Core dimensions include business impact, decision criticality, autonomy, user population, data sensitivity, external exposure, model complexity, explainability, reversibility, regulatory scope, security, robustness, and dependency concentration.

Additional factors may include safety, financial authority, customer vulnerability, geographic reach, and reliance on third-party models or data.

IV. INHERENT RISK SCORING

Inherent risk represents exposure before controls. Each dimension can be scored on a standardized scale and weighted according to enterprise priorities.

High inherent-risk systems should receive stronger independent validation, security testing, human oversight, and deployment controls.

V. CONTROL EFFECTIVENESS ADJUSTMENT

Control effectiveness reflects design adequacy, implementation completeness, operating evidence, automation, monitoring, and remediation performance.

Residual risk is calculated by adjusting inherent risk according to verified control strength rather than simply assuming that documented controls are effective.

VI. RESIDUAL RISK AND ACCEPTANCE

Residual risk should be compared with enterprise risk appetite and use-case-specific tolerance. Scores above tolerance require remediation, compensating controls, restricted deployment, or formal risk acceptance.

Risk acceptance should have an accountable owner, rationale, expiration date, and reassessment trigger.

VII. MODEL VALIDATION AND EVALUATION

Validation should test performance, robustness, bias, explainability, security, privacy, harmful outputs, retrieval quality, prompt behavior, and failure modes according to model type.

Generative and agentic systems require scenario-based testing because deterministic accuracy metrics alone are insufficient.

VIII. CONTINUOUS MONITORING

Production assurance should monitor model quality, drift, data changes, prompt changes, tool usage, abnormal behavior, policy violations, incidents, cost, and business outcomes.

Monitoring thresholds should be linked to risk tier so that critical systems generate faster escalation and more frequent review.

IX. CHANGE AND REASSESSMENT TRIGGERS

Material changes include new model versions, provider updates, prompt redesign, new data sources, new tools, broader user populations, increased autonomy, or changed regulatory scope.

These events should automatically trigger targeted reassessment rather than waiting for a scheduled annual review.

X. ASSURANCE FREQUENCY

Low-risk systems may use periodic automated checks, while high-risk systems require continuous monitoring, frequent recertification, and independent review.

Risk-based assurance frequency concentrates resources where failures would have the greatest enterprise impact.

XI. PORTFOLIO REPORTING

Portfolio dashboards should show risk distribution, residual-risk trends, overdue validations, high-risk exceptions, incident concentration, provider dependency, and control weaknesses.

Executive reporting should focus on material exposure and remediation progress rather than excessive technical detail.

XII. IMPLEMENTATION ROADMAP

Organizations can begin with a simple risk taxonomy and scoring model, validate it against known use cases, and progressively improve calibration using incident and performance data.

Integration with model registries, AI gateways, identity systems, observability platforms, and governance workflows can enable continuous assurance.

XIII. CONCLUSION

The Enterprise AI Model Risk Scoring and Continuous Assurance Framework provides a structured method for quantifying AI risk and maintaining assurance as systems evolve.

By linking risk, controls, monitoring, and reassessment triggers, enterprises can scale AI while keeping residual risk visible, accountable, and aligned with risk appetite.

ACKNOWLEDGMENT

The author acknowledges the contributions of enterprise architecture practitioners, AI governance communities, security professionals, standards bodies, and technology researchers whose published work has informed the concepts synthesized in this paper.

REFERENCES

- [1] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), 2023.

- [2] National Institute of Standards and Technology, AI RMF Generative Artificial Intelligence Profile, 2024.
- [3] ISO/IEC 42001:2023, Information technology — Artificial intelligence — Management system.
- [4] ISO 31000:2018, Risk management — Guidelines.
- [5] ISO/IEC 27001:2022, Information security, cybersecurity and privacy protection — Information security management systems.
- [6] National Institute of Standards and Technology, SP 800-207: Zero Trust Architecture, 2020.
- [7] The Open Group, TOGAF Standard, 10th Edition, 2022.
- [8] ISACA, COBIT 2019 Framework: Governance and Management Objectives, 2019.
- [9] OWASP Foundation, OWASP Top 10 for Large Language Model Applications.
- [10] MITRE, ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems.
- [11] OECD, Recommendation of the Council on Artificial Intelligence.
- [12] IEEE, Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems.