



# AI Governance and Risk Management

A Practical Guide to effectively  
governing and managing AI Risk

AI is no longer an experimental add-on; it is a production capability that reconfigures how banks create value and concentrate risk. Governing AI as a checklist or a point solution is a strategic error: regulators will expect demonstrable evidence, customers will want reliable outcomes, and vendors will continue to introduce new risks. Banks must therefore treat AI as an enterprise discipline — governed by use case, scaled by materiality, and instrumented by immutable evidence and real-time telemetry — so they can move fast without sacrificing control. This document sets forth a concise, pragmatic framework that banks can deploy immediately to govern AI responsibly and scale it safely.

**Shweta Jain**

Partner, Risk & Compliance Head  
IBM Promontory

## Key Highlights

- **Treat AI as an enterprise capability:** Governing AI effectively requires (i) placing the unit of governance at the level of the AI use case (model + system + deployment), (ii) applying a materiality-based control model that aligns oversight intensity to impact, and (iii) instrumenting governance with immutable evidence and real-time telemetry.
- **Consider AI differs from traditional software and statistical models:** Non-deterministic outputs, reliance on large and often opaque third-party models, evolving behavior (drift), and new failure modes (hallucinations, prompt injection, and agentic actions). Effective governance therefore uses established risk disciplines—model risk, operational risk, vendor management—augmented with AI-specific practices, metrics, and tooling.
- **Govern by use case, not by model:** Risk is a function of context. A foundation model used for internal summarization presents different risks than the same model applied to credit underwriting.
- **Scale controls to materiality:** Resources must be focused on the use cases that can cause material financial, regulatory or reputational harm.
- **Evidence drives defensibility:** Supervisors and auditors will be evaluating documentary and telemetry evidence. Policies without demonstrable implementation will be insufficient.
- **Develop layered defenses:** Deterministic controls, monitoring, and human judgment each address distinct failure modes; they should operate in concert.
- **Enable experimentation, constrain production:** Guarded sandboxes with clear production gates allow innovation without regulatory exposure.
- **Treat third parties as integral to the bank’s program:** Vendor opacity is a primary source of residual AI risk; contractual, technical and operational levers must be configured accordingly.

## AI Governance and Risk Management

### OPERATING MODEL

- **Embed AI oversight and governance effectively in the existing organizational structure.** Setting appropriate delegations of authority and escalations into the board and senior management committee structure is key to effective oversight.
- **Align materiality definitions with the relevant enterprise frameworks** and set up appropriate intake gating for all new use cases that meet materiality thresholds.
- **Three Lines of Defense:** 1LoD builds and operates; 2LoD validates and challenges; 3LoD audits. Define clear roles across the three lines of defense.

- **Incorporate relevant changes to roles and responsibilities** to remove ambiguity on classification, decision rights, deployment holds, and rollback authority.
- **Board’s involvement and role** should entail approving enterprise AI framework, risk appetite and receiving periodic materiality and incident reporting.
- **Senior Management** should oversee and sponsor resourcing, and ensures cross-functional integration (Risk, Legal, Technology, Business). Increasingly firms are opting to appoint a dedicated Chief AI Officer.
- **AI implementation oversight requires dedicated senior management committee**, i.e., AI Council: a management committee with representation from Model Risk, Compliance, InfoSec, Business Heads and the AI Centre of Excellence (CoE). The council adjudicates materiality disputes, approves standards, and triages escalations. The CoE would provide toolkits, templates, training, and technical enablement; runs the intake system and the sandbox environment.

## RISK FRAMEWORK

- **Enterprise Risk Framework** should be expanded to incorporate AI-related Risk considerations across the existing risk taxonomy and processes. Develop enterprise-level risk measures for AI and calibrate AI “risk appetite” across relevant risks.
- **Risk Inventory Extension:** Add AI-specific subcategories under enterprise risks (Model Risk, Operational Risk, Cyber, Third-Party, Data, Conduct, Reputational, Regulatory) but also consider new risks and associated metrics of AI Deployment.
- **Expand existing frameworks and processes** i.e., model development and validation, 3rd party AI usage inventory and monitoring, operational risk event/incident management and loss calculations, regulatory obligations scanning, and regulatory change management to include AI-related scope, coverage, and controls. Deployment Risk and Data Risk should also be considered.
- **Risk appetite:** Calibrate AI considerations in your existing risk identification and measurement processes. Consider key AI KRIs (e.g., hallucination rates, prompt-injection success thresholds, vendor

concentration limits) for use cases in each division and function.

- **AI Deployment Risk:** AI Deployment Risk is the set of real-time, runtime risks that arise from how a model/system is integrated, executed, scaled and operated in production AI Deployment Risk is distinct from model, cyber and other existing risks due to the following:
  - *Time sensitivity:* Deployment failures manifest in minutes/hours and require operational playbooks and automated mitigation (canaries, rollbacks), unlike some model validation findings that are addressed offline.
  - *Environment dependency:* Production behavior depends on integration, pipelines, retrieval corpora, user prompts, third-party updates, and infrastructure — risks that do not appear in isolated model tests.
  - *Emergent interactions:* Multi-component systems (RAG, agents, tool integrations) create cascading or correlated failures that cannot be detected by model-only validation.
  - *Human factors:* Real users, volume, and workflows change exposure and amplification of errors (scale and reputational impact).
  - *Monitoring requirement:* Deployment risk demands continuous telemetry, alerting thresholds, and fast remediation SLAs rather than periodic validation alone.
- **Measures for AI Deployment Risk**
  - *Observability:* Log inputs, prompts, model/version, retrieval context, outputs, decisions, and reviewer actions to an immutable store; correlate logs with business events and customer complaints.
  - *Real-time telemetry & KRIs:* Monitor hallucination rate, prompt-injection attempts & success rate, retrieval coverage, latency, fallback usage, output rejection rates, vendor health, and API error rates. Define thresholds and automated runbooks.
  - *Canary + feature flags:* Use canary deployments, traffic throttling, and feature toggles to limit blast radius for model or pipeline changes.
  - *Automated rollback & kill switches:* Predefine rollback criteria; ensure runbooks and automation can isolate problematic versions within minutes.
  - *Dependency mapping:* Maintain a live dependency graph (models, indexes, agents, vendors, APIs) and run “what-if” impact analysis for vendor outages or dataset corruption.

- Human-oversight design: Define “human-in-the-loop vs. “human-on-the-loop” rules by tier, including sampling methodology, decision authority and surge capacity for reviews during incidents.
- *Adversarial and resiliency testing*: Regular red-team exercises focused on runtime attacks (prompt injection, extraction, data poisoning) and infrastructure stress tests.
- *Incident management & evidence pack*: Standardize incident intake with required artifacts (immutable snapshots, KRI snapshot, model card, validation report, RCA and remediation plan). Treat near-misses as precedence for control tuning.
- *SLAs & contingency*: Contractually require third-party vendors to provide change notices, telemetry, rollback support, and tested contingency plans. Maintain alternative providers or fallback models.

AI Deployment metrics are key to effectively managing existing risks, including the risks below and their associated considerations:

Topic	Considerations
<b>Model Risk</b>	<ul style="list-style-type: none"> <li>• Consider LLMs<sup>1</sup>, RAG systems, prompts and pipelines as model components. Validate conceptual soundness, data fitness, fairness, and sensitivity to prompt variants</li> <li>• Independent 2LoD validation for Medium+/Material models.</li> <li>• Version control, back-testing, challenger models, and re-validation triggers for material change</li> </ul>
<b>Cyber Risk</b>	<ul style="list-style-type: none"> <li>• Harden inference endpoints (private networks, mutual TLS)</li> <li>• Implement prompt/input sanitization and output filtering.</li> <li>• Deploy prompt-injection and extraction detectors, honeypots, and red-team exercises</li> <li>• Include AI telemetry in SIEM; enforce patch &amp; incident SLAs</li> </ul>

<b>Third-Party Risk</b>	<ul style="list-style-type: none"> <li>• Tier vendors by criticality; require model cards/AI Cards, change notification, rollback rights, and telemetry sharing for critical suppliers</li> <li>• Contractual flow-down to subcontractors, right to audit/attestations, and tested contingency/exit plans</li> <li>• Continuous vendor monitoring and quarterly deep dives for material vendors.</li> </ul>
<b>Operational Risk</b>	<ul style="list-style-type: none"> <li>• Update RCSA and control design and tests. AI usage governance may require new controls</li> <li>• Map dependencies and fallbacks; require runbooks, human escalation paths, and SLAs</li> <li>• Canary releases, feature flags, and rollback capability</li> <li>• Define and track near-misses related to AI usage and define loss measures</li> </ul>
<b>Compliance Risk</b>	<ul style="list-style-type: none"> <li>• Consider both existing and emerging regulations, (e.g., NIST AI RMF, ISO/IEC 42001:2023, SR 11-7, MAS FEAT Principles)</li> <li>• Embed in the existing regulatory change management processes by mapping obligations to use cases and monitor regulatory change</li> </ul>
<b>Data Risk</b>	<ul style="list-style-type: none"> <li>• Enforce data lineage, provenance metadata, and fitness-for-purpose assessments</li> <li>• Protect PII with encryption, minimization, and consent checks, log dataset use</li> <li>• Monitor for drift/contamination in retrieval corpora and retraining datasets; retain provenance snapshots for audits</li> </ul>

<sup>1</sup> An LLM as a standalone artifact is a model; more relevant for banks is the system/use-case that embeds that model.

## STRESS TESTING

- **Scenario Design:** Build a focused AI stress-scenario library and map impact paths. Define plausible, high-value scenarios (e.g., mass hallucination in a customer-facing model, RAG index contamination, targeted prompt-injection leading to PII leakage, supplier foundation-model outage, agentic runaway performing unauthorized transactions). For each scenario consider triggers, propagation chains (which downstream systems, channels, or customers are hit), likely regulatory exposures, and qualitative severity tiers (operational, financial, reputational, regulatory).
- **Sensitivity Analysis:** Develop appropriate assumptions and run sensitivity analyses. For each scenario state assumptions (detection delay, percent of traffic affected, success rate of attack, time to rollback). Convert those into quantitative knobs (e.g., hallucination rate × user volume × average remediation cost) and run sensitivity sweeps to show how losses and customer impact change when key unknowns move. Map dependencies and cascading failure logic. (concentration risk). From the maps derive concrete mitigations: fallback model, read-only mode for indexes, pre-approved manual workflows, and vendor diversification thresholds.
- **Stress Testing:** Test both qualitatively (tabletops) and quantitatively (simulations & chaos). Run cross-functional tabletops with tech/cyber/legal/comms/business to validate assumptions, communications and escalation. Complement with technical exercises: adversarial red-team campaigns, synthetic data contamination tests, canary/circuit-breaker deployments, etc. Ensure tests capture detection, containment, rollback and human surge capabilities — and that log/evidence capture works end-to-end.
- **Capital and Contingency Planning:** Convert outcomes into governance actions and resilience investments. Feed stress results into capital and contingency planning where quantified exposures are material. Translate findings into prioritized remediations (telemetry upgrades, additional fallback capacity, etc.). Prepare evidence materials for material scenarios (assumptions, simulations, RCA template, remediation plan) and bring to AI Council and senior management for review.

## TECHNOLOGY

Technology can be used to manage AI Governance and Risk effectively – below are some examples:

