Securing AI in the Supply Chain

A Comprehensive Guide to Third-Party AI Risk Assessment

As we witnessed in 2024, nearly 36% of all data breaches originated from third-party compromises, a 6.5% increase[1] from the previous year that cost organizations an average of $4.91 million per incident. Yet as Artificial Intelligence transforms business operations, traditional Third-Party Risk Management (TPRM) programs face an unprecedented challenge: 61% of companies now experience third-party breaches annually, with AI-powered systems introducing entirely new categories of risk that extend far beyond conventional cybersecurity concerns.[2]

This guide provides risk management professionals with a systematic approach for assessing AI-specific controls within third-party relationships. Unlike traditional TPRM approaches focused on point-in-time assessments, AI risk management demands continuous monitoring of model performance, bias detection, automated decision oversight, and regulatory compliance across an evolving landscape of AI applications.

Contents

The Evolution of Third-Party Risk

Traditional TPRM's New Reality

Traditional Third-Party Risk Management (TPRM) operated on a foundation of predictability. Organizations conducted comprehensive third-party control assessment at the time of onboarding, followed by annual or bi-annual risk reviews, monitored service quality through pre-defined SLAs, and tracked cybersecurity scores from rating providers. This approach assumed that vendor risk profiles remained relatively stable over time, with clear boundaries between vendor operations and organizational impacts. The occasional sanctions screening and periodic security questionnaire seemed adequate for managing typical third-party vendor relationships[3].

However, today's interconnected supply chains have fundamentally altered this risk equation. The digital transformation of business processes has created complex dependencies where a single vendor failure can cascade through entire ecosystems, disrupting operations far beyond the initial point of failure. Supply chain vulnerabilities now exceed network and application security risks in both frequency and impact, with organizations discovering that their cybersecurity posture is only as strong as their weakest vendor link.

The 2024 Change Healthcare breach exemplifies this new reality. When cybercriminals compromised their systems, the attack didn't just affect one company, it exposed 100 million records and disrupted healthcare services nationwide, preventing patients from accessing medications[4]. This incident demonstrated how third-party failures cascade through entire industries, affecting end customers who may never have directly interacted with the compromised vendor.

The AI Transformation Challenge

The integration of artificial intelligence technologies, including machine learning algorithms, generative AI systems, and agentic platforms has accelerated these supply chain risk challenges exponentially. While AI significantly improves operational efficiency, enhances user experiences, and delivers higher-quality services for third-party providers, it simultaneously introduces risk categories that transcend traditional cybersecurity boundaries.[5][6]

Unlike conventional software applications that execute predetermined logic consistently over time, AI systems exhibit dynamic behavior patterns that can evolve, adapt, and make autonomous decisions. These systems process vast amounts of personal and sensitive data, create detailed behavioral profiles, and make consequential decisions about individuals and organizations. The algorithms powering these systems can identify patterns and relationships that extend far beyond their intended scope, potentially influencing human behavior in ways that weren't anticipated during system design.

Consider how a third-party AI system used for customer service might gradually learn to identify frustrated customers and automatically route them to specific representatives to avoid escalation, a pattern that could constitute discriminatory service delivery without anyone explicitly programming such behavior. Or imagine a vendor's AI-powered pricing algorithm that inadvertently learns to associate certain zip codes with higher default rates, effectively implementing redlining practices that violate fair lending regulations.

The challenge becomes more complex when AI systems begin making decisions that affect other AI systems. Fourth-party AI dependencies create cascading effects where model drift, bias

amplification, or security vulnerabilities in one system can propagate through Nth-Parties— multiple layers of the supply chain, potentially affecting thousands of end customers before anyone identifies the root cause.

Regulatory Response and Compliance Imperative

Global regulators are responding to these emerging threats with unprecedented speed and scope. The European Union's AI Act, which began phased implementation in 2024 and continues through 2027, establishes comprehensive requirements for high-risk AI systems and explicitly mandates oversight obligations.[7][8] Organizations that deploy high-risk AI systems must now maintain detailed documentation, conduct conformity assessments, and ensure their third-party AI providers meet specific transparency & accountability standards.

Other forms of guidance such as the NIST AI Risk Management Framework advises the importance of governing AI risks throughout the system lifecycle, including explicit guidance for managing AI risks associated with external entities and third-party AI providers. It recognizes that AI risk management cannot be contained within organizational boundaries, it must extend throughout the entire supply chain.

Financial services regulators are developing sector-specific guidance that addresses algorithmic fairness in lending decisions, explainable AI requirements for credit scoring, and enhanced due diligence requirements for AI-powered fintech partnerships. Healthcare regulators are focusing on AI-enabled medical devices, diagnostic algorithms, and patient safety requirements for AI systems used in clinical decision-making.

Organizations can no longer afford to wait for complete regulatory clarity as the regulatory landscape is evolving too rapidly, and the potential consequences of non-compliance are too severe.

Companies must implement AI-specific risk management controls immediately to avoid compliance violations, operational disruptions, and reputational damage from vendor AI failures.

The Controls Assessment Approach for AI Risks

The following approach provides structured assessment criteria across critical AI risk sub-domains from the Halbarad Nth Party Risk Management Framework. Each sub-domain addresses specific risk characteristics unique to AI systems, with detailed guidance for systematic third-party evaluation, designed to support both strategic decision-making at the executive level and detailed technical review by risk analysts and AI engineers. In the following sections of this guide, we have divided these AI risk sub-domains into 3 categories based on potential impact and how quickly they need to be considered for assessment.

As in a traditional TPRM control assessment, the process begins by carefully defining the scope and objectives of the exercise, determining which vendors and AI solutions are relevant and the business processes they impact. The foundational step involves obtaining and reviewing key governance documents from the third-party AI provider, including the organization's AI policies, ethical guidelines, development and deployment standards, and governance structures. Assessors should pay particular attention to materials that illustrate executive oversight and accountability, such as organizational charts and RACI matrices, to ensure responsible leadership shapes operational behavior.

Next, the assessor thoroughly evaluates how AI models are integrated into vendor business workflows. This involves reading documentation about model capabilities, limitations, training and testing methods, and data usage, especially if any client data is involved. Regulatory compliance is another critical area; the assessor examines how the provider keeps pace with requirements like the EU AI Act, DPDP, NIST AI RMF, and SOC 2, looking for evidence in

compliance matrices, audit reports, and records of actions taken in response to regulatory exams. Understanding how audit findings are translated back into policy or process improvements is essential for gauging ongoing maturity.

A review of quality assurance practices rounds out this stage; assessors should ensure the vendor's QA teams have clear mandates for restricting AI use to approved business cases, monitor for unauthorized deployment sprawl, and are empowered to escalate potential misuse.

Site visits, interviews, or field validation give additional insight into how governance is practiced and how staff responds to issues in real scenarios, including compensating system-level access controls. The exercise concludes by synthesizing findings, highlighting strengths and gaps, and prioritizing which AI risk sub-domains merit further in-depth review and continuous monitoring based on risk exposure and business criticality.

This holistic approach balances the organizational context, documentation evidence, operational realities, and strategic imperatives, ensuring the assessment is actionable and thorough.

[Specific approach and actionable steps are provided in the appendix section of this document.]

Priority Level 1: Immediate Assessment Actions

The priority 1 AI risk sub-domains have the highest potential for immediate operational, legal, or reputational impact. These sub-domains should be prioritized for all critical AI service providers and AI third-parties, within the first 60 days of implementation.

Model Transparency and Explainability

Modern AI systems, particularly those based on deep learning architectures and large language models, often function as "black boxes" where the decision-making process remains opaque even to their creators. This opacity creates significant challenges for organizations that must understand, validate, and potentially defend the decisions made by third-party AI systems on their behalf.

Model transparency refers to the ability to understand how an AI system arrives at its decisions, while explainability focuses on communicating these decisions in terms that stakeholders can comprehend and act upon. When third-party vendors deploy AI systems that influence customer interactions, pricing decisions, or service delivery without providing adequate transparency, organizations face accountability gaps that can lead to regulatory violations, customer disputes, and operational inefficiencies.

The risk extends beyond technical opacity to encompass business accountability. When a vendor's AI system denies a loan application, recommends a medical treatment, or flags a transaction as fraudulent, the organization must be able to explain the reasoning behind that decision to customers, regulators, and internal stakeholders. Without adequate transparency, organizations find themselves defending decisions they cannot understand or explain.

Assessment Approach

Evaluate whether third-party vendors maintain comprehensive and responsible AI development processes, that prioritize transparency from system design through deployment. This includes reviewing documentation standards, explanation methodologies, and the vendor's ability to provide real-time explanations for system decisions. Vendors should demonstrate that AI model decisions can be explained in plain language to end-users and stakeholders within reasonable timeframes, typically within minutes rather than days or weeks.

The assessment should examine whether transparency and explainability criteria are integrated into the vendor's model selection and validation processes. Vendors should have documented procedures that prevent opaque or unaccountable models from reaching production environments through structured gates, peer review, and regular audits of transparency.

Warning Signs

The assessor should be particularly concerned when AI vendors cannot provide sample explanations of model decisions upon request, indicating either poor system design or inadequate preparation for transparency requirements. The absence of documented model selection criteria that address explainability requirements suggests that transparency is not prioritized during the development process. Additionally, one should be wary of vendors whose AI engineers and data scientists lack training on transparency requirements, or whose quality assurance teams cannot escalate models that fail transparency standards.

AI Data Privacy and Usage Risk

Systems utilizing AI, process personal data at unprecedented scale and granularity, creating detailed behavioral profiles that can reveal sensitive information about individuals even when such revelation wasn't the intended purpose. Third-party AI vendors often require access to customer data, transaction histories, behavioral patterns, and other sensitive information to deliver their services effectively. However, this data usage creates privacy risks that extend far beyond traditional database security concerns .

The challenge with AI data privacy lies in the technology's ability to infer sensitive information from seemingly innocuous data. An AI system trained to optimize delivery routes might inadvertently learn to predict personal relationships, health conditions, or financial situations based on address patterns and delivery frequencies. When third-party vendors have access to customer data for AI training or inference, they may unintentionally or deliberately create privacy violations that expose the organization to regulatory sanctions and customer trust erosion.

Modern privacy regulations create specific obligations for AI data usage that differ significantly from traditional data processing requirements. These regulations establish principles of data minimization, purpose limitation, and consent specificity that require careful implementation in AI contexts where data usage patterns may evolve over time.

Assessment Approach

Assessor must evaluate whether third-party AI vendors maintain formal data privacy policies that align with applicable jurisdictional standards and demonstrate practical implementation of privacy-by-design principles throughout the AI system lifecycle. The assessment should examine whether vendors conduct Privacy Impact Assessments before deploying new systems or substantially changing existing usage patterns.

The evaluation should focus on data collection practices, ensuring that vendors restrict data gathering to information that is necessary and directly relevant for stated AI purposes, with specific consent obtained for AI-related processing activities. Organizations should verify that vendors implement appropriate anonymization, de-identification, or pseudonymization techniques before AI model training where feasible, and maintain robust consent management processes that enable data subjects to exercise their rights effectively.

Warning Signs

The absence of Privacy Impact Assessments conducted before AI system deployment indicates inadequate privacy governance and potential regulatory non-compliance. Organizations should be concerned when vendors cannot demonstrate practical data minimization techniques in their AI training datasets or lack automated data deletion capabilities when legal or business purposes are fulfilled. The absence of documented consent management processes specifically designed for AI data usage suggests that vendors may not be prepared for privacy regulation enforcement.

Automated Decision Risk

The autonomous nature of AI systems enables them to make decisions with minimal or no human oversight, creating accountability challenges that don't exist with traditional software applications. While automation can improve efficiency and consistency, it also introduces the risk of scaled errors, biased outcomes, or decisions that violate organizational policies or regulatory requirements without immediate detection.[10][11]. Automated decision risk becomes particularly complex in third-party relationships where organizations may have limited visibility into the decision-making processes used by the vendors. Even when a vendor's AI system automatically approves credit applications, schedules maintenance activities, or routes customer service calls, your organization remains accountable for the outcomes even though it doesn't directly control the decision-making logic.

The challenge extends beyond individual decisions to encompass systemic impacts. An AI system that makes thousands of small decisions daily can create cumulative effects that significantly impact business operations, customer relationships, or regulatory compliance. For example, a vendor's AI-powered scheduling system might gradually optimize for operational efficiency in ways that inadvertently discriminate against customers in certain geographic areas, creating fair lending violations that emerge only through statistical analysis of long-term patterns.

Assessment Approach

Organizations should begin by ensuring that their third-party vendors maintain comprehensive inventories of all automated decision-making systems, with classification schemes that address business impact, regulatory exposure, and the level of human oversight provided. This inventory should include both obvious

decision points and subtle automated processes that might influence customer experiences or business outcomes.

The assessor should examine whether vendors conduct thorough risk assessments before deploying automated decision systems, with particular attention to ethical implications, operational impacts, fairness considerations, and privacy effects. Assessor should also verify that vendors have documented procedures for human oversight of high-risk automated decisions and maintain clear accountability structures through RACI matrices or similar governance frameworks. Critical to this assessment is understanding how vendors handle incidents when automated decisions cause privacy violations, ethical concerns, or reputational impacts. The assessor should evaluate whether vendors maintain incident management processes and have procedures in place to reduce potential regulatory or reputational impact from automated decisions.

Warning Signs

The absence of a comprehensive inventory of automated decision systems or risk classification methodology indicates inadequate governance over automated processes. Assessor should be particularly concerned when critical business decisions are automated without human oversight requirements or when vendors lack incident management processes specifically designed for automated decision failures. The absence of periodic reviews of automated decision risk appetite and oversight levels suggests that vendors may not be adapting their governance processes to evolving risk landscapes.

Priority Level 2: Intermediate Assessment Actions

These level 2 risk sub-domains require systematic actions but pose somewhat lower immediate operational risks. Organizations should address these areas after establishing foundational controls in Priority Level 1 sub-domains.

Bias, Fairness, and Non-Discrimination

Artificial Intelligence based systems can perpetuate, amplify, or create new forms of bias that result in unfair treatment of individuals or groups. Unlike human bias, which typically affects individual decisions, AI bias can scale to impact thousands or millions of decisions consistently, creating systematic discrimination that may violate civil rights laws, fair lending regulations, or equal opportunity requirements.[9]

The challenge with AI bias extends beyond intentional discrimination to include subtle algorithmic bias that emerges from training data patterns, feature selection decisions, or model optimization choices. An AI system trained on historical hiring data might learn to favor certain demographic groups based on past hiring patterns that reflected discriminatory practices, effectively perpetuating historical bias in new contexts.

Third-party AI bias becomes organizational liability when these systems affect customer decisions, service delivery, or access to opportunities. Organizations may find themselves facing discrimination lawsuits, regulatory investigations, or reputational damage based on biased decisions made by vendor AI systems, even when the bias wasn't intentionally programmed or obvious during system testing.

Assessment Approach

Assessor should evaluate whether third-party vendors conduct systematic bias impact analyses during AI model development and prior to deployment, with particular emphasis on high-risk systems that affect customer access to services or opportunities. This assessment should examine the representativeness and diversity of training and validation datasets, ensuring that vendors maintain sufficient data to represent all relevant population segments fairly. The evaluation should focus on fairness metrics and monitoring processes, including demographic parity analysis, equal opportunity assessments, equalized odds calculations, and both individual and group fairness measures. Assessor should verify that vendors conduct independent fairness audits using both internal resources and external experts, with findings reported to senior management and board-level governance structures. Critical to bias management is the implementation of “human-in-the-loop” oversight processes that provide mechanisms for reviewing and escalating bias-related issues. Ensure that the vendors have established clear procedures for bias detection, investigation, and remediation, and escalation paths when bias is identified.

Warning Signs

The absence of documented bias testing procedures or fairness metrics indicates inadequate attention to discrimination risks. Assessor should be concerned when vendors’ training datasets lack demographic diversity or adequate representation, as this can lead to systematic bias in system outputs. The lack of automated bias detection tools and/or manual review processes suggests that vendors may not be equipped to identify bias issues before they affect customers. Additionally, the absence of escalation procedures for bias detection or corrective action protocols indicates inadequate incident management capabilities.

AI Regulatory and Ethical Compliance

The regulatory landscape for AI is evolving rapidly across multiple jurisdictions, creating complex compliance requirements by geography, industry, and model type.

The EU AI Act establishes risk-based requirements for AI systems, with the most stringent requirements applied to high-risk applications such as credit scoring, employment decisions, and law enforcement tools. Financial services regulators are focusing on algorithmic fairness and explainable AI, healthcare regulators are addressing AI-enabled medical devices and clinical decision support systems, and data protection authorities are developing specific guidance for AI data processing. The challenge for organizations is that regulatory non-compliance can result in significant financial penalties, operational restrictions, and market access limitations. Under the EU AI Act, violations can result in fines up to 7% of global annual turnover for the most serious infractions, while sector-specific violations can trigger additional regulatory actions including consent orders, business restrictions, or license revocations.

Assessment Approach

Organizations should evaluate whether third-party vendors have established systematic processes for monitoring AI regulatory developments and updating their policies and procedures as new requirements are published. This assessment should examine the vendor's regulatory tracking capabilities, legal compliance resources, and change management processes for implementing regulatory updates.

The evaluation should focus on AI system documentation practices, ensuring that vendors maintain comprehensive records covering system source, versioning information, training data provenance,

intended use cases, known limitations, risk level classifications, and regulatory disclosure requirements. This documentation must be sufficient to support regulatory audits, customer inquiries, and internal governance processes.

Assessor should assess whether vendors conduct systematic impact analyses based on applicable AI risk management frameworks, with specific attention to legal, social, and ethical risks that could affect regulatory compliance. The assessment should also examine vendor due diligence processes for their own third-party AI providers, ensuring that compliance requirements cascade through the entire supply chain.

Warning Signs

The absence of systematic processes for monitoring AI regulatory changes across relevant jurisdictions indicates potential compliance vulnerabilities. Assessor should be concerned when vendors maintain inadequate AI system documentation that cannot support regulatory audits or examinations. The lack of regular compliance assessments or legal risk evaluations suggests that vendors may not be prepared for regulatory scrutiny. Additionally, the absence of contractual requirements for third-party AI compliance attestations indicates inadequate supply chain risk management.

Model Validation and Performance Drift

Unlike traditional software applications that maintain consistent performance over time, AI models experience performance degradation as real-world conditions change. This phenomenon, known as model drift or concept drift, occurs when the statistical relationships that the model learned during training no longer accurately reflect current reality[12]. Model drift can result from changes in customer behavior, market conditions, regulatory

environments, or data collection processes. A credit scoring model trained before the COVID-19 pandemic might have learned relationships between employment patterns and default risk that no longer hold true in a remote work environment.

Similarly, a fraud detection model might become less effective as criminals adapt their strategies to evade detection. The challenge with model drift is that it often occurs gradually and may not be immediately apparent through standard monitoring. Unlike application errors that typically produce obvious failures, model drift manifests as slowly degrading accuracy or increasing bias that may go unnoticed until significant damage has occurred.

Assessment Approach

Assessor should evaluate whether third-party vendors have established comprehensive model validation processes that include defined performance benchmarks, baseline guidelines, and regular validation schedules for all AI models. Also examine the vendor's approach to maintaining strict separation between validation and test datasets to prevent data leakage and ensure unbiased performance measurement. The evaluation should focus on monitoring capabilities, including statistical tests and visualization tools used to track changes in input feature distributions and target variable relationships over time. Organizations should verify that vendors implement automated monitoring for performance anomalies, data drift indicators, and other signals that might indicate model degradation. Critical to model validation is the vendor's approach to model retraining and redeployment. Assessor should check whether vendors have established automated pipelines for model retraining using recent data, with deployment processes that trigger only after successful validation against defined performance thresholds.

