Third Party Risk Management Guide for AI Vendors

Initial Risk Assessment

Stratyfy x SOLASAI

Table of Contents

1 Introduction

2 Getting Aligned Internally

2 Setting your Organization up for Success

3 Identifying the Type of Technology

5 Questions to Ask AI Vendors

9 Definitions

11 About the Contributors

TPRM Programs Need to Match the Pace of AI Evolution

Many financial institutions recognize the strategic importance of AI, yet struggle to translate executive mandates into practical, secure vendor engagement.

Traditional third-party risk management (TPRM) practices were not built to handle the new risks that generative and autonomous AI introduces, including new dimensions of exposure related to bias, explainability, hallucinations, and training data integrity.

This guide is designed to close that initial governance gap, equipping business leaders and risk professionals with the questions needed to confidently begin assessing AI vendors and differentiating between different levels and focus areas of risk at the pace of AI evolution.

STEP 1

Define AI: Getting Aligned Internally

To effectively manage AI-related risks, it's crucial to first define distinct categories of AI. **The term "AI" is too broad to serve as a meaningful basis for differentiated governance and risk** [See Appendix for the definitions of different types of AI].

When using this guide, please make sure to consider the totality of the AI system components that the vendor is using and the associated risks for vendor assessment.

Critically, establishing shared definitions with the vendor at the outset is essential for mutual understanding and accurate communication about the technology being evaluated.

STEP 2

Know your Principles: Setting your Organization up for TPRM Success

Successful AI adoption begins with clear internal governance. **An effective TPRM program must be adaptable, yet anchored in the organization's core principles.** Examples of core principles include requiring transparency and explainability, or setting rules around risk tiering and defining high risk AI usage. These principles provide direction and continuity as regulations and technologies evolve, allowing compliance programs to adapt without losing alignment or clarity.

Your parameters, the standards that put principles into action, should evolve over time. Parameters define acceptable training data, documentation, human oversight, and performance thresholds. Set them by clarifying your institution's risk tolerance and non-negotiable requirements before engaging vendors. As AI and its risks continue to evolve, keep your assessments grounded in familiar, business-aligned risk categories. Evaluate risks in the legacy processes being replaced and how AI may introduce new risks or mitigate existing ones.

Over time, patterns will emerge across use cases to create more streamlined processes for TPRM for AI. A use case with no customer impact, ongoing human oversight, and limited financial exposure may be considered low risk and reviewed with less rigor. Similar cases can follow that streamlined path, while higher-risk use cases require greater scrutiny. Recognizing these patterns enables consistent risk tiering and scalable governance without reinventing processes and introducing time delays for each new system.

Vendors must be able to demonstrate that their controls align with your institution's standards. If the vendor cannot provide direct access to artifacts or model details, your TPRM process must at least require that the vendor maintain adequate documentation and governance to substantiate compliance.

STEP 3

Define your Track: Identifying the Type of Technology

The TPRM evaluator must clearly identify the type of AI technology the vendor is providing. This distinction, combined with the customer impact of the particular use case, is key to defining the solution's risk profile and the necessary depth and type of due diligence.

For this guide, due diligence is split into two tracks based on the AI system's primary function:

Track 1: Predictive / Decisioning AI - Applies to systems whose primary function is prediction, classification, or decision-making. This includes machine learning (ML) models and deep learning (DL) models, as they share the same fundamental risks under existing Model Risk Management (MRM) frameworks.

Track 2: Generative / Autonomous AI - Applies to systems designed for creation, synthesis, or autonomous operations. These AI systems introduce additional and emergent risks beyond traditional predictive systems and include Generative AI (GenAI) and Agentic AI.

Classification questions to determine the type of AI being provided

If the initial categorization is unclear to you, the questions below are mandatory for the vendor and will direct to the appropriate specialized track for the diligence questions that follow. Ask the vendor:

1) Does your solution rely on algorithms or models?

→ If yes, describe their primary purpose (e.g. prediction/classification vs. content generation/autonomous actions)
→ If predictive, proceed to Track 1
→ If generative or agentic → proceed to Track 2

2) Even if your primary system is ML-based, do you use any Generative or Agentic AI tools within your workflows that could impact our business operations?

→ If yes, Track 1 is required, and evaluators should determine whether the vendor must complete all or a subset of Track 2 and/or if the institution should require the vendor to fully indemnify the organization from any GenAI-related risk when the contracted solution itself does not involve generative functionality.

General questions to ask regardless of track

1. Are you providing off-the-shelf or proprietary models?
2. Are you following any AI governance or risk management frameworks? If so, which one(s)?
3. Do you provide transparency reports, audit rights, or documentation to verify your AI governance practices?
4. What contractual commitments do you make regarding liability, indemnification, and notification if AI-related risks materialize?