- **Automated discovery and a synchronized golden record:** Technology can continuously detect AI assets (models, prompts, embeddings, RAG indexes, agent workflows) across clouds, code repos and APIs and reconcile them into a single inventory. Practically this means scheduled scans, API hooks to CI/CD and vendor portals, and automated alerts for unregistered assets—so you can block production traffic from unknown models. *Guardrail: require human verification and policy approval for any automated classification changes to prevent false positives/negatives from becoming control failures.*
- **Real-time observability, anomaly detection and KRI automation:** Instrument inference pipelines and user channels to stream telemetry (inputs, prompts, model version, retrieval context, outputs) into analytics engines that compute KRIs (hallucination rate, injection attempts, drift, PII flags) in near real time. Use ML detectors and rule engines to surface anomalous patterns, cluster incident signals, and auto-prioritize alerts for human triage. *Guardrail: calibrate thresholds, measure false negative rates, and ensure critical alerts escalate to named humans with enforced SLAs.*
- **Automated validation, adversarial testing and continuous “health checks”:** Use orchestration to automate pre-deploy test suites (factuality, fairness, sensitivity, extraction probes) and scheduled adversarial/red-team runs in sandbox environments; automatically compare results to baseline thresholds and block or quarantine failing builds. Practically this creates repeatable, auditable validation runs for every vendor update, index refresh, or prompt change. *Guardrail: keep independent validation and final sign-off in 2LoD; treat automated passes as necessary but not sufficient evidence.*
- **Policy-as-code, CI/CD gating and fast rollback automation:** Encode governance rules (materiality gates, required artefacts, KRI thresholds, allowed vendors) as executable policies in build/deploy pipelines so non-compliant changes are prevented automatically. Couple this with canaries/feature flags and scripted rollback/kill-switch actions to reduce blast radius when runtime telemetry trips. *Guardrail:*

*require human override logs and make rollback actions themselves auditable and reversible.*

- **Explainability, provenance and evidence automation for audit/readiness:** Automatically assemble regulator-grade evidence bundles: model and prompt versioning, retrieval index snapshots, validation reports, KRI history and immutable input/output logs. Leverage explainability modules (surrogates, source-citing for RAG) to produce human-consumable rationale on demand so reviewers can triage faster. *Guardrail: preserve immutable snapshots separate from production edits and require independent review of automated explanations before they are relied upon in high-impact decisions.*

## WORKFORCE AND TALENT MANAGEMENT

- **Competency Management:** Define clear competency profiles (e.g., Model Owner, Validator/2LoD Reviewer, Agent Supervisor, Platform Engineer, Business User) with mandatory proficiency levels. Role-based competency gates, not one-size training. Require role-specific certification before access to production (e.g., “validator certified” for independent sign-off on Medium+/Material models). Make certifications evidence-based: live scenario assessments, red-team participation, and a capstone review. Enforce recertification cadence.
- **Human Capital Planning:** Prioritize humans where risk concentrates; staff to measured capacity. Allocate human oversight and validator resources by materiality: more frequent sampling, deeper review and on-call surge capacity for material models. Set target staffing ratios (e.g., one validator per X material models, human reviewers sized to handle peak HITL load with defined SLA). Maintain a small “rapid response” pool trained for incident surge and audits.
- **Training and Upskilling:** Train by doing – sandbox, simulation & role play instead of theoretical training. Replace lecture hours with hands-on sandboxes linked to live telemetry (synthetic but realistic data), tabletop incident drills, adversarial-attack simulations and AI governance board role-plays. Measure training effectiveness by activation. Manage talent via rotation, hybrid sourcing and career paths. Combine upskilling of existing staff

(Builders → Validators) with targeted hires for deep technical, adversarial, and legal expertise; formalize rotation/programs to build cross-functional experience. Create clear career ladders (AI Risk Specialist, Agent Supervisor) and link promotion/compensation to demonstrated governance outcomes (quality of validations, remediation closure rate).

- **Performance Management:** Embed AI considerations in the existing performance management framework. Evaluate based on outcomes and auditability, not completion hours. Track operational KPIs tied to workforce effectiveness such as, % material models with certified validator, average time to independent validation, validation defect rate, human override accuracy and timeliness, and training activation metrics (simulation detection rate). Feed these into performance reviews and the AI Council’s resourcing decisions; use audit findings and incident root causes to close capability gaps with targeted retraining.