Warning Signs

The absence of defined performance benchmarks or validation schedules for AI models indicates inadequate model lifecycle management. Organizations should be concerned when vendors lack automated monitoring capabilities for performance anomalies or data drift, as manual monitoring processes are typically insufficient for detecting gradual performance degradation. The absence of documented procedures for root cause analysis of performance degradation suggests that vendors may not be equipped to address drift issues effectively. Additionally, the lack of automated model retraining and validation pipelines indicates that vendors may not be able to respond quickly to performance issues.

Priority Level 3: Strategic Assessment Actions

These risk sub-domains represent advanced AI governance capabilities that enhance long-term security posture and operational resilience. Act on these controls after establishing foundational and intermediate risk management capabilities.

AI Security and Adversarial Threats

AI systems face unique security vulnerabilities that don't exist in traditional applications. AI adversarial attacks involve deliberately crafted inputs designed to fool AI models into making incorrect decisions, while data poisoning attacks attempt to corrupt training datasets to influence model behavior. Model inversion attacks can extract sensitive information from training data, and model extraction attacks can steal proprietary algorithms through carefully crafted queries.

These threats are particularly concerning in third-party relationships because organizations may have limited visibility into the security measures protecting vendor systems. An adversarial attack against a vendor's fraud detection system could potentially allow fraudulent transactions to bypass detection, while a data poisoning attack against a vendor's system could manipulate customer behavior in ways that benefit competitors or cause reputational damage.

The sophistication of AI-specific attacks is increasing rapidly, with researchers regularly discovering new attack vectors and defensive countermeasures. Organizations must ensure that their third-party vendors stay current with emerging threats and implement appropriate defensive measures throughout the AI system lifecycle.

Assessment Approach

Assessor should evaluate whether third-party vendors conduct dedicated threat modeling for their AI systems, covering adversarial attacks, model theft attempts, data poisoning scenarios, model inversion risks, and privacy-related threats. The assessment should examine the vendor's understanding of AI-specific attack vectors and their implementation of appropriate defensive measures.

There should focus on adversarial training practices, examining whether vendors integrate adversarial examples and perturbations into their model training pipelines to enhance model resilience against evasion and manipulation attacks. Assessor should verify that vendors implement robust input validation and real-time data sanitization to filter suspicious, malformed, or anomalous inputs before they reach AI models.

Access control implementation represents another critical area for assessment, including granular controls on model endpoints, query frequency restrictions, and API output limitations designed to prevent model extraction and inference attacks. Evaluate whether vendors apply cryptographic protections to models and datasets, with monitoring systems that can detect unauthorized changes and enable rapid rollback when necessary.

Warning Signs

The absence of AI-specific threat modeling or security assessments indicates inadequate preparation for AI-targeted attacks. Organizations should be concerned when vendors lack adversarial training programs or model hardening techniques, as these represent fundamental defensive measures against common attack types.

The absence of input validation or anomaly detection for AI systems suggests vulnerability to adversarial input attacks. Additionally, inadequate access controls or API protection for model indicates potential exposure to model extraction attempts.

Data Quality and Training Data Risk

The quality of training data directly determines AI model performance, fairness, and compliance characteristics. Poor data quality can introduce bias, create security vulnerabilities, compromise model reliability, and lead to regulatory violations. In third-party relationships, organizations often have limited visibility into training data quality processes, yet they remain accountable for the outcomes produced by vendor AI systems.

Data quality challenges in AI contexts extend beyond traditional data management concerns to include representativeness across relevant population segments, temporal stability of data relationships, and the absence of spurious correlations that could lead to discriminatory outcomes. Training datasets must accurately reflect the operational environment where AI systems will be deployed while avoiding historical biases that could perpetuate unfair treatment. The challenge becomes more complex when vendors use synthetic data generation, data augmentation techniques, or transfer learning approaches that may introduce subtle quality issues that aren't apparent through standard data validation processes.

Assessment Approach

Assessor should evaluate whether third-party vendors maintain training datasets that are representative of the problem space and operational environment, with adequate coverage across relevant segments, time periods, and geographic regions. This assessment

should examine the vendor's data collection methodologies, sampling strategies, and validation processes for ensuring dataset representativeness. The evaluation should focus on data preprocessing practices, including deduplication procedures, anomaly removal techniques, error correction processes, and missing value imputation strategies.

Organizations should verify that vendors implement automated data quality pipelines that check for schema mismatches, invalid values, referential integrity violations, and suspicious distribution patterns in both historical datasets and ongoing data feeds. Version control and change management represent critical areas for assessment, ensuring that vendors maintain complete documentation of dataset modifications, annotation processes, and update rationales. Organizations should also evaluate access management, encryption, and monitoring controls for all training datasets, both at rest and in transit.

Warning Signs

Training datasets that lack representativeness of the operational environment or target populations indicate potential bias and performance issues. Organizations should be concerned when vendors lack documented data preprocessing or quality assurance procedures, as this suggests inadequate attention to data quality management. The absence of automated data quality validation pipelines indicates potential vulnerability to data quality degradation over time. Additionally, inadequate access controls or audit trails for training data management suggests potential security and compliance vulnerabilities.

Human-in-the-Loop Governance

Effective human oversight provides essential safeguards against AI errors, bias detection, edge case management, and accountability maintenance. Human-in-the-loop processes ensure that critical decisions receive appropriate human judgment while maintaining the efficiency benefits of AI automation. The challenge in third-party relationships is ensuring that vendors implement appropriate human oversight without creating operational bottlenecks or transferring inappropriate decision-making authority to vendor personnel.

Organizations must balance the need for human oversight with the practical realities of scaled AI operations and ensure that human reviewers have adequate training, authority, and support to make effective decisions. Human-in-the-loop governance becomes particularly complex when dealing with high-volume, low-latency AI applications where traditional approval processes may not be feasible. Organizations must work with vendors to develop innovative approaches that provide meaningful human oversight without compromising system performance or customer experience.

Assessment Approach

Assessor should evaluate whether third-party vendors maintain comprehensive policies that mandate human oversight in AI systems, with clear specifications of when human involvement is required and under whose authority. This assessment should examine the vendor's systematic identification and documentation of AI workflows and decision points that require human review, particularly focusing on high-impact, legally regulated, or ethically consequential decisions.

The evaluation should focus on structured processes for human review, approval, and escalation, including defined roles, threshold criteria, and documentation requirements.

Assessor should verify that vendors restrict decision-making authority through appropriate role-based or attribute-based access controls, with periodic reviews and comprehensive logging of all human-in-the-loop activities. Training and certification programs represent another critical assessment area, ensuring that personnel serving as human reviewers receive appropriate education on risks, bias recognition, escalation procedures, privacy requirements, and ethical compliance standards.

Warning Signs

The absence of formal policies requiring human oversight for AI decision-making indicates inadequate governance structures. Organizations should be concerned when critical AI decisions are automated without documented human review requirements or when vendors lack comprehensive training programs for human reviewers. The absence of audit trails or logging for human interventions and approvals suggests inadequate accountability and traceability.

Over-Reliance Risk Management

Organizations and their employees can develop unhealthy dependencies on AI systems, leading to reduced critical thinking skills, acceptance of incorrect outputs, and inability to operate effectively when AI systems fail. Over-reliance risk is particularly challenging because it develops gradually and may not become apparent until AI systems produce significant errors or become unavailable .

In third-party relationships, over-reliance risk can compound when organizations become dependent on vendor AI capabilities without maintaining adequate internal expertise to evaluate AI outputs or manage operations during vendor system failures. This dependency can create single points of failure that expose organizations to operational disruptions, poor decision-making, and competitive disadvantages. The challenge extends beyond technical dependencies to encompass skill atrophy, where employees lose the ability to perform tasks manually or make decisions without AI assistance. Organizations must ensure that their third-party vendors support appropriate levels of human skill maintenance while delivering the efficiency benefits of AI automation.

Assessment Approach

Assessor should evaluate whether third-party vendors conduct regular assessments to identify where and how users might develop over-reliance on AI systems, with monitoring processes that track critical tasks for over-reliance indicators such as lack of human intervention, unchecked output acceptance, or degraded manual capabilities. The evaluation should focus on training and education programs that provide ongoing AI literacy instruction, model limitation awareness, cognitive bias recognition, and critical thinking skill development. Organizations should verify that vendors

implement user interfaces and communication processes that clearly convey AI capabilities, uncertainties, and known limitations while providing rationales for key outputs.

User experience design represents another critical assessment area, examining whether vendors embed features that encourage users to pause, reflect, and review AI-driven outputs before acceptance or action. Organizations should also assess whether vendors maintain regular exercises, simulations, and manual procedures that prevent skill atrophy and build resilience for AI system failures.

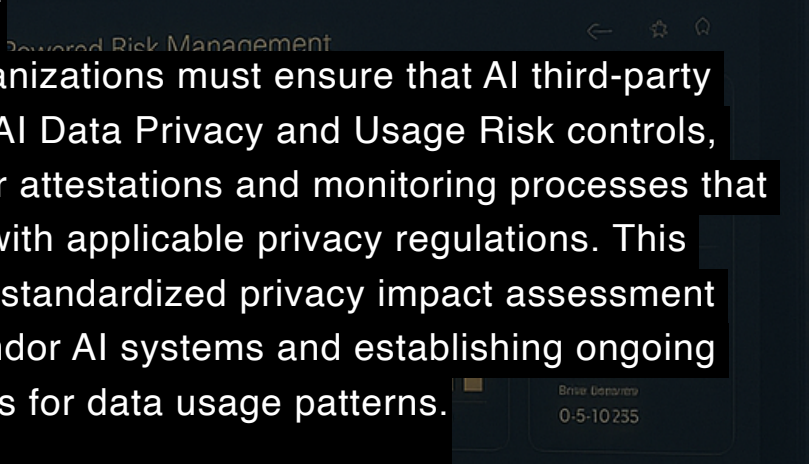
Warning Signs

The absence of systematic assessments for over-reliance indicators suggests inadequate attention to human factors in AI deployment. Organizations should be concerned when vendors lack comprehensive AI literacy training programs or critical thinking skill development initiatives. The absence of backup procedures or manual processes for AI system failures indicates potential operational vulnerabilities during system outages or failures.

Assessment Roadmap

Phase 1: Foundational Capabilities

The initial assessment phase focuses on establishing fundamental AI risk management capabilities that address the most critical immediate risks are effective. Organizations should begin by conducting comprehensive assessments of Model Transparency and Explainability for all critical AI vendors, ensuring that they can understand and explain AI-driven decisions that affect customers or business operations.



Simultaneously, organizations must ensure that AI third-party provider has robust AI Data Privacy and Usage Risk controls, implementing vendor attestations and monitoring processes that ensure compliance with applicable privacy regulations. This includes developing standardized privacy impact assessment requirements for vendor AI systems and establishing ongoing monitoring processes for data usage patterns.

The foundation phase concludes with Automated Decision Risk inventory and oversight requirements, ensuring that all third-party and vendor automated decision systems are identified, classified, and subject to appropriate human oversight based on their risk levels and business impacts.

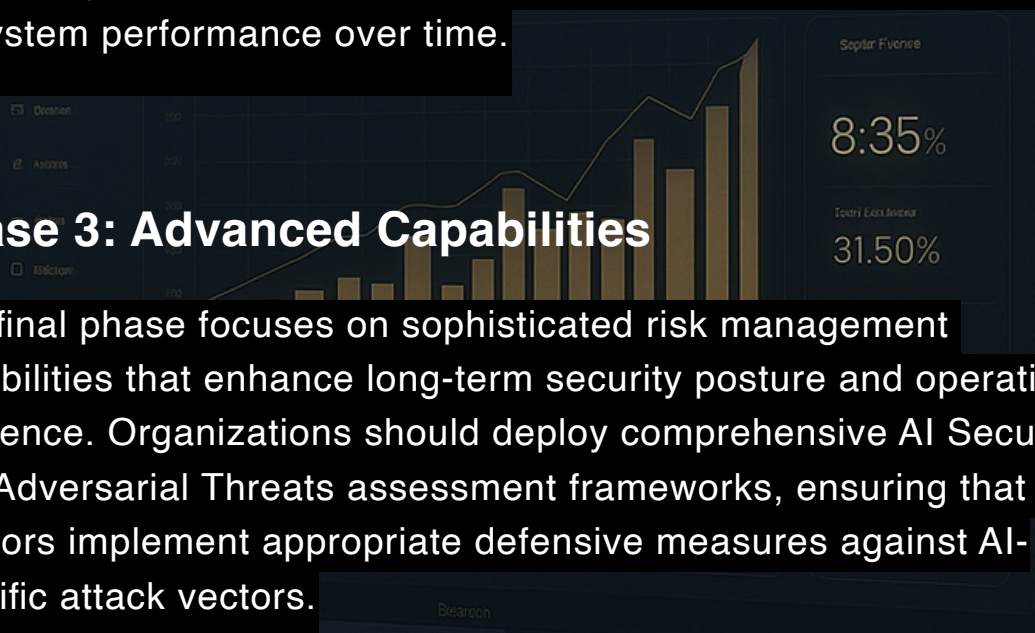
Phase 2: Strategic Capabilities

The second phase builds upon foundational controls by ensuring more sophisticated risk management capabilities that address fairness, compliance, and performance issues. Organizations should roll out comprehensive Bias, Fairness, and Non-

Discrimination testing protocols, establishing regular assessment processes and remediation procedures for identified bias issues.

During this phase, organizations must review third-party AI Regulatory and Ethical Compliance monitoring processes that track regulatory developments across relevant jurisdictions and ensure vendor compliance with evolving requirements. This includes establishing documentation standards, procedures for audit, & change management processes for regulatory updates.

The expansion phase concludes with ensuring Model Validation and Performance Drift detection capabilities, implementing automated monitoring systems and response procedures that ensure continued AI system performance over time.



Phase 3: Advanced Capabilities

The final phase focuses on sophisticated risk management capabilities that enhance long-term security posture and operational resilience. Organizations should deploy comprehensive AI Security and Adversarial Threats assessment frameworks, ensuring that vendors implement appropriate defensive measures against AI-specific attack vectors.

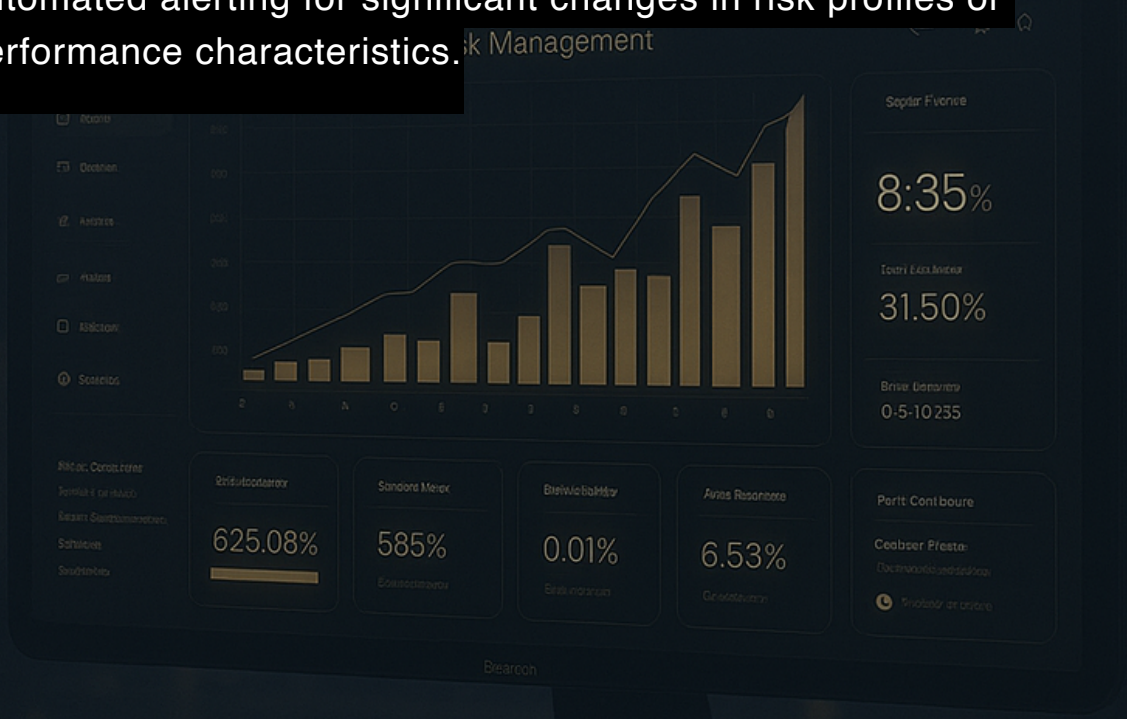
This phase includes checking Data Quality and Training Data Risk management processes that ensure ongoing reliability and fairness of vendor AI systems through rigorous data governance and quality assurance procedures.

The advanced implementation concludes with establishing Human-in-the-Loop Governance and Over-Reliance Risk controls that maintain appropriate human oversight and prevent unhealthy dependencies on AI systems.

Ongoing Operations and Continuous Improvement

Following initial due-diligence, organizations must maintain continuous oversight through ongoing monitoring that focus on regulatory changes, emerging risks, and evolving threat landscapes. Annual comprehensive reviews should examine all AI risk domains with updated control requirements and lessons learned from operational experience.

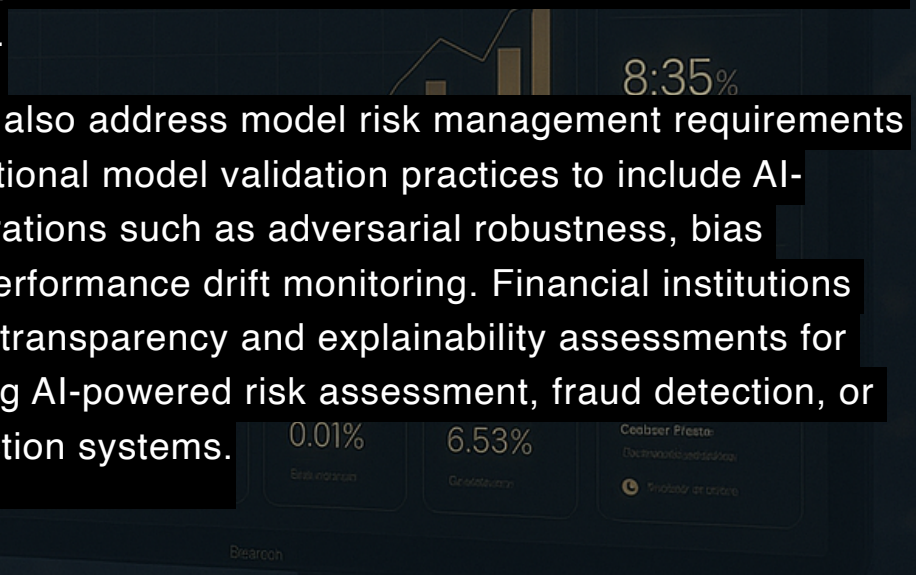
Continuous monitoring processes should track AI performance metrics and risk indicators across all third-party relationships, with automated alerting for significant changes in risk profiles or performance characteristics.



Industry-Specific Considerations

Financial Services Sector

Financial services organizations face unique regulatory requirements and risk considerations when implementing AI controls for third-party relationships. Banking regulators emphasize algorithmic fairness in lending decisions, requiring explainable AI implementations for credit scoring systems and enhanced due diligence for AI-powered fintech partnerships. Organizations must ensure that vendor AI systems comply with fair lending regulations, consumer protection requirements, and anti-discrimination laws while maintaining the efficiency and accuracy benefits of automated decision-making.



The sector must also address model risk management requirements that extend traditional model validation practices to include AI-specific considerations such as adversarial robustness, bias detection, and performance drift monitoring. Financial institutions should prioritize transparency and explainability assessments for vendors providing AI-powered risk assessment, fraud detection, or customer interaction systems.

Healthcare Industry

Healthcare organizations must address patient safety considerations, clinical decision transparency requirements, and HIPAA compliance obligations when working with AI vendors. The sector faces unique challenges in balancing AI innovation with patient protection, requiring enhanced oversight for AI-enabled medical devices, diagnostic algorithms, and clinical decision support systems.

Healthcare organizations should prioritize data privacy and security assessments for AI vendors, with particular attention to protected health information handling, consent management, and data minimization practices. The sector must also address liability and accountability considerations for AI-assisted clinical decisions while ensuring that human oversight remains appropriate and effective.

Technology Sector

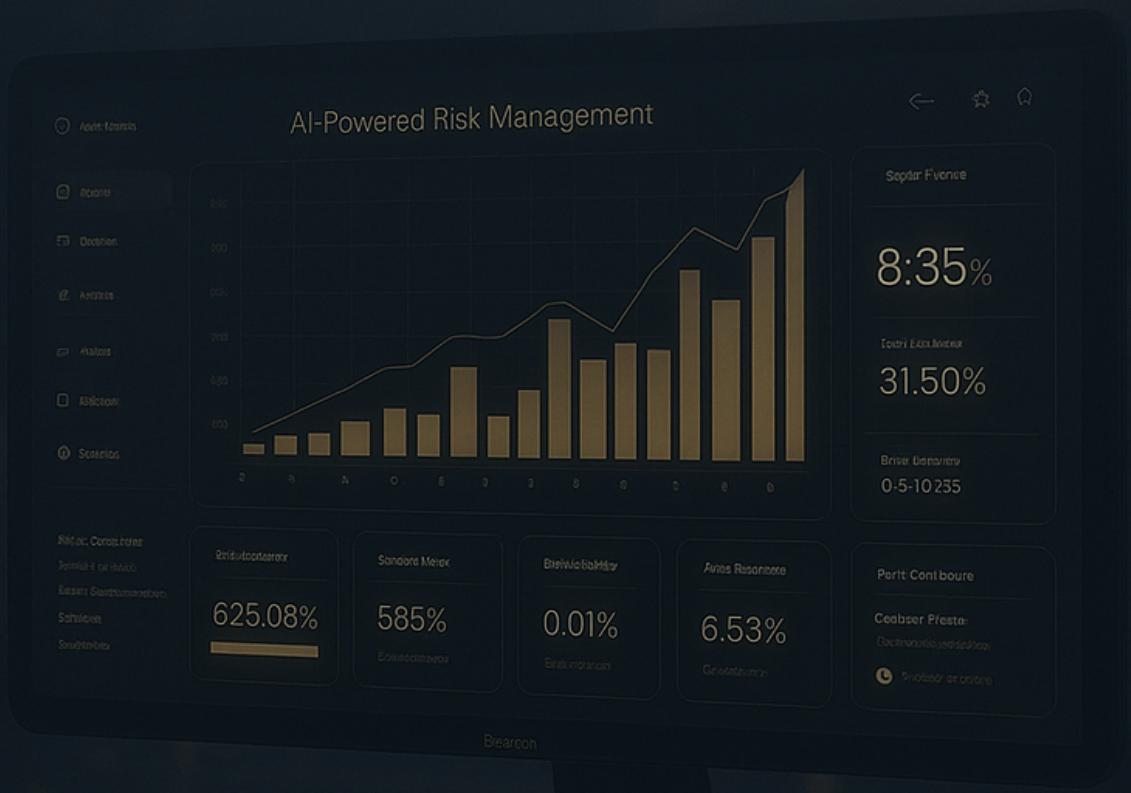
Technology companies face complex multi-jurisdictional compliance requirements and must address algorithmic transparency obligations across various markets and applications. The sector must implement comprehensive data privacy controls that address global privacy regulations while supporting AI innovation and competitive positioning.

Technology organizations should emphasize security and adversarial threat assessments for AI vendors, given the high-value targets these systems represent and the sophisticated threat actors that typically target technology companies. The sector must also address intellectual property protection and competitive intelligence concerns when working with AI vendors.

Retail and E-commerce

Retail organizations must address algorithmic bias in pricing and recommendation systems, customer data protection requirements, and fair lending practices for AI-powered financial services. The sector faces particular challenges in balancing personalization benefits with privacy protection and anti-discrimination requirements.

E-commerce companies should prioritize bias and fairness assessments for AI vendors providing pricing, recommendation, or customer service systems, ensuring that these systems don't create discriminatory outcomes or violate consumer protection regulations.



Conclusion

The integration of artificial intelligence into third-party business relationships represents both unprecedented opportunity and uncharted risk territory. As this assessment framework demonstrates, traditional Third-Party Risk Management approaches are insufficient for addressing the unique challenges posed by AI systems that can learn, adapt, and make autonomous decisions at scale.

The transformation of TPRM from static, point-in-time assessments to dynamic, AI-aware risk management requires fundamental changes in how organizations approach vendor relationships. The ten critical risk domains outlined in this guide—ranging from model transparency and bias management to adversarial threats and human oversight—collectively address the full spectrum of AI-specific risks that organizations must now manage throughout their supply chains.

The statistics are compelling: with 36% of data breaches originating from third-party compromises and 61% of companies experiencing third-party breaches annually, organizations cannot afford to ignore AI-specific risks in their vendor relationships. Organizations that delay implementation of AI-focused TPRM controls face escalating risks on multiple fronts. Customer expectations for algorithmic fairness and transparency are rising, particularly in sectors like financial services and healthcare where AI decisions significantly impact individual outcomes. Competitive pressures are increasing as organizations seek to leverage AI innovation while managing associated risks effectively.

The phased Assessment roadmap presented in this guide recognizes that organizations must balance thoroughness with practical constraints. The prioritization framework enables

organizations to focus initial efforts on the highest-impact risk areas while building comprehensive capabilities over time.

Organizations must recognize that AI risk management in third-party relationships is not a destination but an ongoing journey that requires continuous adaptation to evolving technologies, regulations, and threat landscapes. The framework and guidelines presented here provides a structured approach to this journey, but organizations must customize and adapt these guidelines to their specific industry context, risk tolerance, and operational requirements.

Success in AI-aware vendor risk management requires collaboration across multiple organizational functions, including risk management, legal and compliance, information technology, procurement, and business operations. The framework must be supported by appropriate governance structures, training programs, and technological capabilities that enable effective implementation and ongoing management.

Looking Ahead

As artificial intelligence continues to evolve, new risk categories will undoubtedly emerge. These challenges will not be simple third-party issues; they will be complex, cascading **Nth-Party** risks. An AI system's vulnerability is no longer just its own code but the code and data from its *entire* supply chain—dependencies that are often invisible. Organizations that establish robust AI risk foundations, ones that provide **Nth-Party visibility**, will be better positioned to adapt to these future challenges.

The investment in AI-focused risk capabilities represents more than mitigation—it enables organizations to engage confidently with AI-powered vendors **and their hidden dependencies**. This supports innovation and maintains competitive positioning in an AI-driven marketplace. Organizations that master these capabilities will be equipped to capitalize on the transformative potential of AI technologies throughout their **entire Nth-Party ecosystem**.

The time for action is now. The risks are real, the regulatory requirements are emerging, and the competitive implications are significant. Organizations that implement comprehensive AI risk management frameworks that **look beyond their immediate vendors** today will be the organizations that thrive in tomorrow's AI-enabled economy.

Appendix

AI Assessment Approach: High-Level Checklist

1. Define Assessment Scope and Objectives

- Identify in scope vendors, business units, and AI services and document the purpose of the assessment (onboarding, periodic, regulatory, or ad hoc).
- Work with procurement or IT teams to map all AI solutions used by third parties, especially those integrated into sensitive business processes.

2. Gather Foundational AI Governance Documents

- Request the vendor's AI policy, AI risk management framework, and related ethical standards to assess organization-wide commitment.
- Collect AI development and deployment standards, including SDLC, model training/testing criteria, and privacy-by-design statements.
- Obtain organizational charts and governance structures showing oversight for AI activities and relevant RACI matrices.

3. Review AI Integration and Implementation Evidence

- Documentation describing how AI models are embedded in business workflows, including impact analysis, model capabilities, intended use, and limitations.
- Review training/testing reports, validation logs, and data usage agreements (especially if your organization's data is utilized).
- Request self-attestation questionnaires and supporting evidence for claims on algorithm fairness, model explainability, and data protection.

4. Assess Regulatory and Compliance Alignment

- Examine the vendor's procedures for tracking and implementing regulations (EU AI Act, NIST AI RMF, SOC 2, ISO/IEC 42001, GDPR/DPDP).
- Review recent audit or regulatory exam reports if available, documenting lessons learned and corrective actions taken.
- Validate incident management policies for AI model failures or compliance breaches.

5. Evaluate Organizational Culture and Oversight

- Assess whether executive leadership supports and enforces AI governance through policy enforcement, resource allocation, and business alignment.
- Verify that AI risk is owned at the right level, and that culture supports ethical/secure AI use, not just technical compliance.

6. QA, Testing, and Model Performance Monitoring

- Request test cases and QA protocols demonstrating control over AI deployment sprawl; ensure controls exist to prevent unauthorized use or unintended process expansion.
- Review mechanisms for reporting, escalation, and remediation of AI misuse or drift outside intended business cases.

7. Prepare for Field Validation or Onsite Assessment

- Plan interviews and site visits with QA staff, developers, and compliance teams to validate documentation, claims, and observe governance in practice.
- Assess physical and system-level access controls protecting AI assets and sensitive data.

8. Synthesize Results and Establish Next Steps

- Summarize strengths, gaps, and risk themes; prioritize deep-dive reviews into sub-domains according to domain criticality, regulatory requirements, and exposure level.
- Develop an action plan, scheduling additional detailed evaluations in high-risk areas.

About Halbarad Risk Intelligence Inc.

Redefining Vendor Risk Management

AI risk is one of the 40 risk domains of the assessment framework developed by Halbarad, the first AI-native platform purpose-built for Nth-party risk management.

Unlike traditional solutions, which **check** direct vendors, Halbarad's deep, AI-powered engine provides unparalleled visibility into every layer of the supply chain, mapping not only immediate partners but also their subcontractors and beyond (Nth-Party). This gives organizations proactive control over risks that typically go undetected until a breach or compliance issue surfaces.

Currently, companies often struggle to keep up with the velocity of vendor onboarding, facing **3-6 months of** back-and-forth questionnaires just to finish a controls assessment. Halbarad transforms this experience, compressing tedious risk reviews into mere hours through intelligent automation.

The platform generates targeted assessments, auto-fills responses using public and proprietary data with coverage spanning 40 risk domains, ranging from compliance to emerging AI governance requirements. Halbarad's comprehensive framework not only accelerates due diligence but also ensures organizations remain resilient against evolving supply chain threats through effective ongoing monitoring.

Halbarad's AI-assessor capabilities and **built-in** assessment guidance mean organizations can eliminate dedicated TPRM teams with specialized skillsets to understand controls across **a** wide range of domains, from cybersecurity, ESG, **and** BCP to emerging **technologies** like AI and Crypto.

References

1. IBM Security. (2025). “Cost of a Data Breach Report 2025.” Statistic referenced: 36% of all data breaches originated from third-party compromises—a 6.5% increase from previous year.
2. Secureframe Blog. (2025). “110+ of the Latest Data Breach Statistics to Know for 2026.” Industry analysis on breach sources and impact.
3. Mitratech. (2025). “Third-Party Data Breaches: What You Need to Know.” Discussion of breach frequency and supply chain risk.
4. SecurityScorecard. (2025). “Global Third Party Breach Report.” Comprehensive third-party breach data and trends.
5. Sprinto Blog. (2025). “2025 Data Breach Report: Costs, Risks & AI-Driven Threats.” Current statistics and discussion about vendor risk & breach costs.
6. 8isoft. (2024). “2024 Top Third-Party Data Breaches and Lessons Learned.” Source for case studies and vendor risk incidents.
7. SCMR. (2024). “Analyzing the Supply Chain Risks Behind the Top Data Breaches in 2024.” Supply chain and vendor attack vectors.
8. IBM Insights. (2024). “Third-party breaches hit 90% of top global energy companies.” Insights on breach exposure by industry.

Author's Note: Research and synthesis for this whitepaper were supported by Perplexity AI.

About the Author

Shirish R. Korgaonkar is an accomplished risk management leader with over three decades of experience across Information Technology, Cybersecurity, and Risk Management. For more than a decade, he has specialized in designing and implementing comprehensive third-party and vendor risk management frameworks for modern enterprises. Shirish has guided organizations through the evolving landscape of vendor risk, business continuity, and cybersecurity. His expertise includes developing scalable control libraries, due diligence programs, building risk management and assessment products, and regulatory alignment strategies that enable enterprises to strengthen resilience across complex, multi-tier supplier ecosystems.