AI Diligence: Questions to Ask Vendors

Once you have assessed the type of AI and risk profile, you can begin to evaluate the vendor in the initial stage of diligence [See Appendix for detailed AI descriptions].

The specialized questions below are organized by the core risk domains as outlined by NIST AI RMF (Risk Management Framework) for Track 1 and NIST GenAI 600-1 (Generative AI Intelligence Profile) for Track 2. These frameworks provide a standardized means of organizing AI risks into the broadest and important characteristics that businesses need to address. The authors of the NIST AI Frameworks have drawn heavily from SR 11-7 as a basis for their work.

Track 1: Predictive / Decisioning AI

RISK	DEFINITION	TOP QUESTIONS
Valid & Reliable	Production of accurate, consistent, and reliable outputs, designed to achieve the intended purpose	• How did you test the model to prove it works accurately? • How do you monitor for and address performance degradation (data drift) over time?
Safe	Business continuity and positive user experiences	• What safeguards and testing are in place to prevent dangerous or incorrect decisions? • Does your system automatically flag predictions that are unusual or based on information it has never encountered before?
Secure & Resilient	Protection of data and systems that maintain confidentiality, integrity, and availability, and remain robust against attacks or failures	• What security testing have you performed to identify vulnerabilities to attacks by bad actors? • What defensive measures are implemented to detect and prevent these types of attacks? • What is your incident response plan if a security breach occurs, including your detection capabilities, containment procedures, and client notification timelines?

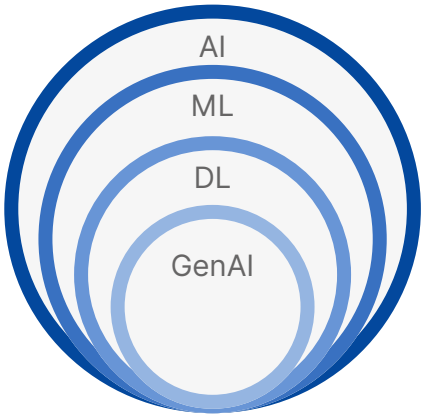
RISK	DEFINITION	TOP QUESTIONS
Accountable & Transparent	Governance practices that provide documentation, human oversight, and controls for AI/ML risks	• What artifacts can you provide to demonstrate the transparency of the technology?
Explainable & Interpretable	Effective design, monitoring, and auditing regimes that explain how and why a model is working	• What explainability methods does your model provide to show which input features influenced individual predictions? • Can you explain in business terms how your model makes predictions?
Privacy-Enhanced	Privacy and data protection compliance across all stages of a model lifecycle	• What sources and categories of personal or sensitive data does your model use as input or training data? Do you use PII? • How do you ensure compliance with data protection regulations (GDPR, CCPA, HIPAA) throughout the model lifecycle? • Will our data be segregated from other customers' data or used to retrain models that serve other organizations?
Bias Management	Identification and mitigation of data and/or decision-making bias	• What bias testing have you conducted, and can you provide performance metrics to demonstrate fairness? • What processes do you have for ongoing fairness monitoring in production, and what remediation procedures are triggered if bias is detected?

Track 2: Generative / Autonomous AI

RISK	DEFINITION	TOP QUESTIONS
Errors / Low Performance	Inaccurate or unreliable outputs such as false statements, coding errors, or hallucinations	• How do you address or prevent hallucinations (where the model generates false or nonsensical content), and what are your accuracy rates for our specific use case? • Do you provide tools for users to verify the accuracy of generated content?
Data Privacy	Unintentional leakage, unauthorized disclosure, or malicious de-anonymization of sensitive data	• How is our data used or stored? Is our data shared with others or used to train your model? • Can we delete our queries and inputs from your system?
Human-AI Interactions	Risks from user overreliance, misuse, and various forms of emotional entanglement	• How does your system indicate uncertainty and flag potential errors to ensure appropriate human oversight for high-stakes decisions? • What documentation and training materials do you provide to help our employees understand the system's limitations and avoid over-reliance on AI outputs?
Information Security	Compromise of confidentiality, integrity or availability of the system; or the abuse and misuse of GenAI for hacking (e.g., improved phishing or enhanced malware)	• What security testing have you conducted to identify vulnerabilities to prompt injection, data poisoning, model extraction, and other GenAI-specific attacks? • What is your incident response plan if your system is compromised or used to facilitate a cyberattack against our institution?
Intellectual Property	Creation or distribution of content that infringes on existing copyright, patents, trademarks, or trade secrets	• Can you warrant that your training data was lawfully obtained and does not infringe on copyrights, trademarks, or other intellectual property rights?

RISK	DEFINITION	TOP QUESTIONS
Third Parties / Supply Chain	Exposure to risks from external data sources, open-source tools, or vendors with limited transparency or assurance	• Can you provide complete transparency about all third-party components in your system, including their sources, licenses, and any known vulnerabilities?
Bias & Disparities	Generation of unfair, offensive, or homogenized content that reflects unequal representation across languages, cultures, or groups	• What specific testing have you conducted to identify and mitigate biases, and can you provide performance metrics to demonstrate fairness? • When using synthetic data, how do you monitor for and prevent the model's performance from degrading (data drift) or its outputs from becoming overly uniform (homogenization)?
Mis- and Dis-information	Generation and distribution of false, misleading, or manipulated content such as deepfakes	• Do you have methods to help users distinguish AI-generated content from human-created content?
Dangerous or Criminal Recommendations	Generation and distribution of content that is violent, threatening, or radicalizing or that provides instructions for criminal activities or self-harm	• What guardrails are in place to prevent the generation of violent, threatening, hateful, or harmful content, and how frequently are these tested? • How do you monitor for, detect, and respond to attempts to use your system to create harmful materials?

Definitions*



Artificial Intelligence (AI): AI represents the entire field of computer science dedicated to creating systems that can simulate human intelligence to perform tasks like problem-solving, decision-making, and understanding language.

Model: The output of a training process that has learned patterns, rules, and relationships from data to perform a specific task, such as making predictions, classifications, or generating new data. It encapsulates the knowledge gained during training and is used to process new, unseen inputs to produce a desired output.

- In financial services, typical examples include models that predict attrition, default, fraud, and AML.

Machine Learning (ML): Refers to computer systems that identify and learn patterns from data to improve their performance on specific tasks, without being explicitly programmed for every rule or outcome.

- In financial services, ML is commonly applied in fraud detection, credit risk modeling, marketing analytics, and other activities that rely on recognizing complex data patterns.

Deep Learning (DL): A subset of machine learning that uses multi-layered neural networks to identify complex patterns in large amounts of data, such as text, speech, images, or transaction records.

- In financial services, deep learning supports applications like voice authentication, document and identity verification (KYC), transaction monitoring for anomaly or fraud detection, and automated document processing.

* Definitions are aligned with established frameworks: **ISO/IEC 22989:2022**, *Artificial intelligence* and **NIST AI RMF 1.0**, Risk Management Framework.

Generative AI (GenAI): Systems trained to create new content, such as text, code, images, or audio, based on patterns learned from existing data. **Foundation models** are the large, general-purpose "engines" that power a broad ecosystem of generative AI applications. A foundation model is any model that is trained on broad data that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks.

- In financial services, generative AI is used for applications like client chatbots, summarizing or drafting reports, assisting with code development, generating customer communications, and supporting staff through knowledge or research assistants. Foundational models are typically used directly by analysts or developers, for chatbots, or built into productivity apps.

Agentic AI (AAI): Systems that can autonomously pursue goals or complete tasks by planning, reasoning, taking actions, and adapting to feedback; often by coordinating multiple models or tools. Unlike standard AI systems that respond to single prompts or inputs, agentic AI systems chain together actions, use memory or context, and may interact with external systems or data sources to achieve an outcome.

- In financial services, investment is being directed towards agentic AI for applications such as fraud detection and intervention, dynamic credit assessment, and customer servicing.

About the Contributors

Stratyfy is an industry leader in decision optimization for financial institutions, empowering lenders to make more accurate, efficient, and fair decisions using transparent ML. The fintech was named a 2024 Banking Tech Awards USA winner, 2024 Benzinga Fintech Awards Best Lending Solution finalist, and AIFintech100 honoree in 2023 and 2024. Stratyfy's patent-pending technology is trusted by top-10 U.S. banks, core providers, and industry leaders in responsible AI governance and policy.

SolasAI leverages over 75 years of collective experience in providing AI-based compliance SaaS and machine learning solutions to address bias and discrimination at enterprise-grade scale. SolasAI's methodologies are used by more than half of Fortune 50 companies and provide the only end-to-end, industry-agnostic AI software solution helping banks, fintechs, insurers, and healthcare companies achieve fairer outcomes for their customers by fixing their models rather than replacing them.

Have questions about this guide?

Laura Kornhauser
CEO & Co-Founder | Stratyfy

info@stratyfy.com

Deniz Johnson
COO & CIO | Stratyfy

info@stratyfy.com

Larry Bradley
CEO | SolasAI

larry.bradley@solas.ai

Nicholas Schmidt
CTIO | SolasAI

nicholas.schmidt@solas.ai