## Action Plan: What Banks Should Do Now?

### 30 DAY PLAN – STABILIZE

- **Design and implement a fit-for-purpose AI Oversight Structure** – not all organizations have the same risk profile. Consider your organization (i.e., global and large vs. regional and specialized) and develop an oversight model which embeds oversight in the existing governance structure and stands up required new committees and COEs.
- **Develop a customized AI Framework and amend required existing frameworks / policies** – frame AI risk appetite and strategy and amend existing policies to reflect AI-related considerations. Develop an AI framework across 1LoD and 2LoD and establish roles and responsibilities.
- **Lock down discovery & intake of AI usage:** Create the golden record. Deploy a mandatory intake form and register gate (simple portal or workflow) that collects owner, purpose, risk tier, vendor, data sensitivity and initial KRI set. Enforce policy that any unregistered AI cannot be promoted to production.

### 90 DAY PLAN – DEFEND

- **Stand up the top-of-stack telemetry & KRI dashboard:** Instrument production endpoints (or sample logs for hosted services) to produce baseline values for 4–6 priority KRIs (hallucination flags per

10k, prompt-injection attempts, PII detection rate, % material models with logging). Publish a weekly report to the AI Council.

- **Operationalize pre-deploy validation and gating for Medium+/Material Use Cases:** Develop and publish a use-case validation checklist (conceptual soundness, data/grounding fitness, factuality/hallucination tests, bias/fairness checks, adversarial probe checks). Integrate the checklist into gating so Medium+/Material builds cannot promote without appropriate sign-offs.
- **Enable immutable logging and automated alerting for material use cases:** Configure tamper-evident logging for material models (inputs/prompts + model version + retrieval context + outputs + reviewer actions) and connect to SIEM/monitoring for KRI alerts (auto-notify owner & SOC on threshold breaches).
- **Triage vendors and begin contractual & capability fixes;** launch role-based certifications: Conduct a vendor criticality triage (top vendors by volume/criticality), issue an AI Card / information request for Material suppliers, and negotiate basic change-notice + rollback clause amendments for critical vendors where possible. Launch role-based certification program with certification for production privileges.

### 180 DAY PLAN – OPERATIONALIZE

- **Automate safe deployment controls and rapid remediation paths:** Implement canary and feature-flag patterns for Material models, automated rollback runbooks (scripted rollback and kill-switch), and a tested incident playbook tied to KRIs (including forensic snapshot generation). Ensure rollback/kill actions are auditable and have a human escalation path.
- **Build adversarial testing, stress scenarios and assurance loop:** Launch scheduled red-team/adversarial program for material models and ingestion pipelines; incorporate 3–4 AI stress scenarios (mass hallucination, RAG contamination, vendor outage, prompt-injection at scale) into enterprise stress testing and business continuity plans. Feed findings into remediation and capital/resilience planning.
- **Institutionalize governance, staffing and audit readiness:** make the AI Council cadence

operational, certify the validator and agent-supervisor population to production access standards, and develop evidence bundles (model cards, validation reports, immutable logs, KRI history, RCA templates) for top material use cases.

## How IBM Promontory Can Help

IBM Promontory is a leading advisory firm staffed by former senior officials of regulators and industry practitioners. IBM Promontory's regulatory depth combined with IBM Consulting's implementation capability and IBM Technology, particularly WatsonX, are leading the conversation on AI Risk and Governance. IBM can help your organization with the following:

- ✓ Readiness Assessment
- ✓ AI Governance and Training:
- ✓ AI Risk Framework and Risk Appetite Development
- ✓ Model Risk Management
- ✓ Third Party Risk Management
- ✓ Cyber Risk Management
- ✓ Compliance Framework
- ✓ Regulatory Change Management
- ✓ IBM WatsonX
- ✓ IBM AI Solutions

## About the author



### **Shweta Jain**

Partner  
Risk & Compliance Head  
IBM Promontory

## Contributors

### **Manish Goyal**

IBM Consulting,  
Enterprise AI Strategy &  
Governance Global  
Offering Lead

### **Miles Ravitz**

IBM Promontory  
Associate Partner

### **James Pastro**

IBM Promontory  
Associate Partner

© Copyright IBM Corporation 2026

Produced in the  
United States of America  
March 2026

IBM, the IBM logo, BM Trademarks List are trademarks or registered trademarks of International Business Machines Corporation, in the United States and/or other countries. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on [ibm.com/trademark](http://ibm.com/trademark).

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS" WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT