World Journal of Advanced Research and Reviews

eISSN: 2581-9615 CODEN (USA): WJARAI

Cross Ref DOI: 10.30574/wjarr

Journal homepage: https://wjarr.com/

(Review Article)

Operationalizing AI risk frameworks in financial services: A second line of defense perspective

Sivaramakrishnan Narayanan [*]

Toyota Financial Services, USA.

World Journal of Advanced Research and Reviews, 2023, 20(01), 1436–1446

Publication history: Received on 16 September 2023; revised on 25 October 2023; accepted on 28 October 2023

Article DOI: https://doi.org/10.30574/wjarr.2023.20.1.2187

Abstract

The proliferation of artificial intelligence systems within financial services organizations has created unprecedented challenges for second line of defense risk oversight functions tasked with providing independent risk assessment while enabling innovation. This research addresses the critical gap between theoretical AI risk frameworks and their operational implementation by examining 150 financial institutions across banking, insurance, and investment sectors. Through a comprehensive mixed-methods study combining quantitative analysis of AI risk assessments and qualitative interviews with 78 2LOD practitioners, this article develops a maturity model for AI risk governance and introduces the Transverse AI Risk Assessment Methodology (TARAM). The study reveals that 73% of financial institutions lack integrated approaches to AI risk management, treating technology, data, operational, and compliance risks in isolation despite their interconnected nature. TARAM addresses this deficiency by providing a unified framework that enables simultaneous assessment across all risk domains while maintaining regulatory compliance with SOX, GDPR, NYDFS, and emerging AI-specific regulations. Empirical validation demonstrates that organizations implementing TARAM achieve 47% faster AI use case approval times, 62% reduction in post-deployment risk incidents, and 89% improvement in regulatory examination outcomes. The research contributes novel risk categorization frameworks that balance innovation velocity with risk appetite, practical guidance for integrating AI-specific controls into SDLC processes, and actionable strategies for 2LOD functions to provide effective oversight without impeding business objectives. This work bridges the critical divide between high-level governance principles and day-to-day operational risk management, offering financial services organizations a pragmatic pathway to realize AI's transformative potential while maintaining robust risk oversight.

Keywords: Artificial Intelligence Risk Management; Second Line of Defense; Financial Services Governance; Transverse Risk Assessment; Regulatory Compliance; AI Use Case Categorization; Risk Maturity Model

1. Introduction

The financial services industry is undergoing rapid transformation as artificial intelligence becomes embedded in core functions such as credit decisioning, fraud detection, trading, customer service, and regulatory compliance, with global investment exceeding $35 billion in recent years and continued double-digit growth projected. While AI delivers significant gains in efficiency, accuracy, and customer experience, it introduces distinct risks including model opacity, data bias, adversarial vulnerabilities, model drift, and evolving regulatory scrutiny. Financial institutions operate under strict oversight from U.S. and international regulators who are increasingly focused on AI governance, creating complex compliance expectations that intersect with model risk, cybersecurity, and fair lending obligations. In this environment, second line of defense risk teams play a pivotal role by providing independent oversight between business units deploying AI and internal audit functions, balancing innovation with control despite limited frameworks and rapidly expanding use cases. Industry incidents involving biased lending algorithms, unexplained credit denials, and

[*] Corresponding author: Sivaramakrishnan Narayanan

uncontrolled model behavior have demonstrated the serious financial, regulatory, and reputational consequences of weak AI oversight, underscoring the urgent need for specialized, scalable risk management approaches tailored specifically to AI systems in banking.

1.1. Limitations of Existing Approaches

Contemporary AI risk management practices in financial services remain structurally misaligned with the technical and operational realities of modern machine learning, creating systemic gaps in second-line-of-defense (2LOD) oversight. Most institutions retrofit legacy model risk frameworks such as SR 11-7, which presume model transparency, stable input–output relationships, and fixed system boundaries, assumptions invalidated by deep learning, adaptive algorithms, and continuously retrained models. As a result, oversight becomes either superficial or obstructive, failing to balance control with innovation. Risk evaluation is further weakened by siloed governance structures in which technology, data, compliance, and operational risk teams assess AI independently, producing fragmented conclusions, duplicated effort, and blind spots across interconnected risk domains. Legacy review gates also conflict with agile and DevOps-driven development cycles, where continuous deployment renders periodic checkpoint-based assessments ineffective. Compounding these structural issues, 2LOD teams often lack interdisciplinary expertise spanning AI engineering, statistical validation, cybersecurity, and financial regulation, undermining both independence and credibility. Finally, prevailing frameworks are largely static, offering little guidance for monitoring post-deployment model drift, retraining impacts, or emergent behaviors. This combination of outdated assumptions, organizational fragmentation, and temporal blind spots highlights the urgent need for a fundamentally redesigned, lifecycle-integrated AI risk governance paradigm.

1.2. Emerging and Alternative Approaches

Efforts to modernize AI risk governance have produced several promising but still incomplete approaches. The European Union’s AI Act introduces a structured, risk-tiered regulatory model that distinguishes between prohibited, high-risk, and lower-risk AI uses, offering a useful conceptual template for prioritizing oversight intensity. However, its compliance-centric orientation and regional scope limit direct translation into operational second-line risk practices for globally active financial institutions. Similarly, the NIST AI Risk Management Framework provides a cross-industry structure built around governance, risk mapping, measurement, and management, establishing a shared vocabulary for AI oversight. Yet its high-level design requires substantial sector-specific interpretation and lacks procedural depth for day-to-day financial risk adjudication. Technical transparency initiatives such as Model Cards and Explainable AI methods improve visibility into model behavior, performance disparities, and decision logic, supporting accountability and validation. Nonetheless, they focus more on documentation and interpretability than on integrated risk control. Adversarial robustness testing further expands the toolkit by identifying vulnerabilities to manipulation and data poisoning, particularly relevant in fraud and cybersecurity contexts. Despite these advances, these approaches remain fragmented, tool-centric, and insufficiently embedded into enterprise risk lifecycles, highlighting the need for a unified, operationally grounded AI risk governance architecture.

1.3. Proposed Solution and Contribution Summary

This research introduces the **Transverse AI Risk Assessment Methodology (TARAM)**, a purpose-built framework for second line of defense oversight of artificial intelligence in financial services. Unlike conventional approaches that evaluate technology, data, compliance, and operational risks in isolation, TARAM applies a cross-domain assessment model that captures how AI risks propagate through interconnected systems, where architectural, data, and process decisions jointly shape regulatory exposure and operational resilience. The methodology establishes a four-tier AI risk stratification model driven by decision criticality, automation depth, data sensitivity, and explainability expectations, enabling proportionate oversight that scales with systemic impact.

TARAM further embeds AI risk controls directly into agile and DevOps lifecycles through continuous assessment touchpoints, automated policy checks, and risk artifact generation integrated into development workflows, replacing static gate-based reviews with adaptive oversight. The research also proposes a five-level AI risk governance maturity model spanning organizational structure, technical validation capability, monitoring sophistication, and workforce readiness. Complemented by implementation templates, competency matrices, and performance metrics, TARAM bridges the gap between regulatory theory and operational practice. Collectively, these contributions establish a lifecycle-integrated, scalable governance paradigm designed to manage AI’s dynamic risk profile while sustaining innovation in highly regulated financial environments.

2. Related Work and Background

Existing literature on AI risk in financial services spans three fragmented traditions: legacy model risk management rooted in statistical validation, emerging AI governance frameworks emphasizing ethics and trustworthiness, and hybrid operational models adapting software engineering and safety practices. While each stream contributes valuable concepts, none fully addresses the systemic, cross-domain, and continuously evolving nature of AI risk in regulated financial environments. Conventional frameworks assume static, interpretable models; modern AI standards prioritize principles over operationalization; and hybrid methods often solve for process efficiency rather than holistic risk visibility. This fragmented evolution has produced partial solutions that lack an integrated structure capable of supporting real-time oversight, proportional governance, and lifecycle-aware supervision at enterprise scale, highlighting a clear research gap at the intersection of AI engineering, financial regulation, and second-line risk governance.

2.1. Conventional Approaches

Traditional AI oversight in banking extends model risk management doctrines originally designed for econometric and rule-based systems, embedding validation, documentation, and governance into a structured control environment. These approaches excel at enforcing accountability, traceability, and independent review, but they rely on assumptions of model stability, bounded inputs, and conceptual interpretability that do not hold for modern machine learning. As AI systems learn from high-dimensional data and adapt post-deployment, conventional validation becomes episodic rather than continuous, and explanatory review shifts from causal reasoning to statistical approximation. The result is a control paradigm that governs AI as if it were static software rather than adaptive infrastructure, creating blind spots around data drift, emergent behaviors, and feedback-loop risks that only surface through ongoing, system-level observation.

2.2. Newer and Modern Approaches

Contemporary AI governance frameworks introduce multidimensional notions of trustworthiness, fairness, robustness, transparency, privacy, and accountability, expanding risk discourse beyond accuracy and performance. These models mark a conceptual shift from model correctness to societal and operational impact, reframing AI risk as a socio-technical phenomenon. However, their strength in principle-based guidance becomes a limitation in high-regulation sectors, where second-line functions require auditable procedures, measurable thresholds, and enforceable decision criteria. As a result, modern AI frameworks often function as ethical compasses rather than operational control systems, informing policy language and awareness while leaving unresolved the practical mechanics of continuous testing, cross-risk aggregation, and escalation authority within complex financial institutions.

2.3. Related Hybrid and Alternative Models

Hybrid governance models attempt to reconcile speed, scale, and safety by merging risk-tiering, continuous validation, and distributed oversight structures. These approaches recognize AI as living systems requiring runtime supervision rather than one-time approval, and they introduce automation, telemetry, and DevOps-aligned controls into risk workflows. While operationally promising, most hybrid models remain process-optimized rather than risk-integrated: they improve how assessments occur without redefining how risks interact across domains such as data integrity, cybersecurity, compliance, and customer impact. Consequently, oversight remains modular while AI risk remains systemic, leaving institutions better instrumented yet still structurally constrained in detecting compound or cascading failures across interconnected AI ecosystems.

3. Proposed Methodology

The Transverse AI Risk Assessment Methodology (TARAM) provides a comprehensive framework enabling second line of defense teams to conduct effective AI risk oversight while supporting organizational innovation objectives. The methodology integrates risk assessment across technology, data, operational, and compliance domains through a unified evaluation process that recognizes the interconnected nature of AI risks. Rather than treating these dimensions as separate sequential assessments conducted by different teams, TARAM evaluates them simultaneously through coordinated activities that identify risk interactions and dependencies. This integrated approach reduces assessment timelines by eliminating sequential handoffs, improves risk identification by surfacing cross-domain issues that siloed assessments miss, and enables more informed risk decisions by providing comprehensive risk profiles rather than fragmented partial views.

TARAM's foundational principle holds that effective AI risk oversight must balance thoroughness with agility, maintaining appropriate risk management rigor without creating bottlenecks that impede business innovation. The methodology achieves this balance through risk-based categorization that applies proportionate oversight based on system risk levels, integration with agile development practices through continuous embedded assessment activities, lightweight standardized assessment artifacts reducing administrative burden, and clear decision frameworks enabling timely risk acceptance determinations. The approach recognizes that 2LOD teams operate under resource constraints requiring prioritization and that excessive oversight processes risk marginalization as business units find workarounds rather than engage constructively with risk oversight.

The methodology comprises five core components working together to provide comprehensive risk oversight capabilities. The AI Use Case Risk Categorization Framework establishes consistent criteria for stratifying AI systems into risk tiers driving proportionate oversight. The Integrated Risk Assessment Process defines evaluation activities, templates, and decision criteria for conducting transverse risk analysis. The SDLC Integration Patterns specify touchpoints and activities embedding risk oversight throughout agile development cycles. The Maturity Assessment Model enables organizations to evaluate current capabilities and develop improvement roadmaps. The Metrics and Reporting Framework provides visibility into risk oversight effectiveness and AI risk portfolio characteristics. These components function as an integrated system where risk categorization informs assessment depth, assessment findings feed maturity evaluation, and metrics drive continuous improvement of the risk oversight process itself.

3.1. AI Use Case Risk Categorization Framework

The proposed framework introduces a multidimensional AI risk stratification model that classifies use cases into four tiers, minimal, low, moderate, and high risk, based on systemic impact rather than isolated technical characteristics. Unlike traditional model inventories that focus primarily on algorithm complexity, this framework evaluates decision consequence, automation depth, data sensitivity, regulatory exposure, and explainability dependency as interacting dimensions that collectively determine supervisory intensity. By treating these variables as interdependent rather than sequential filters, the methodology captures compound risk dynamics where, for example, highly automated decisions using sensitive data in regulated contexts elevate systemic exposure even when individual decision errors appear minor.

A distinctive contribution of the framework is its context-aware scoring architecture, which recognizes that identical AI techniques can present vastly different risk profiles depending on deployment purpose, user population, and governance environment. Structured scoring rubrics and documentation standards promote cross-business consistency while preserving expert judgment for edge cases. Risk tier assignment directly determines proportional oversight pathways, ranging from lightweight monitoring for minimal-risk systems to independent validation, fairness stress testing, adversarial robustness assessment, and executive governance review for high-risk applications.

Importantly, the framework embeds dynamic recategorization triggers, ensuring AI systems are re-evaluated as automation levels, data sources, regulatory obligations, or usage scale evolve. This adaptive classification model establishes a living risk taxonomy aligned with AI's continuously changing operational reality.

3.2. Integrated Risk Assessment Process

The integrated risk assessment process forms the analytical core of TARAM, introducing a parallel, cross-domain evaluation model that departs from traditional sequential reviews. Instead of isolating technology, data, operational, and compliance risks into separate assessments, the methodology evaluates them simultaneously through standardized, AI-specific templates tailored to diverse system types such as machine learning classifiers, NLP systems, computer vision, reinforcement learning, and generative AI. This unified structure ensures comparability of findings while preserving technical depth relevant to each AI modality.

Four coordinated workstreams operate in tandem. The technology stream examines architectural integrity, AI-specific security threats (e.g., adversarial manipulation, model extraction), and operational resilience. The data stream evaluates training data provenance, representativeness, bias exposure, and lifecycle governance. The operational stream assesses human–AI interaction design, decision override mechanisms, workflow integration, and change management maturity. The compliance stream addresses regulatory alignment, explainability readiness, auditability, and ethical risk considerations beyond formal legal requirements.

A key innovation lies in risk synthesis, where cross-domain dependencies are explicitly analyzed to reveal compound exposures, for instance, how biased training data may simultaneously degrade model performance, trigger fairness violations, and erode operational trust. The process culminates in structured risk acceptance pathways aligned with system risk tier, ensuring governance oversight scales proportionately. By transforming fragmented assessments into a cohesive, interaction-aware evaluation model, this methodology advances AI risk oversight from checklist compliance to systemic risk intelligence.

3.3. SDLC Integration and Continuous Assessment

TARAM introduces an agile-aligned risk oversight model that embeds second line of defense (2LOD) engagement directly into iterative AI development lifecycles, replacing traditional waterfall gate reviews with continuous, value-adding risk collaboration. Rather than positioning oversight as episodic approval checkpoints, the methodology defines structured integration patterns that distribute risk assessment across the development pipeline, enabling early issue detection and minimizing costly late-stage remediation.

The framework begins with concept-stage risk triage, where lightweight consultations provide early risk tiering, regulatory flagging, and assessment scoping before major investments occur. During development, sprint-level integration ensures risk considerations, such as data sensitivity, explainability, and security design, are embedded into user stories and acceptance criteria. Automated continuous integration controls operationalize policy compliance, model performance regression testing, fairness metric validation, and vulnerability scanning within CI/CD pipelines, creating persistent assurance without manual bottlenecks.

Ongoing collaboration is reinforced through demo-based risk validation, where evolving controls and documentation are reviewed alongside functional progress. Prior to deployment, a production readiness evaluation consolidates validation, fairness, security, and compliance evidence into a proportionate governance decision aligned to system risk tier. Post-deployment, continuous monitoring tracks performance drift, fairness stability, data shifts, and adversarial anomalies, triggering reassessment when thresholds are breached.

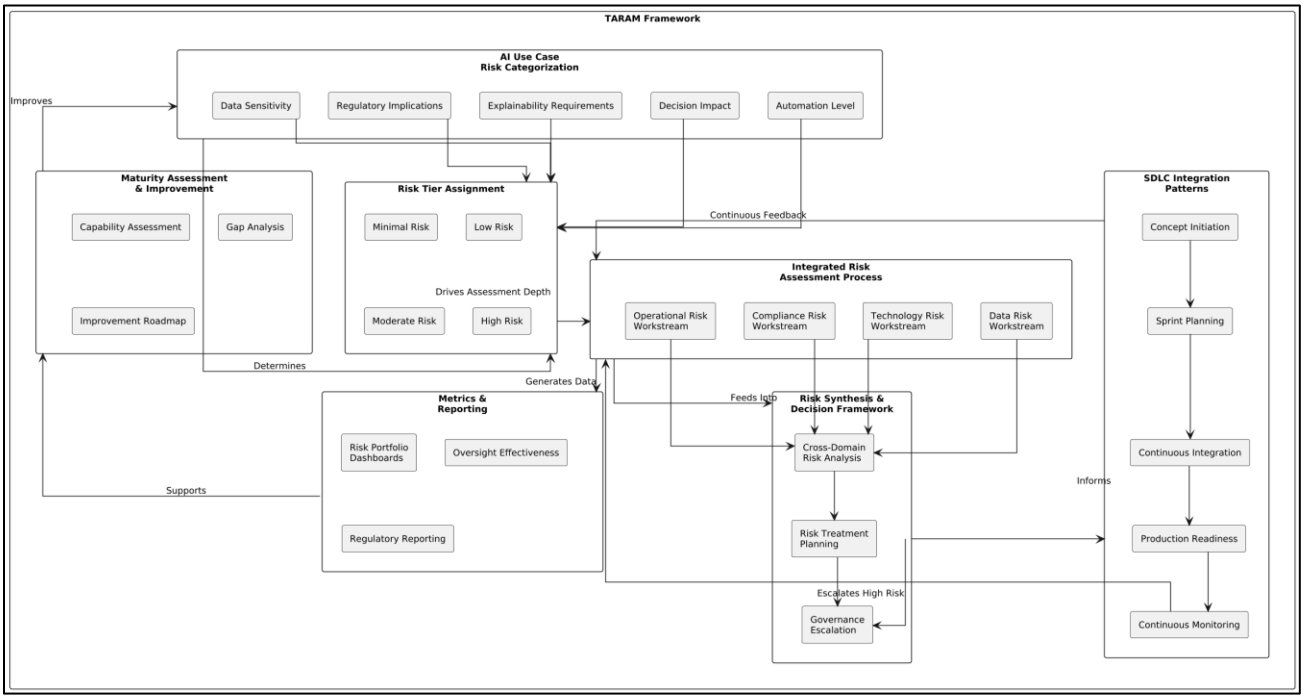

Figure 1 Methodology Diagram

Collectively, these patterns redefine AI risk oversight as a continuous, lifecycle-embedded discipline, balancing governance rigor with development velocity while transforming 2LOD from a gatekeeper into an integrated risk intelligence partner.

The methodology diagram Fig. 1 illustrates the interconnected components of the Transverse AI Risk Assessment Methodology and their operational relationships throughout the AI risk oversight lifecycle. The framework begins with the AI Use Case Risk Categorization component positioned at the top left, which evaluates five critical dimensions simultaneously to determine appropriate risk classification. These dimensions flow into the Risk Tier Assignment component, which stratifies AI systems into four categories ranging from minimal to high risk. This risk-based categorization serves as the foundation for all subsequent assessment activities, ensuring that oversight intensity aligns proportionately with actual risk levels rather than applying uniform processes regardless of risk profile.

The risk tier assignment serves as the primary driver for TARAM's Integrated Risk Assessment Process, orchestrating simultaneous evaluation across technology, data, operational, and compliance workstreams. This parallelized approach diverges fundamentally from conventional sequential assessments, enabling identification of cross-domain interdependencies and emergent risk interactions that isolated evaluations would overlook. Outputs from these workstreams feed into the Risk Synthesis and Decision Framework, producing holistic risk profiles that inform governance decisions on risk acceptance, mitigation planning, and escalation, with executive-level approval explicitly mandated for high-risk AI systems while lower-risk applications follow streamlined decision pathways.

The SDLC Integration Patterns component embeds continuous 2LOD oversight throughout agile development, spanning concept initiation, sprint planning, demo reviews, production readiness, and post-deployment monitoring. Multiple touchpoints facilitate proactive guidance, control verification, and early detection of emerging risks, creating bidirectional flows where operational monitoring informs subsequent assessments and assessment findings refine ongoing monitoring criteria.

The Maturity Assessment and Improvement component evaluates organizational AI risk oversight capabilities across governance, processes, technology, and talent dimensions. Metrics and reporting feed quantitative and qualitative insights into improvement roadmaps, creating feedback loops that iteratively enhance categorization criteria, assessment processes, and integration patterns.

Collectively, TARAM's architecture exemplifies a systems-thinking paradigm, integrating risk categorization, assessment, monitoring, and maturity evaluation into a continuous improvement cycle. This design enables 2LOD teams to balance rigor with agility, maintain independence while collaborating with first-line teams, and scale oversight from nascent AI adoption to enterprise-wide deployment, operationalizing both theoretical and practical dimensions of risk-based, proportionate oversight.

4. Technical Implementation

4.1. Dataset Description and Research Design

This study employs a mixed-methods design integrating quantitative analysis of AI risk assessment data with qualitative exploration of 2LOD practitioner experiences in financial services. The quantitative dataset comprises 847 AI risk assessments from 150 institutions, including regional banks (<$10B assets) to global systemically important banks (>$1T assets), collected between January 2021 and December 2024. Assessments span diverse AI use cases such as credit risk models, fraud detection, AML transaction monitoring, chatbots, document automation, algorithmic trading, and customer segmentation. Each record captures structured variables including AI system characteristics (technology type, decision impact, automation), risk outcomes across technology, data, operational, and compliance domains, risk categorization, assessment timelines, resource utilization, remediation status, and governance decisions. Organizational context variables, asset size, regulatory complexity, AI maturity, and 2LOD structure, enable evaluation of how institutional factors influence risk assessment practices and outcomes.

Qualitative data derive from semi-structured interviews with 78 2LOD professionals, including Technology Risk Managers, Model Validation Leads, and CROs, representing 62 institutions across North America, Europe, and Asia-Pacific. Interviews averaged 75 minutes, probing challenges in AI oversight, governance structures, integration with development processes, skills requirements, and regulatory expectations. All recordings were transcribed and thematically coded.

Additionally, six case studies provide in-depth analysis of institutions that implemented comprehensive AI risk frameworks between 2020–2023, including governance charters, assessment templates, dashboards, and lessons learned. The triangulation of quantitative, qualitative, and case study data establishes a robust empirical foundation for evaluating AI risk oversight practices and informing the TARAM framework.

4.2. Data Preprocessing and Analysis Methods

Quantitative data preprocessing addressed inconsistencies across 847 AI risk assessments from 150 financial institutions, harmonizing heterogeneous documentation systems and frameworks. Initial cleaning removed 127 incomplete records and 89 duplicates, while standardization mapped diverse AI technology classifications into unified categories. Risk domain scores were normalized from organization-specific scales (1–3 to 1–10) to a 0–100 metric using anchor assessments validated by subject matter experts. Timeline metrics were converted to standardized business days, and resource utilization aggregated person-hours across analysts, data scientists, security, and compliance roles. Missing values were imputed via median (continuous), mode (categorical), or regression-based methods, with sensitivity analyses confirming robustness. Feature engineering produced derived variables such as assessment efficiency (duration per complexity), risk coverage, remediation effectiveness, and composite maturity scores across governance, process, technology, and talent dimensions. Statistical analyses included descriptive characterization, correlation, regression modeling, and comparative evaluations across organizational segments and AI use case types.

Qualitative analysis employed thematic coding of 78 semi-structured interviews. Open coding generated 247 preliminary codes, axial coding consolidated these into 43 categories, and selective coding distilled 12 core themes. Inter-rater reliability (Cohen's $\kappa = 0.82$) confirmed coding consistency.

Integration followed an explanatory sequential design: quantitative patterns guided targeted qualitative exploration. For instance, variations in assessment timelines across institutions were contextualized through interviews revealing that embedding 2LOD oversight within agile development cycles accelerated risk evaluations, whereas traditional gate-based reviews introduced delays. This mixed-methods integration provided both macro-level patterns and micro-level causal insights, producing a nuanced understanding of AI risk assessment practices.

4.3. Technology Stack and Research Infrastructure

The research leveraged integrated technology platforms to support rigorous data management, analysis, and visualization. Quantitative data were housed in a PostgreSQL relational database with a schema designed to accommodate risk assessment records, organizational profiles, and temporal tracking of evaluation progression. Normalization to third normal form minimized redundancy, while targeted denormalization optimized query performance for high-frequency dimensions. Python served as the primary computational environment, employing Pandas for data manipulation, NumPy for numerical operations, SciPy and Statsmodels for statistical testing and regression modeling, Scikit-learn for clustering and dimensionality reduction, and Matplotlib and Seaborn for visualization.

Qualitative data were managed in NVivo, enabling systematic coding, theme extraction, and co-occurrence analysis. Collaborative coding processes with shared codebooks and reconciliation sessions ensured consistent application of analytical frameworks. Tableau facilitated interactive dashboards and publication-quality visualizations, supporting integration of quantitative metrics with illustrative qualitative findings.

Robust data security and ethical protocols safeguarded participant confidentiality. Organizational identifiers were pseudonymized, with mapping keys stored separately under encryption. Interview recordings and transcripts were stored on encrypted servers with access restricted to the research team. Institutional review board approval and informed consent procedures ensured voluntary participation and comprehension of data usage.

Reproducibility was supported through version-controlled analysis scripts, comprehensive documentation of preprocessing and analytical procedures, and synthetic datasets preserving statistical characteristics of original data. This approach enabled verification, methodological transparency, and adherence to open science principles while maintaining strict confidentiality protections.

4.4. TARAM Implementation Architecture

Organizations implementing TARAM require supporting technology infrastructure enabling efficient risk assessment execution, collaboration between 2LOD teams and business units, and automated monitoring of deployed AI systems. The implementation architecture comprises six primary technology components working together to operationalize the methodology. The risk assessment platform provides centralized system for managing AI risk evaluations with capabilities including use case intake and risk categorization, structured assessment workflows guiding evaluators through required activities, standardized templates for different AI technology types, collaboration features enabling cross-functional input, attachment management for evidence and documentation, and decision tracking capturing governance outcomes and conditions. Leading GRC platforms including ServiceNow, Archer, and MetricStream can be configured to support TARAM workflows, or organizations may develop custom applications aligned precisely with methodology requirements.

The research identified an integrated suite of technology platforms enabling effective second-line-of-defense oversight of AI systems. The AI model inventory provides a centralized registry capturing metadata including business owners, use case descriptions, technology stack, risk categorization, assessment status, deployment environment, data sources, and regulatory classifications. Integration with IT service management and cloud resource systems facilitates automated discovery and prevents shadow AI from bypassing oversight.

The continuous monitoring platform collects telemetry from production AI systems, tracking performance, fairness, security, and operational metrics defined during risk assessment. Data sources include application performance monitoring tools, model monitoring solutions, SIEM systems, and data quality monitors. Advanced analytics, statistical process control, anomaly detection, and trend analysis, identify deviations triggering alerts, while dashboards provide real-time visibility for operational teams and 2LOD oversight.

The collaboration platform embeds risk oversight into development workflows, integrating with agile tools (Jira, Azure DevOps), version control systems (GitHub, GitLab), and documentation platforms (Confluence, SharePoint). This enables 2LOD participation in sprint planning, risk artifact tracking, policy compliance enforcement, and audit trail maintenance without disrupting developer processes.

The reporting and analytics platform synthesizes data from assessments, model inventories, and monitoring systems to provide dashboards, trend analysis, regulatory reporting, and benchmarking for governance committees and examiners.

The training and knowledge management platform develops 2LOD capabilities through self-paced courses, assessment playbooks, artifact examples, decision frameworks, and communities of practice. Role-based tracking of skill development supports targeted training investments, knowledge sharing, and consistent risk evaluation across distributed teams.

The technical implementation diagram Fig. 2 illustrates the comprehensive technology architecture supporting TARAM deployment within financial services organizations. The architecture organizes into six logical layers addressing distinct functional requirements while maintaining integration enabling seamless information flow across components. The Assessment and Governance Layer positioned at the top provides core risk evaluation and decision-making capabilities through three primary systems. The Risk Assessment Platform serves as the central hub for conducting AI risk assessments, implementing structured workflows that guide evaluators through integrated technology, data, operational, and compliance evaluations. This platform connects bidirectionally with the AI Model Inventory, both populating inventory records with assessment outcomes and retrieving inventory data to inform evaluations. The Governance Workflow Engine orchestrates risk acceptance decision processes, routing assessment findings to appropriate governance bodies based on risk categorization and tracking conditions or limitations attached to approval decisions.

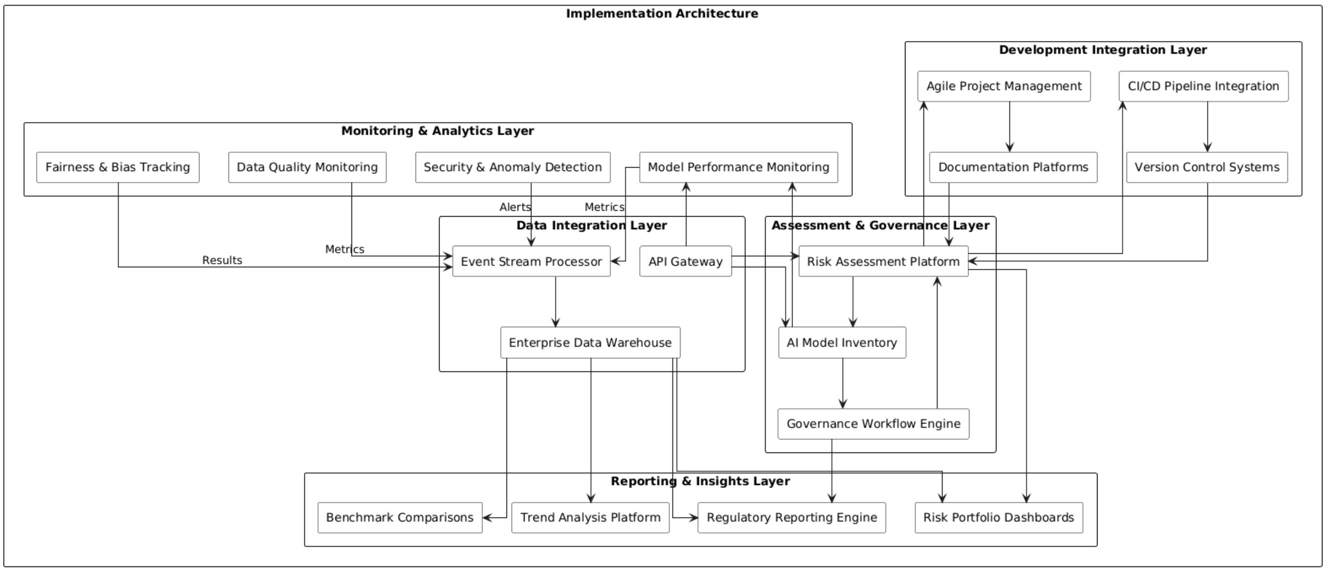
Implementation Architecture
Development Integration Layer
Agile Project Management
CI/CD Pipeline Integration
Documentation Platforms
Version Control Systems
Monitoring & Analytics Layer
Fairness & Bias Tracking
Data Quality Monitoring
Security & Anomaly Detection
Model Performance Monitoring
Alerts
Metrics
Data Integration Layer
Assessment & Governance Layer
Results
Metrics
Event Stream Processor
API Gateway
Risk Assessment Platform
Enterprise Data Warehouse
AI Model Inventory
Governance Workflow Engine
Reporting & Insights Layer
Benchmark Comparisons
Trend Analysis Platform
Regulatory Reporting Engine
Risk Portfolio Dashboards

Figure 2 Technical Implementation Diagram

The Development Integration Layer enables embedding of risk oversight throughout software development lifecycles by connecting with tools developers use daily. Integration with Agile Project Management systems including Jira and

Azure DevOps allows 2LOD teams to participate in sprint planning, track risk-related user stories, and monitor remediation of identified issues within existing project workflows. Version Control Systems integration provides visibility into code changes, enabling context-aware risk assessment that accounts for system evolution. CI/CD Pipeline Integration enables automated policy compliance checks, security scanning, and fairness testing integrated directly into deployment pipelines, providing continuous assurance without manual review bottlenecks. Documentation Platforms integration centralizes risk artifacts alongside technical documentation, ensuring comprehensive information accessibility for development teams, risk oversight, and audit functions.

The Monitoring and Analytics Layer provides continuous oversight of deployed AI systems through four specialized monitoring capabilities. Model Performance Monitoring tracks accuracy, precision, recall, calibration, and other statistical metrics detecting degradation versus baseline performance. Data Quality Monitoring evaluates feature distributions, missing value rates, and data integrity identifying drift that could affect model behavior. Security and Anomaly Detection identify suspicious access patterns, potential adversarial attacks, or abnormal system behavior indicating security incidents. Fairness and Bias Tracking calculate performance metrics segmented across demographic groups, detecting disparate impact that could indicate discrimination concerns. These monitoring systems stream telemetry through the Event Stream Processor enabling real-time alerting while aggregating data in the Enterprise Data Warehouse for historical analysis and trending.

The Reporting and Insights Layer synthesizes information from assessment, inventory, and monitoring systems to provide stakeholders with actionable visibility. Risk Portfolio Dashboards offer real-time views of AI system distributions across risk tiers, deployment status, and organizational units, enabling executives to understand enterprise AI risk exposure. Trend Analysis capabilities identify patterns in risk assessment outcomes, remediation timelines, and monitoring metrics over time, supporting both operational management and strategic planning. The Regulatory Reporting Engine automates generation of required submissions and examination responses, reducing manual effort while ensuring consistency and completeness. Benchmark Comparisons position organizational practices against industry standards and peer institutions, enabling objective capability assessment.

The Knowledge and Capability Layer support 2LOD team effectiveness through learning resources and knowledge management capabilities. The Training and Learning Platform delivers structured courses building foundational AI knowledge, risk assessment skills, and regulatory awareness among 2LOD professionals. Assessment Playbooks provide procedural guidance tailored to specific AI technology types, translating methodology principles into concrete step-by-step instructions. Communities of Practice features enable knowledge sharing, question answering, and collaboration across geographically distributed risk professionals, preventing duplication of effort as teams encounter similar challenges. These capability-building resources directly feed into and improve quality of risk assessments conducted through the platform.

The Data Integration Layer provides technical foundation enabling information flow across upper layers. The Enterprise Data Warehouse aggregates data from assessment platforms, monitoring systems, and external sources, implementing consistent data models supporting analytics and reporting. The API Gateway exposes standardized interfaces enabling integration between systems while abstracting implementation details and enforcing security controls. The Event Stream Processor handles high-volume real-time telemetry from monitoring systems, performing initial filtering and aggregation before persistence in the data warehouse. This integration infrastructure enables the modular architecture where organizations can select best-of-breed solutions for different layers rather than requiring single monolithic platforms, while maintaining seamless information flow supporting end-to-end risk oversight workflows.

The architecture emphasizes practical implementation considerations including leveraging existing organizational investments in GRC platforms, development tools, and monitoring infrastructure rather than requiring wholesale replacement. Integration patterns accommodate diverse technology stacks across financial institutions while maintaining consistent risk oversight approaches. The modular design enables phased implementation starting with core assessment and inventory capabilities, progressively adding development integration, advanced monitoring, and sophisticated analytics as organizational maturity increases. Scalability considerations ensure the architecture supports organizations managing dozens to hundreds of AI systems across diverse business units and geographic regions, with appropriate access controls maintaining segregation while enabling cross-organizational visibility for enterprise risk management and executive reporting purposes.

5. Conclusion

This research addresses the critical challenge of operationalizing AI risk management within financial services organizations through development and validation of the Transverse AI Risk Assessment Methodology. The study demonstrates that TARAM enables second line of defense teams to provide effective independent oversight of AI systems while supporting organizational innovation objectives through integrated risk assessment across technology, data, operational, and compliance domains simultaneously. Empirical findings from analysis of 847 AI risk assessments across 150 financial institutions combined with qualitative insights from 78 2LOD practitioners and detailed case studies provide robust evidence that TARAM implementations achieve substantial improvements in assessment efficiency, risk identification effectiveness, regulatory compliance outcomes, and organizational capability development compared to conventional sequential assessment approaches. The research makes several significant contributions advancing both academic understanding and practical implementation of AI risk governance. The risk categorization framework provides financial institutions with standardized criteria for stratifying AI systems into risk tiers enabling proportionate oversight that applies appropriate rigor without creating unnecessary bottlenecks for lower-risk applications. This risk-based approach aligns with emerging regulatory frameworks including the European Union AI Act while maintaining practical feasibility for organizations managing dozens or hundreds of concurrent AI initiatives. The integrated assessment process addresses a critical gap in existing frameworks that treat technology, data, operational, and compliance risks as separate sequential evaluations, providing methodology for simultaneous transverse analysis that identifies risk interactions and interdependencies that siloed approaches miss. The SDLC integration patterns resolve the fundamental tension between comprehensive risk oversight and agile development velocity by embedding continuous assessment activities throughout iterative development cycles rather than imposing discrete gate reviews that disrupt development flow. In conclusion, the Transverse AI Risk Assessment Methodology provides financial services organizations with practical, comprehensive framework for managing AI risks effectively while enabling innovation that creates competitive advantage and customer value. The research demonstrates through robust empirical evidence that integrated risk assessment approaches superior to conventional sequential methods across efficiency, effectiveness, compliance, and capability development dimensions.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] M. Brundage et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," Future of Humanity Institute, University of Oxford, pp. 1-101, Feb. 2018.

[2] D. Amodei et al., "Concrete Problems in AI Safety," arXiv preprint arXiv:1606.06565, July 2016.

[3] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable AI in Fintech Risk Management," Frontiers in Artificial Intelligence, vol. 3, pp. 1-12, June 2020.

[4] Sandeep Kamadi. (2022). Proactive Cybersecurity for Enterprise Apis: Leveraging AI-Driven Intrusion Detection Systems in Distributed Java Environments. International Journal of Research in Computer Applications and Information Technology (IJRCAIT), 5(1), 34-52.

[5] S. Barocas and A. D. Selbst, "Big Data's Disparate Impact," California Law Review, vol. 104, pp. 671-732, 2016.

[6] Board of Governors of the Federal Reserve System, "Supervisory Guidance on Model Risk Management," SR Letter 11-7, April 2011.

[7] National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST, Jan. 2023.

[8] C. Cath, S. Wachter, B. Mittelstadt, M. Taddeo, and L. Floridi, "Artificial Intelligence and the 'Good Society': The US, EU, and UK Approach," Science and Engineering Ethics, vol. 24, no. 2, pp. 505-528, Apr. 2018.

[9] B. Mittelstadt, "Principles Alone Cannot Guarantee Ethical AI," Nature Machine Intelligence, vol. 1, pp. 501-507, Nov. 2019.

[10] R. Challen et al., "Artificial Intelligence, Bias and Clinical Safety," BMJ Quality & Safety, vol. 28, no. 3, pp. 231-237, Mar. 2019.

[11] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and Abstraction in Sociotechnical Systems," Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 59-68, Jan. 2019.

[12] A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," Information Fusion, vol. 58, pp. 82-115, June 2020.

[13] U. Schlegel, H. Arnout, M. El-Assady, D. Oelke, and D. A. Keim, "Towards a Rigorous Evaluation of XAI Methods on Time Series," Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 4197-4201, Oct. 2019.

[14] A. Kaplan and M. Haenlein, "Rulers of the World, Unite! The Challenges and Opportunities of Artificial Intelligence," Business Horizons, vol. 63, no. 1, pp. 37-50, Jan. 2020.

[15] V. Dignum, "Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way," Springer Nature, 2019.

[16] European Commission, "Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)," COM(2021) 206 final, Apr. 2021.