ISSN: 2756-2638
Transactions on Engineering and Computing Sciences; Vol. 14 No. 05 (2026) (19-62)
https://doi.org/10.14738/tecs.1405.12205

From Model Risk Management to AI Assurance: Integrating AI Risk Management, Independent Validation, and AI Auditing for Generative and Agentic AI

Tim Godlove[1], & John Buchanan[2]

1. Long View Review, Virginia, United States
2. Yddrasil, Montana, United States

Abstract: Generative and agentic artificial intelligence challenge assumptions embedded in traditional model risk management (MRM), including stable system boundaries, direct observability, controllable change, and access to model-development evidence. Yet the enduring purposes of MRM—inventory, materiality assessment, independent validation, monitoring, documentation, effective challenge, and accountable governance—remain necessary. This conceptual and applied article develops a socio-technical assurance architecture for contemporary AI systems by integrating governmental and regulatory guidance, international standards, professional assurance practices, and scholarly literature. The proposed Integrated AI Assurance Framework connects four distinct but interdependent functions: AI governance; AI risk management and MRM; independent AI/model validation; and AI auditing. The article also introduces the AI Assurance Boundary, which defines the technical and organizational components whose evidence is material to justified reliance on an AI-enabled capability, and the Assurance Sufficiency Principle, which addresses circumstances in which complete transparency, traceability, or explainability is unattainable. The resulting architecture is risk-based, with assurance intensity varying according to materiality, consequentiality, autonomy, data sensitivity, regulatory exposure, observability, topology, and third-party dependence. Lending, employment, and agentic-enterprise illustrations demonstrate how the framework supports accountable ownership, effective challenge, evidence, escalation, and independent assurance.

Keywords: AI assurance, agentic AI, generative AI, independent validation, model risk management, trustworthy AI

INTRODUCTION

Artificial intelligence (AI) has evolved from statistical models and purpose-built machine learning systems into an expanding ecosystem of deep learning, generative AI, foundation models, retrieval-augmented generation (RAG), and increasingly autonomous agentic AI. This evolution is changing not only what AI systems can do, but also how organizations must govern, validate, monitor, and independently assure their use. Traditional models were generally developed or acquired for defined purposes, operated within identifiable system boundaries, and evaluated against relatively stable assumptions, inputs, outputs, and performance measures. Generative and agentic AI challenge many of these characteristics. Their behavior can be probabilistic and context-dependent; their capabilities may originate from externally developed foundation models; their outputs may depend on dynamically retrieved information; and agentic systems may be authorized to interact with other systems or initiate actions with varying degrees of human intervention [1], [2].

These developments present an important challenge for model risk management (MRM). Traditional MRM provides durable principles for model identification and inventory, risk classification, independent validation, performance monitoring, change management, documentation, and effective challenge. These principles remain relevant to AI. However, their application becomes more difficult as AI systems become less transparent, more externally dependent, more adaptive, and more autonomous. The central issue is therefore not whether traditional MRM remains useful, but whether MRM alone provides a sufficiently broad structure for managing and assuring the risks of contemporary generative and agentic AI [3], [4].

The location, ownership, and access model of AI further complicate this challenge. Organizations may develop and host AI internally, operate internally hosted large language models (LLMs), consume externally provided foundation-model services, or use AI embedded within enterprise software and cloud platforms. Each arrangement creates different governance and assurance conditions. Organizations may remain accountable for AI-enabled outcomes while having limited visibility into model architecture, training data, model weights, development practices, or data lineage. Organizational data may also cross boundaries through prompts, application programming interfaces, RAG architectures, agents, and third-party services. Consequently, where AI resides, who operates it, how it is accessed, what data it can reach, and what actions it can perform have become important determinants of risk and assurance requirements [1], [5], [2].

AI risk also crosses traditional organizational boundaries. MRM, enterprise risk management (ERM), IT risk, cybersecurity, third-party risk management, privacy, legal and compliance functions, AI governance, independent validation, and internal audit may each have legitimate responsibilities for aspects of an AI system. Yet overlapping responsibilities do not necessarily produce comprehensive assurance. Without clearly defined ownership and decision rights, organizations risk both control duplication and control gaps. One function may evaluate technical vulnerabilities, another model performance, another vendor risk, and another the effectiveness of the control environment, while no single function has responsibility for integrating these perspectives [6], [7], [8], [9].

Generative and agentic AI therefore raise a broader question about the boundary between risk management and assurance. Organizations must determine not simply whether an AI system qualifies as a model under an existing MRM program, but which risks must be managed, which systems require independent validation, what evidence is necessary, who provides effective challenge, and who independently assesses whether governance and controls are operating as intended. The evolving treatment of generative and agentic AI within U.S. federal banking MRM guidance provides an important example of this changing boundary. Although financial services offer a mature model-risk environment, the underlying challenge extends across industries as organizations deploy AI capabilities whose risks span technology, models, data, cybersecurity, compliance, operations, and third-party relationships [3], [1], [10].

Organizations are not without guidance. Governmental, regulatory, professional, and international frameworks increasingly address aspects of AI governance and assurance, including the National Institute of Standards and Technology (NIST) AI Risk Management Framework and Generative AI Profile, ISO/IEC AI management and risk standards, model risk guidance, sector-specific requirements, and internal-audit approaches to AI. The challenge

is translating these overlapping expectations into organizational practice. Frameworks may describe what organizations should accomplish, but organizations must still determine who owns each responsibility, how functions interact, how independence is preserved, what evidence is required, and how exceptions and disagreements are escalated. This creates an emerging AI assurance gap between the existence of governance and risk frameworks and their coordinated implementation within organizations [6], [1], [11], [12], [13].

This article addresses that gap through the following primary research question:

- How can organizations integrate AI risk management, model risk management, independent validation, and AI auditing to create an effective assurance framework for generative and agentic AI systems?

A secondary research question examines the associated organizational challenge:

- What challenges do organizations face in governing, validating, and independently assuring generative and agentic AI when traditional model risk management frameworks do not adequately address these technologies?

To address these questions, this article develops an Integrated AI Assurance Framework that connects AI governance, AI risk management and MRM, independent AI/model validation, and AI auditing while recognizing the complementary roles of ERM, IT risk, cybersecurity, third-party risk management, legal and compliance functions, and governance or steering committees. The framework distinguishes risk ownership and management from effective challenge and independent assurance, clarifying where these functions should coordinate and where organizational independence and accountability should be preserved.

From an engineering and computing perspective, the assurance problem is socio-technical rather than model-only. The relevant system may combine a foundation model with retrieval infrastructure, enterprise data, APIs, identity and access controls, policy layers, agent tools, cloud services, monitoring, and human decision points. The paper therefore treats assurance as a system-level computing problem: defining the technical and organizational boundary over which claims must be supported, identifying observable and unobservable dependencies, and establishing evidence sufficient to validate and govern the complete AI-enabled capability.

The framework is intentionally not presented as a universal organizational model. AI assurance must reflect organizational and technological context. A regulated financial institution with mature MRM and validation functions may require a different structure from an enterprise primarily consuming external AI services or an organization developing and hosting its own AI systems. Assurance requirements should therefore vary according to materiality, data sensitivity, potential harm, regulatory exposure, autonomy, AI architecture and topology, and third-party dependency. The objective is a coordinated assurance architecture that accommodates different organizational designs while preserving clear accountability, effective challenge, appropriate evidence, escalation, and independence [6], [12].

The article first examines the ongoing relevance and limitations of traditional MRM, as well as the expansion of the AI risk landscape. It then considers the emerging governance and regulatory environment and distinguishes the responsibilities of AI risk management, MRM, independent validation, AI auditing, and related organizational functions. Building on

these foundations, the article examines organizational design and assurance requirements for generative and agentic AI, then presents the Integrated AI Assurance Framework and alternative approaches for operationalizing it across organizational and deployment contexts. High-risk use cases illustrate its application, followed by implications for practice, limitations, and opportunities for future research.

The central proposition is that the transition to generative and agentic AI does not eliminate the need for MRM; it requires MRM to operate within a broader, deliberately coordinated AI assurance architecture. No single function can independently provide complete assurance over increasingly complex AI systems. Effective AI assurance requires clear risk ownership, coordinated specialized functions, effective challenge, appropriate evidence and escalation, and independent assurance that governance, risk management, validation, and controls operate as intended [10], [9].

APPROACH AND FRAMEWORK DEVELOPMENT

This article uses a conceptual and applied framework-development approach to examine how organizations can integrate established model risk management principles with emerging AI governance, risk management, independent validation, and internal audit practices. The analysis draws on four complementary sources of evidence: (1) governmental and regulatory guidance relevant to model and AI risk management; (2) international AI governance and risk-management standards; (3) professional guidance addressing internal audit, assurance, and organizational accountability; and (4) relevant scholarly literature concerning model risk, AI governance, generative AI, agentic AI, validation, and assurance [3], [6], [1], [11], [12], [13].

Sources were selected purposively rather than through a systematic-review protocol. Selection emphasized authoritative and directly applicable materials that addressed one or more of the article's two research questions and contributed to at least one core assurance function: governance and accountability; risk identification and management; independent validation and effective challenge; or independent audit and assurance. Priority was given to primary governmental and regulatory guidance, international standards, professional assurance guidance, and peer-reviewed or publicly available scholarly literature addressing model risk, AI governance, generative AI, agentic AI, validation, and assurance. Sources were comparatively reviewed for their stated scope, allocation of responsibility, independence expectations, evidence requirements, lifecycle coverage, and treatment of third-party or partially observable AI systems [3], [6], [1], [11], [12], [14], [13].

The framework-development process then used comparative synthesis. Recurring requirements and unresolved organizational gaps were mapped across the selected sources, with particular attention to ownership, effective challenge, independent assurance, evidence availability, architecture and hosting, third-party dependence, transparency, data access, autonomy, and organizational use. These cross-source patterns were used to derive the AI Assurance Boundary, the Assurance Sufficiency Principle, and the four-layer Integrated AI Assurance Framework. The resulting framework was assessed analytically through contrasting organizational contexts and illustrative high-risk use cases rather than through statistical hypothesis testing or empirical causal inference [1], [2], [15], [9].

This synthesis provides the basis for the Integrated AI Assurance Framework developed in this article. Framework construction proceeded by mapping recurring assurance requirements to four functions—AI governance; AI risk management and model risk management; independent AI/model validation; and AI auditing—and then testing whether those functions remain sufficient when the assured object is expanded from a model to a composite AI-enabled system. The framework was evaluated analytically against contrasting computing topologies and high-risk use cases, including internally controlled AI, externally provided foundation-model services, hybrid/RAG architectures, and agentic systems. The evaluation asks whether the framework can identify material technical dependencies, assign accountable ownership, specify evidence and validation needs, preserve effective challenge and audit independence, and support escalation when observability or evidence is insufficient. Adjacent functions—including ERM, IT risk, cybersecurity, third-party risk management, legal and compliance functions, and governance committees—are incorporated where their responsibilities intersect these layers [6], [7], [8], [10].

The article makes three related contributions. First, the AI Assurance Boundary defines the models, data, controls, technologies, external dependencies, human actors, and organizational processes whose behavior or evidence is material to justified reliance on an AI-enabled capability. Second, the Assurance Sufficiency Principle provides a risk-based decision rule for circumstances in which complete transparency, traceability, or explainability is unavailable. Third, the Integrated AI Assurance Framework (IAAF) defines the organizational architecture through which governance, AI risk management and MRM, independent validation, and internal audit govern, manage, challenge, and independently assure that boundary. Together, these contributions extend traditional MRM principles into a broader assurance architecture for generative and agentic AI systems.

FROM TRADITIONAL MODEL RISK TO AI RISK

Traditional Model Risk Management—and the Adequacy of the Legacy Model

The durability of MRM is most evident throughout its lifecycle. Inventory establishes the population of models or AI-enabled systems on which the organization relies. Materiality and risk classification determine the intensity of governance, validation, monitoring, and approval. Development and acquisition standards establish intended use, design claims, data requirements, limitations, and accountable ownership. Independent validation challenges those claims before reliance; ongoing monitoring tests whether performance and assumptions remain acceptable; change management identifies when modifications require reassessment; documentation preserves institutional knowledge and evidence; and retirement addresses residual dependencies, retained data, access, and replacement risk [3].

Contemporary AI does not invalidate these purposes. It changes the conditions under which they must be achieved. A foundation model may be externally developed, dynamically updated, and embedded within a composite service whose outputs depend on prompts, enterprise data, retrieval sources, policy controls, interfaces, and human decisions. An agent may additionally possess identities, permissions, tools, and authority to initiate

consequential actions. The object requiring assurance is therefore frequently larger than the model identified in a conventional inventory.

The legacy assumptions requiring adaptation are not formal requirements that every model be simple or perfectly transparent. They are practical expectations that the organization can identify a sufficiently stable object, obtain evidence about its construction and operation, reproduce material results, control changes, and isolate responsibility for deficiencies. Where those conditions no longer hold, MRM must coordinate with broader AI, technology, cybersecurity, privacy, third-party, legal, compliance, and operational assurance rather than treating unavailable evidence as proof of acceptability [6], [1], [5].

Traditional MRM provides a durable foundation for governing quantitative and algorithmic decision systems. Established practices emphasize model identification and inventory, documented development, materiality and risk classification, independent validation, ongoing monitoring, change management, limitations, governance, and retirement. U.S. banking guidance further emphasizes the need for effective challenge by objective experts and recognizes that model risk can arise from interactions and dependencies among models [3].

The question is not whether generative and agentic AI make MRM obsolete. It is whether implementation expectations developed around comparatively bounded and inspectable models remain achievable when the object being assured becomes a composite, externally dependent, dynamically changing AI-enabled system. We contend that the increasing opacity, externalization, scale, and architectural complexity of contemporary foundation-model systems may render some traditional expectations of model transparency impractical or unattainable, particularly where the organization consuming the model does not own or control its development.

This does not justify abandoning transparency, traceability, or explainability. Instead, it requires a risk-based determination of what level of evidence is sufficient for justified reliance. The objective should not be perfect knowledge of every computational operation, parameter, byte, or internal representation. It should be sufficient evidence to determine fitness for the intended use, identify material limitations and uncertainty, support effective challenge, and meet applicable legal and regulatory obligations.

The Expansion of the AI Risk Landscape

Generative and agentic AI broaden risk beyond conventional predictive performance to include hallucination and output reliability, bias and discrimination, privacy and confidentiality, cybersecurity and prompt injection, RAG-source integrity, third-party and concentration risk, autonomous actions, identity and permissions, human oversight, and insufficient traceability [6], [1]. Risk may reside in the interaction among components rather than in a defect isolated to the foundation model.

AI Architecture, Location, and Access as Risk Drivers

AI topology materially affects control, observability, evidence availability, and the feasibility of validation methods. Internally hosted systems may permit direct access to infrastructure, configuration, logs, and testing evidence; externally operated foundation-

model APIs or AI-enabled SaaS can place critical behavior outside the customer's direct control. Hybrid systems distribute responsibility further across internal data, external models, RAG, SaaS integrations, and agentic tools. Table 1 summarizes how these operating contexts affect typical control, observability, and the primary assurance implication.

Table 1: AI Operating Context, Control, and Observability

Operating Context	Typical Control	Observability	Primary Assurance Implication
Internally hosted/controlled AI	High relative control over infrastructure and configuration	Potentially high, subject to expertise and tooling	Direct testing can be emphasized; organization bears greater operational responsibility.
Externally provided foundation-model/API	Provider controls core model and many changes	Often limited to interfaces, documentation, contractual/provider evidence, and observed behavior	Evidence sufficiency and provider dependency become central; compensating controls may be required.
AI embedded in enterprise SaaS/product	Shared/opaque control; AI may be one component of a larger service	Variable and often indirect	TPRM, application assurance, privacy, security, MRM, and vendor evidence must be coordinated.
Hybrid/composite AI	Distributed across internal and external components	Mixed	Assurance Boundary must follow material dependencies and data/action paths rather than organizational ownership alone.

Generative and Agentic AI as an Assurance Challenge

Traditional Models → Machine Learning → Deep Learning → Generative AI → Foundation Models/RAG → Agentic AI

A useful progression is increasing opacity -> external dependency -> variability/adaptability -> autonomy -> consequential action. Traditional models principally produce outputs; agentic systems can increasingly produce consequences. As autonomy increases, validation must address not only what a model predicts or generates but what identities, permissions, tools, external systems, transaction limits, and intervention mechanisms can convert an output into an action [1], [15].

The AI Assurance Boundary

The boundary should be established for an intended use rather than for a model name alone. A practical boundary record should identify the business purpose; accountable owner; model and provider; hosting and access topology; enterprise and external data; retrieval sources; prompts and configurations; policy or safety controls; interfaces; agent identities, permissions, and tools; human decision points; downstream systems; monitoring; and evidence repositories. A component is material when its failure, change, or unavailability could alter the reliability, legality, security, or consequences of the approved use.

Boundary governance is also dynamic. Provider releases, new data sources, changed prompts, expanded permissions, additional tools, altered human-review practices, and downstream integrations can move the effective boundary even when the named model remains unchanged. Material boundary changes should therefore trigger a reassessment of classification, controls, validation scope, approval, monitoring, and residual risk acceptance.

This article proposes the AI Assurance Boundary as a practical construct for defining the scope of models, data, controls, technologies, external dependencies, human actors, and organizational processes whose behavior or evidence is materially necessary to establish justified reliance on an AI-enabled capability for its intended use. The Assurance Boundary may therefore be materially broader than the conventional model boundary.

Components remain within the Assurance Boundary even when they are externally controlled or unavailable for direct inspection. Lack of observability is an assurance limitation, not a reason to exclude a dependency from scope. An opaque dependency remains within the boundary precisely because the organization relies upon something it cannot fully assure.

This leads to a core proposition: opacity does not eliminate accountability; it increases uncertainty. The uncertainty created by unavailable evidence should be identified, evaluated, mitigated where feasible, and explicitly accepted or rejected according to the consequentiality of the intended use.

Assurance Under Incomplete Observability

Assurance sufficiency should be evaluated claim by claim. The organization should identify the claim necessary for reliance—for example, that a system is sufficiently accurate for a defined task, does not expose protected data, produces legally adequate reasons, or cannot execute transactions beyond approved limits—and map that claim to available evidence. Direct testing and reproducible primary evidence ordinarily provide the strongest support; scoped independent assurance, provider technical evidence, contractual commitments, and self-attestation provide progressively more indirect support and should be weighed according to relevance, independence, recency, and scope.

Unavailable evidence creates a documented uncertainty, not neutral conditions. The assessment should state what cannot be known, why it matters, which compensating controls reduce the uncertainty, what monitoring could reveal a failure, and who has authority to accept the residual exposure. Compensating controls may include narrower intended use, reduced autonomy, isolated data, transaction limits, human approval, behavioral testing, redundant verification, enhanced monitoring, contractual remedies, or an exit plan.

Sufficiency is therefore a decision threshold rather than a quantity of documentation. Evidence may be sufficient for a reversible, low-consequence use while remaining inadequate for credit denial, employment exclusion, clinical decisions, or autonomous financial actions. When material claims cannot be supported and compensating controls cannot reduce uncertainty to an accepted level, the defensible assurance decision is to restrict, defer, or reject the use.

Contemporary AI assurance cannot assume that complete transparency is attainable. Instead, the organization must determine whether the available evidence is sufficient to justify reliance on the AI-enabled system for a defined use. Missing evidence should be treated as uncertainty rather than silently converted into evidence of acceptability.

The required assurance threshold should increase with materiality, consequentiality, autonomy, data sensitivity, regulatory exposure, opacity, third-party dependence, scale, and reversibility of harm. A low-consequence internal summarization tool may be acceptable with behavioral testing, data controls, monitoring, and human review even when the foundation model cannot be inspected. The same evidence may be inadequate for autonomous credit denial, employment screening, or another consequential regulated decision.

This article refers to this as an Assurance Sufficiency Principle: where perfect transparency, traceability, or explainability is technically, economically, or contractually unattainable, organizations should establish the degree of evidence necessary for defensible reliance on the intended use; characterize material deficiencies as uncertainty; apply compensating controls where feasible; and subject residual uncertainty to explicit risk acceptance. Table 2 illustrates this evaluation across several assurance dimensions, showing the governing question, possible evidence, and residual-uncertainty signal for each.

Table 2: Assurance Dimensions Under Incomplete Observability

Assurance Dimension	Question	Possible Evidence	Residual-Uncertainty Signal
Model transparency	What is known about model design, intended use, limitations, and changes?	Provider documentation, model cards, testing, contractual notices	Core design/changes unavailable or unverifiable
Data transparency	What is known about training, enterprise, RAG, and derived data?	Datasheets, lineage, data-flow mapping, RAG inventory	Unknown provenance or unobservable transformations
Decision traceability	Can a material outcome be reconstructed?	Versions, inputs/outputs, prompts/configuration, control logs, overrides	Outcome cannot be reproduced or causal pathway cannot be reconstructed
Outcome explainability	Can the organization give a meaningful reason for a consequential result?	Decision factors, validated explanations, adverse-action rationale, human review	Explanation is generic, post hoc, or disconnected from actual factors
Control observability	Can guardrails, policy engines, and other controls be tested?	Control testing, false-positive/negative rates, exceptions, monitoring	Provider control exists but behavior/thresholds cannot be evaluated
Change visibility	Can material provider/system changes be detected?	Versioning, notices, regression testing, monitoring	Behavior changes without identifiable release/change event

THE EMERGING AI GOVERNANCE AND REGULATORY LANDSCAPE

The governance of artificial intelligence is developing through a combination of regulatory guidance, voluntary risk-management frameworks, international standards, professional assurance practices, and sector-specific requirements. Collectively, these sources establish increasingly clear expectations for responsible AI governance, risk management,

documentation, testing, oversight, and assurance. The practical challenge for organizations is that these expectations do not constitute a single integrated governance regime. Organizations must determine how requirements from different sources apply to their AI systems and translate them into operational responsibilities, controls, validation activities, evidence, and independent assurance [3], [6], [11], [12], [13].

Federal Reserve/OCC/FDIC Model Risk Management Guidance

The April 2026 interagency *Revised Guidance on Model Risk Management* issued by the Federal Reserve, Office of the Comptroller of the Currency (OCC), and Federal Deposit Insurance Corporation (FDIC) represents an important development in the relationship between traditional model risk management (MRM) and emerging AI governance. The guidance replaced the longstanding SR 11-7 framework and emphasizes a risk-based approach tailored to an institution's model risk profile, size, complexity, model purpose, exposure, and materiality. It retains fundamental MRM concepts such as sound model development, testing, validation, monitoring, governance, documentation, and effective challenge [3].

Significantly, however, the revised guidance states that generative and agentic AI models are novel and rapidly evolving and therefore are not within its scope. At the same time, the agencies state that banking organizations' broader risk-management and governance practices should determine the appropriate governance and controls for tools and systems that fall outside the guidance. The guidance continues to apply to traditional statistical and quantitative models and non-generative, non-agentic AI models [3].

This distinction illustrates the central problem this article addresses. Excluding generative and agentic AI from a particular MRM framework does not eliminate the risks associated with these technologies or the organization's responsibility to manage them. Instead, it shifts attention toward the broader governance architecture through which AI risk, validation, control, and assurance must be coordinated.

NIST AI Risk Management Framework

The NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0) provides a broader, technology-oriented approach. Designed for voluntary and cross-sector use, the AI RMF organizes AI risk management around four functions: **Govern, Map, Measure, and Manage**. *Govern* is cross-cutting, while Map, Measure, and Manage support contextual risk identification, assessment, prioritization, and treatment throughout the AI lifecycle. NIST emphasizes that AI risk management should be continuous rather than treated as a one-time compliance activity [6].

This lifecycle orientation complements traditional MRM while extending attention beyond quantitative model performance. It incorporates characteristics of trustworthy AI including validity and reliability, safety, security and resilience, accountability and transparency, explainability and interpretability, privacy, and management of harmful bias [6].

NIST Generative AI Profile

NIST further extended the AI RMF through the *Generative Artificial Intelligence Profile* (NIST AI 600-1). The profile applies specifically to the Govern, Map, Measure, and Manage functions for risks unique to or intensified by generative AI. It is intended to help organizations incorporate generative AI risk considerations across design, development, deployment, use, and evaluation [1].

The profile is especially relevant to assurance because generative AI may depend on foundation models, external providers, cloud services, organizational data, and downstream applications that an organization does not fully control. Consequently, governance increasingly must address not simply the model itself but the larger system and operating context in which the model is used [1], [2].

ISO/IEC AI Management and Risk Standards

International standards provide another complementary layer. ISO/IEC 42001:2023 establishes an AI management system standard based on a Plan-Do-Check-Act approach. Its organizational orientation emphasizes policies, procedures, governance, responsibilities, and continuous improvement for managing AI-related risks and opportunities [11].

ISO/IEC 23894:2023 guides organizations that develop, produce, deploy, or use AI-enabled products, systems, and services. Importantly, the standard is intended to integrate AI risk management into broader organizational activities and can be adapted to organizational context. Together, these standards reinforce the principle that AI governance is an enterprise management responsibility rather than exclusively a technical or model-management function [12].

Internal Audit and the IIA AI Auditing Framework

Risk management alone does not provide independent assurance that governance and controls operate effectively. The Institute of Internal Auditors' Artificial Intelligence Auditing Framework addresses this distinction by providing principles-based guidance for internal auditors assessing AI governance, management, risk, and controls. The framework draws upon the IIA's Three Lines Model and broader professional internal-audit standards [13].

Its inclusion in the AI governance landscape is important because internal audit occupies a different organizational role from AI risk management or model validation. Management owns and manages risk; specialized risk and validation functions provide oversight and effective challenge; internal audit independently evaluates whether governance, risk management, and controls are appropriately designed and operating effectively [14], [8], [13].

Sector-Specific and Emerging Regulatory Requirements

Employment uses illustrate the interaction between AI governance and existing law. Title VII and the Uniform Guidelines on Employee Selection Procedures apply to selection procedures regardless of whether a human, conventional software, or AI implements the

procedure. The Uniform Guidelines address adverse impact, job-relatedness, recordkeeping, and technical standards for validity studies. EEOC materials further emphasize that employers remain responsible when automated systems make or inform selection decisions and that AI tools can also create disability-discrimination risk under the Americans with Disabilities Act. The use of a vendor does not transfer the employer's accountability for the resulting employment practice [16].

Credit decisions create a parallel requirement. The Equal Credit Opportunity Act and Regulation B apply to AI-assisted credit decisions, including requirements for specific and accurate reasons when adverse action is taken. A creditor cannot rely on model complexity or opacity as a justification for an explanation that is generic, inaccurate, or disconnected from the factors that actually drove the decision. This makes reconstruct ability and outcome-level explanation legal and assurance requirements rather than optional transparency features [17].

Organizations must also address requirements arising from specific industries, jurisdictions, and AI use cases. The European Union AI Act establishes risk-based obligations and identifies uses including certain employment and recruitment systems and systems evaluating individual creditworthiness as high-risk applications. In the United States, requirements are developing through federal, state, and local mechanisms. For example, New York City's automated employment decision tool requirements mandate bias audits and specified disclosures for covered employment technologies [18], [19].

These developments illustrate why AI governance cannot be reduced to compliance with a single AI framework. The applicable requirements depend on jurisdiction, industry, intended use, affected individuals, data, materiality, and potential consequences.

Translating External Requirements into Organizational Practice

The proliferation of frameworks creates an implementation problem. NIST may describe risk-management outcomes; ISO standards may establish management-system expectations; regulators may impose sector-specific requirements; MRM may require validation and effective challenge; and internal-audit standards may require independent assurance. Yet none automatically determines how an organization should distribute these responsibilities among AI governance, MRM, ERM, IT risk, cybersecurity, privacy, compliance, third-party risk management, validation, and internal audit [3], [6], [11], [12], [14].

Organizations therefore require a translation mechanism that maps external expectations to internal ownership, controls, evidence, validation requirements, monitoring, escalation, and assurance. Simply demonstrating that an organization has adopted a framework or policy is insufficient if responsibilities remain fragmented or controls cannot be evidenced.

The Emerging AI Assurance Gap

The resulting challenge can be characterized as an **AI assurance gap**. Standards, regulatory guidance, and professional frameworks increasingly explain *what* organizations should govern or manage, but organizations must still determine *who* owns the risks, *who* evaluates the AI system, *who* provides effective challenge, *what evidence* demonstrates appropriate

control, and who independently determines whether the overall governance and risk-management system is working as intended [7], [10], [9].

This gap becomes particularly important for generative and agentic AI, where responsibility may remain with the organization even when the underlying foundation model is externally developed or only partially transparent. Closing the gap therefore requires more than additional requirements. It requires deliberate alignment among the organizational functions responsible for AI governance, risk management, validation, cybersecurity, compliance, and independent assurance [1], [5], [2].

DISTINGUISHING AND ALIGNING THE FUNCTIONS OF AI ASSURANCE

Effective AI assurance depends on multiple organizational functions performing different but complementary roles. AI governance, AI risk management, model risk management (MRM), independent validation, and internal audit address different questions and degrees of independence. ERM, IT risk, cybersecurity, third-party risk management, legal and compliance functions, and governance committees provide additional expertise and oversight. The objective is therefore not to consolidate AI assurance within a single function, but to establish clear boundaries while enabling coordinated risk management, effective challenge, and independent assurance [6], [7], [8].

AI Risk Management

AI risk management addresses enterprise- and system-level risks arising from the development, acquisition, deployment, and use of AI. It extends beyond model performance to consider operational, cybersecurity, privacy, legal, compliance, fairness, reputational, and third-party risks. Consistent with the NIST AI RMF, risk management should be continuous and contextual, with governance, risk identification, measurement, treatment, and monitoring occurring throughout the AI lifecycle [6].

AI risk management does not eliminate responsibility at the business or system-owner level. Rather, it outlines the processes for identifying, classifying, evaluating, mitigating, monitoring, and escalating risks.

AI Model Risk Management

MRM provides specialized lifecycle management for model-related risks. Established MRM practices include model inventory, risk classification, documentation, development and testing standards, validation, monitoring, change management, and retirement [3].

These principles remain relevant to AI, but their applicability depends on the technology and operating context. Generative and agentic AI may introduce risks extending beyond conventional model boundaries, including externally controlled foundation models, dynamic retrieval sources, cybersecurity vulnerabilities, and autonomous actions. AI MRM should therefore operate as a specialized component of broader AI risk management rather than being assumed to encompass every AI-related risk [3], [1].

Independent Model/AI Validation

Independent validation provides effective challenge of claims made about an AI system. Depending on materiality and intended use, validation may evaluate design, data, assumptions, performance, limitations, robustness, fairness, explainability, and ongoing reliability. It should determine whether evidence supports the system's approved use rather than simply confirm that required documentation exists.

The validation approach may change when organizations use externally developed foundation models. Limited access to training data, model weights, or architecture may require greater reliance on use-case testing, output evaluation, vendor evidence, system configuration, monitoring, and documented limitations. Independence remains important because those responsible for developing or operating a system should not be the sole judges of its acceptability [3], [1].

AI Auditing

AI auditing provides independent assurance regarding the effectiveness of the broader governance and control environment. Internal audit may assess whether AI governance, risk management, validation, monitoring, third-party oversight, cybersecurity controls, documentation, and escalation processes are appropriately designed and operating effectively [14], [13], [10].

This role differs fundamentally from risk management and validation. Internal audit should not own AI risks, establish risk acceptance on management's behalf, or become responsible for operating controls it may later audit. Its value derives from organizational independence and its ability to evaluate whether the overall assurance structure functions as intended [14], [8].

ERM, IT Risk, Cybersecurity, Third-Party Risk, Legal/Compliance, and Steering Committees

The contribution of these functions should be defined through evidence and decision rights. ERM determines how AI exposures are incorporated into enterprise risk appetite, aggregation, reporting, and acceptance. IT risk and architecture evaluate resilience, integration, change, and dependency risks. Cybersecurity assesses adversarial manipulation, identity and access management, data protection, and containment. Third-party risk management evaluates provider governance, security, performance evidence, contractual protection, concentration, change notification, and exit feasibility. Privacy and legal/compliance functions determine applicable obligations and the acceptability of data use, outcomes, explanations, notices, and retained evidence. Business and system owners remain accountable for the use and its controls.

These functions should produce evidence that can be reused without treating consultation as transferred ownership. For example, a cybersecurity assessment may support validation of an agent's security-relevant behavior, but it does not establish overall fitness for use. A vendor assurance report may support third-party due diligence, but it does not demonstrate that the organization's configuration, data, workflow, and intended use

are acceptable. A steering committee may resolve cross-functional disagreement or approve residual risk, but it should not obscure the named individual accountable for the decision.

AI rarely fits neatly into a single organizational department because its risks cross established functional boundaries. ERM connects AI exposure to enterprise risk appetite and aggregation. IT risk addresses technology dependencies and resilience; cybersecurity addresses adversarial threats, identity, access, and data protection; third-party risk evaluates providers and external dependencies; and legal, privacy, and compliance functions interpret applicable obligations. Governance or steering committees can coordinate policies, priorities, risk acceptance, and escalation across these functions [7], [20], [5].

Coordination does not require merging these functions. Each contributes specialized knowledge while responsibility for particular risks and decisions should remain identifiable.

Where the Functions Overlap—and Where They Should Not

Appropriate overlap occurs through controlled handoffs. Cybersecurity may conduct prompt injection and permissions testing and provide the results for independent validation. Validation may determine that unresolved findings exceed approved tolerances and require management risk treatment or formal risk acceptance. Legal or compliance specialists may determine the meaning of an obligation, while validation tests whether the implemented system produces evidence consistent with that meaning. Internal audit may later evaluate whether these handoffs occurred, whether authorized parties resolved objections, and whether the validation program itself operates effectively.

Inappropriate substitution occurs when a contributing function is treated as completing another function's responsibility. A completed vendor review is not model or system validation; validation is not management's acceptance of risk; a committee vote is not evidence that controls work; and internal audit should not design, own, or operate the validation and control processes it will later assess. The framework therefore requires coordinated evidence while retaining identifiable ownership, challenge, acceptance, and assurance.

Core distinction to preserve: AI/MRM manages risk; independent validation provides effective challenge of model/system claims; internal audit independently assesses whether governance, risk management, validation, and controls are appropriately designed and operating. Adjacent specialist functions contribute evidence and domain-specific challenge without erasing accountable ownership. The functions necessarily share information, evidence, and expertise, but overlap should not obscure accountability or independence. Table 3 compares AI risk management/MRM, independent validation, and AI audit across purpose, timing, ownership, independence, required expertise, evidence, and escalation.

The principal boundary is therefore straightforward: AI risk management and MRM manage AI and model risk; independent validation challenges whether evidence supports acceptable use; and AI audit independently assesses whether governance, risk management, validation, and associated controls are working effectively [14], [8], [10]. Effective AI assurance requires all three perspectives without allowing coordination to erase their distinct accountabilities.

Table 3: AI Risk Management/MRM, Independent Validation, and AI Audit Compared

Dimension	AI Risk Management / MRM	Independent AI/Model Validation	AI Audit
Purpose	Identify, assess, mitigate, monitor, and document risk	Provide effective challenge of the AI/model and intended use	Provide independent assurance over governance, risk management, validation, and controls
Primary question	What can go wrong and how should the risk be managed?	Does evidence support acceptable use and performance?	Are governance and controls appropriately designed and operating effectively?
Timing	Continuous throughout lifecycle	Before approval and periodically/event-driven thereafter	Periodic and risk-based
Ownership	Management/risk functions	Organizationally independent from development/use	Internal audit
Independence	Risk management oversight; not third-line assurance	Independent from model/system development and ownership	Independent third-line assurance
Core KSAs	AI/model, risk, business and regulatory knowledge	AI/model, data, testing, domain and validation expertise	Audit, governance, risk, controls and sufficient AI expertise
Evidence	Inventories, assessments, controls, monitoring, exceptions	Testing, data, performance, limitations and validation findings	Policies, risk records, validation evidence, control testing and audit evidence
Escalation	Material risk and control exceptions	Validation findings and unresolved limitations	Governance/control deficiencies escalated to management/board

THE ORGANIZATIONAL DESIGN PROBLEM: WHERE AI GOVERNANCE AND ASSURANCE LIVE

AI governance creates an organizational design problem because its risks rarely align with the boundaries of a single existing function. An AI system may simultaneously create model, technology, cybersecurity, privacy, legal, compliance, operational, and third-party risks. Generative and agentic AI further complicate accountability because organizations may consume externally developed models, embed AI within existing products, or authorize AI agents to access data and initiate actions. Effective assurance therefore depends not on establishing one universal organizational home for AI, but on deliberately allocating ownership, oversight, effective challenge, and independent assurance [6], [7], [9].

The Three Lines Model and AI

The Three Lines Model provides a useful foundation for separating responsibilities. The first line includes business owners, developers, product owners, and AI system owners responsible for deploying and operating AI systems in accordance with approved requirements and controls. These parties remain accountable for understanding intended use, maintaining appropriate controls, monitoring performance, and managing risks associated with their activities [8].

The second line may include AI governance, enterprise risk management (ERM), IT risk, cybersecurity risk, compliance, model risk management (MRM), third-party risk

management, and independent validation, depending on organizational structure and the nature of the AI system. These functions establish or oversee risk-management requirements and provide challenge. The third line, internal audit, provides independent assurance regarding whether governance, risk management, validation, and controls are appropriately designed and operating effectively. Maintaining these distinctions is particularly important when organizations consolidate AI responsibilities for efficiency [14], [8].

Organizational Placement Options for AI Governance

No single organizational placement is appropriate for every organization. A financial institution with an established MRM function may place substantial AI oversight within model risk, particularly for AI supporting material decisions. An organization heavily dependent on cloud platforms and externally provided AI may place greater responsibility within technology or IT risk. Organizations facing significant legal or regulatory exposure may emphasize compliance, while others may coordinate AI governance through ERM [3], [6], [5]. A cross-functional structure offers another option by bringing these perspectives together without transferring all AI risk to a new centralized department. Organizational placement should reflect industry, regulatory exposure, existing risk-management maturity, AI architecture and topology, data sensitivity, materiality, autonomy, and third-party dependency [6], [12], [7].

Organizational placement may vary across ERM, MRM, technology/IT risk, compliance, dedicated AI governance, federated structures, or hybrids. The organizational home may matter less than whether authority, accountability, resources, evidence flow, escalation, decision rights, and independence are explicit and effective.

AI Governance Committees, Steering Committees, and Decision Rights

Cross-functional AI governance or steering committees can coordinate policies, risk classification, deployment decisions, exceptions, and escalation. Membership may include representatives from technology, business, risk, cybersecurity, privacy, legal, compliance, MRM, and internal audit, as appropriate [6], [7].

Committees, however, should not become substitutes for individual accountability. Their value lies in coordination and decision-making across organizational boundaries. Policies should therefore identify who owns an AI system, who accepts residual risk, who can approve deployment, who can require remediation or suspension, and where unresolved disagreements are escalated. Internal audit may advise on governance considerations while preserving the independence required for subsequent assurance [14], [8].

Independence and Effective Challenge

AI assurance requires both specialized expertise and sufficient independence. Developers or business owners cannot provide independent validation of the systems they are responsible for. Similarly, internal audit cannot assume management responsibility for designing or operating the controls it later audits [14], [8].

Independent validation provides effective challenge by questioning intended use, assumptions, data, performance, limitations, robustness, fairness, and other relevant characteristics. Internal audit subsequently assesses whether governance, risk management, validation, and control processes function effectively. Organizational structures may vary, but these boundaries should remain identifiable [3], [14], [10].

Third-Party Foundation Models and the Validation Problem

Third-party foundation models create a particularly difficult organizational problem. An enterprise may remain accountable for AI-enabled outcomes even if it lacks access to training data, model weights, development processes, architecture, or complete lineage. Traditional validation approaches based on detailed access to model construction may therefore be impractical [1], [5], [2].

Validation must consequently become more evidence- and use-case-oriented. Organizations can evaluate intended use, outputs, testing results, data flows, configurations, access controls, contractual assurances, vendor documentation, monitoring, and known limitations even when the underlying model cannot be fully inspected. The inability to obtain traditional validation evidence should itself be considered a risk rather than an assumption that validation is unnecessary [1], [5], [21].

Control Gaps, Duplication, and Boundary Failures

Fragmented responsibility creates two opposing risks: control gaps and unnecessary duplication. Cybersecurity may assume that MRM addresses AI risk, while MRM may view prompt injection or data leakage as cybersecurity responsibilities. Compliance may focus on regulatory requirements, while third-party risk evaluates the provider and AI governance evaluates acceptable use. Each activity may be reasonable in isolation while important risks remain collectively unowned [7], [9].

An enterprise AI inventory linked to clearly identified owners, risk classifications, assurance responsibilities, and escalation paths can help reveal these boundary failures. Coordination should eliminate neither specialization nor independence; rather, it should establish where responsibilities intersect and who remains accountable when they do [6], [7].

Contextual Organizational Design Patterns

Three broad patterns illustrate how organizational design may differ. A regulated financial institution may build on existing MRM and independent validation capabilities while integrating AI governance, cybersecurity, compliance, and third-party risk management. A large enterprise that consumes external AI services may rely more heavily on IT risk, cybersecurity, privacy, procurement, and third-party risk when establishing centralized AI governance requirements. An organization developing or internally hosting AI may place greater responsibility on technology and AI development functions while requiring independent risk oversight and validation proportionate to materiality [3], [6], [5].

These patterns are illustrative rather than prescriptive. The essential requirement is not where AI governance appears on an organizational chart, but whether the resulting structure establishes clear ownership, appropriate expertise, effective challenge, evidence, escalation, and independent assurance. Organizational design should therefore follow the risks and architecture of AI rather than force emerging AI systems into organizational boundaries created for earlier technologies [6], [12], [7]. Table 4 outlines the likely organizational emphasis and primary assurance challenge for four illustrative contexts.

Table 4: Contextual Organizational Design Patterns

Context	Likely Organizational Emphasis	Primary Assurance Challenge
Regulated financial institution	Existing MRM/validation + compliance + fair-lending + AI governance + TPRM	Integrating mature model governance with broader AI/system assurance and regulatory evidence
Large enterprise consuming external AI services	TPRM + security/privacy + architecture + business ownership + centralized AI governance	Limited model observability and provider leverage; consistent use-case classification and compensating controls
Internally hosted/private AI	Engineering + security + data governance + MRM/validation + business owner	Greater technical control but greater responsibility for operations, model lifecycle, data, and monitoring
Smaller organization without dedicated MRM	Business owner + IT/security + legal/compliance + external expertise	Resource/skills constraints; defining proportionate assurance without creating unmanageable bureaucracy

RISK, VALIDATION, AND ASSURANCE REQUIREMENTS FOR GENERATIVE AND AGENTIC AI

The dimensions below are evaluated through a common assurance structure. For each material risk, the organization should identify the failure mode and consequence; the accountable owner; preventive, detective, and corrective controls; the independent validation method and acceptance criteria; the evidence required to support reliance; continuous monitoring and key risk indicators; escalation and suspension triggers; and the aspect of governance or control operation subject to independent audit. Applying this structure prevents a catalog of AI risks from being mistaken for an assurance program.

Generative and agentic AI require assurance that extends beyond conventional assessments of model accuracy and performance. Assurance must consider the complete AI system: data, models, retrieval sources, interfaces, security controls, human oversight, third parties, and, increasingly, the ability of AI agents to access resources and initiate actions. The intensity of validation and assurance should be risk-based, reflecting materiality, intended use, data sensitivity, potential harm, autonomy, regulatory exposure, architecture, and third-party dependency [6], [1].

Data Quality, Provenance, and Lineage

Organizations should understand what data an AI system uses, its origin, how it is transformed, and where it flows. Validation should assess whether the data are appropriate for the intended use and whether their provenance and lineage are sufficiently documented. Evidence may include data inventories, lineage records, quality testing, access controls, and

documentation of external sources. Data owners and system owners manage these risks, while validation and audit independently challenge the adequacy of the associated processes and controls [1], [22].

Transparency and Explainability

Transparency requirements should correspond to the consequences of AI-supported decisions. Organizations should document system purpose, capabilities, limitations, dependencies, and appropriate uses. Where AI influences consequential decisions, reviewers should determine whether outputs can be sufficiently explained to users, affected individuals, management, regulators, and other stakeholders. Limited transparency in third-party models should be explicitly recognized as a risk rather than obscured by unsupported claims of explainability [6], [1], [21].

For high-impact uses, general documentation about how a model works is not equivalent to explaining why a specific outcome occurred. Assurance should seek sufficient evidence to reconstruct relevant inputs, provenance, transformations, model/control versions, contextual information, material intermediate decisions, human interventions, and the factors that produced the final outcome [6], [21], [22].

The standard should be proportional rather than absolute: traceability sufficient to support the consequence and obligation of the use case, not an impossible demand to account for every internal bit-level operation.

Bias, Fairness, and Discrimination

AI systems used in lending, employment, healthcare, or other consequential applications require evaluation for potentially discriminatory outcomes. Testing should consider relevant populations, data representation, performance differences, intended use, and applicable legal requirements. Evidence may include fairness testing, impact assessments, validation results, complaints, and monitoring data. Legal and compliance expertise may be necessary alongside technical validation because statistical measures alone cannot determine legal or organizational acceptability [6], [1], [23].

This article uses the proposed term control-induced model risk (CIMR) to describe the risk that mechanisms introduced to constrain, secure, align, govern, or otherwise reduce AI risk may themselves materially degrade accuracy, consistency, fairness, explainability, availability, or intended utility. Accordingly, safety and governance controls should be validated for both their risk-reducing effect and their potential unintended impact rather than presumed effective by design.

Hallucination and Output Reliability

Reliability should be defined at the end-to-end system and use-case level. A foundation model's benchmark performance does not establish that a configured application, retrieval process, policy layer, user interface, or human workflow will produce acceptable outcomes. System owners should define material error categories, acceptable rates, severity thresholds, authoritative sources, prohibited uses, and required human verification.

Validation should use representative and adversarial cases, stratified by consequence and user population. It should examine citation or grounding fidelity, consistency, refusal behavior, uncertainty communication, and the effectiveness of controls intended to prevent unsupported outputs [1], [24].

Evidence should include test design, expected answers or evaluation criteria, model and configuration versions, results, error analysis, control failures, exceptions, and approvals. Monitoring should capture material complaints, corrections, unsupported citations, override patterns, and changes following provider, prompt, RAG, or control updates. Threshold breaches should trigger investigation, restricted use, revalidation, or suspension rather than being absorbed as an undefined characteristic of generative AI.

Model and Performance Drift

Drift in contemporary AI may arise from changing input populations, concepts, prompts, configurations, retrieval corpora, policy controls, provider models, interfaces, tools, or human use. Some changes are observable releases; others appear only through altered behavior. Monitoring should therefore combine statistical performance measures with behavioral regression tests, source and configuration inventories, provider-change notices, incident trends, user feedback, and control-effectiveness measures.

Owners should define material-change criteria and event-driven reassessment triggers. Changes affecting consequential outcomes, data access, autonomy, legal explanations, safety controls, or known limitations should prompt targeted or full revalidation. Where the organization cannot identify when an external dependency changes, the resulting change uncertainty should influence assurance intensity, monitoring frequency, and whether the use remains acceptable [3], [6], [1].

Cybersecurity, Prompt Injection, and Adversarial Manipulation

Assurance should connect threat scenarios to the AI Assurance Boundary. Testing should examine direct and indirect prompt injections, poisoned retrieval content, unauthorized instructions, data exfiltration, tool misuse, privilege escalation, insecure output handling, model or API abuse, and manipulation of human operators. For agentic systems, the critical question is not merely whether an attack changes text but whether it can cross an identity, permission, data, transaction, or system boundary and produce a consequential action [1], [25], [15]. Cybersecurity owns threat assessment and security controls; system owners own secure operation; and independent validation challenges whether the combined controls support the approved use. Evidence should include threat models, architecture and data flows, access and permission tests, adversarial test cases, containment results, incident records, residual-risk decisions, and remediation. Material control failures, unexpected tool use, boundary crossing, or inability to reconstruct an attack should trigger escalation and reassessment.

RAG and Knowledge-Source Integrity

RAG assurance should address source governance and retrieval behavior separately. Source governance includes authority, provenance, currency, ownership, access, segregation,

change control, and protection against poisoning or unauthorized modification. Retrieval validation should examine whether relevant sources are selected, whether permissions are enforced at query time, whether conflicting or stale content is handled appropriately, and whether generated answers remain faithful to retrieved evidence [1], [26].

Evidence should link corpus and index versions, retrieval configuration, permissions, retrieved passages, generated outputs, and citations for material transactions. Monitoring should identify source changes, retrieval failures, ungrounded answers, permission anomalies, and concentration on unreliable sources. Data governance and security may manage source and access controls, while independent validation determines whether the complete RAG implementation supports reliable use.

Privacy, Confidentiality, and Data Leakage

The relevant data pathways include prompts, uploaded files, embeddings, retrieval stores, fine-tuning or training data, system and provider logs, model outputs, caches, monitoring repositories, and information accessed or generated by agents. Assurance should determine purpose, authority, classification, minimization, retention, location, access, onward use, and deletion for each pathway. Provider terms and technical configurations should be examined to determine whether enterprise data can be retained, used for training, disclosed to subprocessors, or exposed through other users or system behavior.

Controls may include data loss prevention, prompt and output filtering, isolation, encryption, access restrictions, redaction, retention limits, contractual prohibitions, and human review. Validation should test representative sensitive-data scenarios and control bypasses rather than relying solely on policy statements. Evidence of unauthorized disclosure, memorization, access, retention, or unexplained provider-side use should trigger privacy and security escalation, incident response, and reassessment of the use [6], [1], [5].

Human Oversight and Human-in-the-Loop Controls

Human involvement is substantive only when the reviewer has sufficient competence, authority, time, information, and practical ability to question or change the result. A nominal approval click after an opaque recommendation is not an effective control. The workflow should identify which decisions require review, what information the reviewer receives, when independent verification is required, who may override the system, and when the AI output must be disregarded or escalated. Validation should test the human-AI system rather than the technical component alone. Evidence may include reviewer qualifications, decision time, agreement and override rates, reasons for overrides, errors accepted or corrected, escalation, and monitoring for automation bias or rubber-stamping. Persistent nonuse of override authority or unexplained convergence with automated recommendations may indicate that the human control is ceremonial rather than effective [27], [28], [29].

Agent Identity, Permissions, and Autonomous Actions

An agent should be treated as a delegated actor with a defined identity, owner, purpose, authority, and termination condition. Permissions should be scoped by least privilege across

data, tools, systems, transactions, and environments. High-consequence actions may require step-up authorization, segregation of duties, dual control, transaction limits, allow-listed tools or destinations, sandboxing, or human approval. Credentials should be attributable, rotatable, revocable, and unavailable to the model except through controlled interfaces.

Validation should examine normal, erroneous, adversarial, and ambiguous scenarios, including attempts to exceed authority, chain tools, persist access, conceal activity, or continue after a stop condition. Evidence should permit reconstruction of planning, tool selection, authorization, action, response, and intervention. Emergency suspension and credential revocation should be tested, not merely documented. Assurance intensity should increase sharply when an agent can produce irreversible or externally consequential actions [1], [25], [15].

Third-Party, Vendor, and Concentration Risk

Third-party assurance should trace dependencies beyond the contracting provider to foundation-model developers, cloud infrastructure, data or retrieval services, embedded tools, and material subcontractors. Due diligence should evaluate the scope and recency of independent reports, provider testing, data practices, security, resilience, model changes, incident notification, audit rights, service termination, portability, and the organization's ability to preserve evidence and continuity when a service changes or fails [5].

Concentration exists when multiple critical uses depend on the same model, provider, cloud, identity service, retrieval source, or control component. The resulting common-mode failure may be invisible when systems are assessed individually. Weak vendor evidence should not automatically be converted into a higher risk rating followed by approval. For high-consequence uses, unavailable evidence, insufficient contractual rights, unobservable changes, or impractical exits may make reliance unacceptable, regardless of compensating documentation.

Traceability, Logging, and Evidence Retention

Evidence requirements should be derived from the decisions and consequences that may need to be reconstructed. Material records may include system, model, prompt, policy, and retrieval versions; relevant inputs and outputs; source references; automated and human decisions; approvals; permissions; tool calls; exceptions; overrides; monitoring; changes; and incidents. Logs should be time-correlated, access-controlled, integrity-protected, and linked to accountable identities.

Retention must balance assurance, legal, privacy, security, and operational requirements. The organization should specify which events require full content versus metadata, how long evidence is retained, who may access it, and whether provider-side evidence can be obtained when needed. If a consequential outcome cannot be reconstructed sufficiently to validate, investigate, explain, or challenge it, the evidence gap should be treated as a limitation on the use—not merely as an inconvenience for auditors [21], [22].

Required Knowledge, Skills, and Abilities

Effective AI assurance requires multidisciplinary knowledge. MRM and validation personnel may require expertise in model performance, statistics, data, machine learning, generative AI, and validation techniques. Cybersecurity specialists contribute knowledge of adversarial threats, identity, access, and system security. Privacy, legal, compliance, and domain specialists address obligations and consequences that technical testing alone cannot determine. Internal auditors require sufficient AI knowledge to evaluate governance and controls while maintaining assurance independence. No single professional discipline is likely to possess all required capabilities [14], [13].

Core Questions by Assurance Function

The functions ultimately ask different but complementary questions. AI systems and business owners ask whether the systems can be operated safely and effectively for their intended purpose. AI risk management and MRM ask what can go wrong, how material the risks are, and whether they are appropriately controlled. Independent validation asks whether claims concerning design, data, performance, limitations, robustness, fairness, and intended use are supported by evidence. Cybersecurity and other specialized risk functions challenge risks within their respective domains. Internal audit asks whether governance, risk management, validation, and controls are appropriately designed and operating effectively [3], [14], [8], [10]. These questions demonstrate why generative and agentic AI cannot be assured through a single control function. Effective assurance requires coordinated evidence and specialized expertise while preserving ownership, effective challenge, escalation, and independence. These requirements provide the functional foundation for the Integrated AI Assurance Framework presented in the following section [7], [10], [9]. Table 5 summarizes the core question each function asks.

Table 5: Core Questions by Assurance Function

Function	Core Question
Business / AI System Owner	Can the system be operated safely and effectively for its intended purpose?
AI Risk / MRM	What can go wrong, how material is it, and is it appropriately controlled?
Independent Validation	Are claims about design, data, performance, limitations, fairness, robustness, and intended use supported by evidence?
Cyber / Privacy / TPRM / Legal	Are domain-specific risks and obligations appropriately evaluated and controlled?
Internal Audit	Are governance, risk management, validation, and controls appropriately designed and operating effectively?

PROPOSED INTEGRATED AI ASSURANCE FRAMEWORK

The preceding sections demonstrate that no single organizational function can provide complete assurance over generative and agentic AI. AI governance establishes organizational expectations but does not independently validate AI systems. AI risk management and model risk management (MRM) identify and manage risks but remain part of the risk-management structure. Independent validation provides effective challenge but does not own the risks being validated. Internal audit provides independent assurance but cannot substitute for

management, risk ownership, or validation. The proposed Integrated AI Assurance Framework (IAAF) connects these functions through four interdependent layers while preserving their distinct responsibilities and necessary independence [7], [14], [8], [10].

Organizations may assign responsibilities differently depending on industry, size, regulatory exposure, AI maturity, architecture and topology, materiality, autonomy, data sensitivity, potential harm, and reliance on third parties. The essential principle is that the four assurance functions must remain identifiable and coordinated even when they reside within different organizational structures [6], [12].

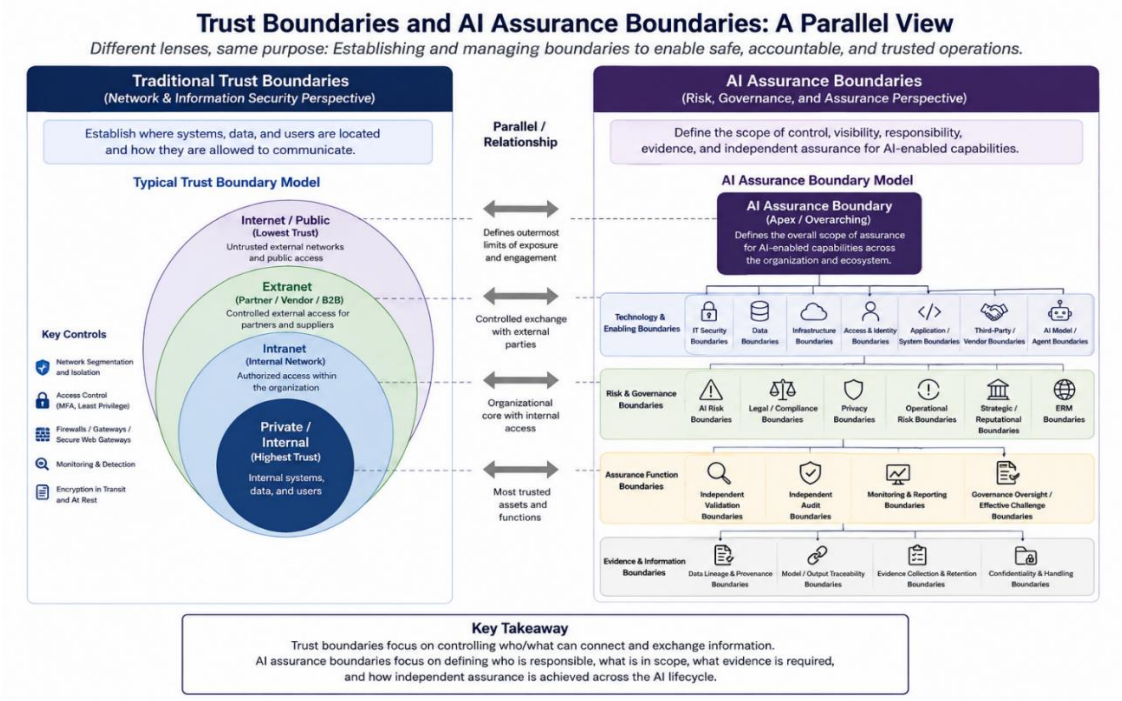

Fig. 1: AI Assurance Boundary Framework.

The Integrated AI Assurance Framework operates within an overarching AI Assurance Boundary. The boundary defines the scope of organizational responsibility, control, visibility, evidence, dependencies, and independent assurance associated with an AI-enabled capability. As illustrated in Figure 1, this boundary encompasses foundational and enabling domains—including security, data, infrastructure, identity, applications, third parties, and the AI model or agent itself—as well as the risk and governance domains that connect these technical dependencies to organizational assurance functions.

The proposed Integrated AI Assurance Framework is a coordinated organizational architecture rather than a rigid organizational chart. It combines four vertical assurance layers: AI Governance; AI Risk Management and Model Risk Management; Independent AI/Model Validation; and AI Auditing. Horizontal specialist capabilities include cybersecurity, privacy, TPRM, ERM, IT risk, architecture, data/AI engineering, legal/compliance, and business/domain expertise.

The AI Assurance Boundary defines the scope over which these functions must collectively establish evidence. Governance establishes authority, accountability, policy,

risk appetite, classification, decision rights, and oversight. Risk functions identify, assess, measure, mitigate, monitor, and document risk. Independent validation provides effective challenge of claims concerning fitness for use and control effectiveness. Internal audit independently assesses whether governance, risk management, validation, and controls are appropriately designed and operating.

The framework is cyclical: Identify -> Assess -> Control -> Validate -> Deploy -> Monitor -> Evidence -> Assure -> Escalate -> Correct -> Reassess. Specialist functions both provide evidence to and receive requirements or findings from one another. The objective is integrated assurance without collapsing risk ownership, effective challenge, or independence.

The boundary domains depicted in Figure 1 should not be interpreted as additional organizational layers within the IAAF. Rather, they identify the technical, governance, assurance, and evidentiary dimensions that establish the scope within which the four IAAF layers operate. The boundary defines what falls within the assurance responsibility; the four layers define how that responsibility is governed, managed, challenged, and independently assured.

Layer 1 — AI Governance

AI governance forms the foundation of the framework. It establishes organizational authority, accountability, policies, risk appetite, classification criteria, decision rights, and oversight mechanisms for AI. Governance should define what constitutes an AI system for organizational purposes, who owns each system, what uses are permissible or prohibited, and which systems require heightened review [6], [11], [7].

Governance also establishes escalation and risk-acceptance authority. Higher-risk AI systems should not proceed merely because technical testing has been completed; appropriate organizational authorities must determine whether residual risks are acceptable. Boards and senior executives provide oversight proportionate to organizational exposure, while AI governance or steering committees can coordinate cross-functional decisions without assuming responsibilities assigned to individual risk owners [6], [11], [7].

Layer 2 — AI Risk Management and Model Risk Management

The second layer converts governance expectations into risk-management activities. AI risk management identifies, assesses, mitigates, monitors, and documents the broader risks arising from AI. These may include operational, legal, compliance, privacy, cybersecurity, reputational, third-party, fairness, and autonomy-related risks [6], [12].

MRM operates within this broader layer where model-specific risks require specialized treatment. Its established principles—including inventory, materiality assessment, documentation, performance monitoring, change management, and lifecycle controls—remain valuable. However, generative and agentic AI may require risk management extending beyond conventional model boundaries [3].

The relationship is therefore complementary rather than competitive: MRM provides specialized model-risk capabilities, while broader AI risk management addresses risks arising from the complete AI system, its dependencies, and its organizational use. Management and

system owners remain responsible for risk; the existence of an AI governance, MRM, or risk function does not transfer that accountability away from the business [3], [6].

Layer 3 — Independent AI/Model Validation

The third layer provides effective challenge. Independent validation evaluates whether claims concerning an AI system's design, data, performance, limitations, robustness, fairness, security-relevant behavior, and intended use are adequately supported by evidence [3], [10].

Validation requirements should be proportionate to risk and adaptable to architecture. Internally developed models may permit examination of development data, methodologies, assumptions, and technical design. Externally provided foundation models may offer substantially less transparency. In these circumstances, validation should not disappear simply because traditional evidence is unavailable. Instead, validation may shift toward use-case testing, output evaluation, configuration review, vendor evidence, data-flow analysis, control testing, documented limitations, monitoring, and evaluation of the organization's implementation and use of the external model [1], [5], [21].

For agentic AI, validation should additionally consider permissions, tools, decision boundaries, autonomous actions, human intervention requirements, transaction limits, and failure behavior. The objective is to effectively challenge whether the system can reasonably be relied upon for its approved use—not an unrealistic expectation that every aspect of a complex foundation model can be independently reconstructed [1], [25], [15].

Layer 4 — AI Audit

The fourth layer provides independent assurance over overall governance, risk management, validation, and the control environment. Internal audit evaluates whether AI governance, risk management, validation, and associated controls are appropriately designed and operating effectively [14], [13], [10].

AI audit therefore asks a fundamentally different question from validation. Validation asks whether evidence supports the reliability and appropriate use of a particular AI system. Audit asks whether the organization has established and consistently operates an effective system for governing and managing AI risk. This may include evaluating inventories, ownership, risk classification, approvals, validation processes, monitoring, cybersecurity controls, third-party oversight, exception management, evidence retention, and escalation [14], [13], [10]. Maintaining this distinction helps prevent internal audit from becoming the owner of AI risk or performing management responsibilities that would compromise its independence [14], [8].

Connecting the Four Layers

The framework should operate as a connected assurance cycle rather than four isolated activities:

AI Governance → AI Risk Management/MRM → Independent Validation → AI Audit

Information also moves in the opposite direction. Validation findings can identify weaknesses requiring additional risk treatment or governance decisions. Audit findings may reveal systemic deficiencies requiring changes to policies, responsibilities, validation standards, or risk appetite. Incidents, monitoring results, regulatory developments, and material system changes may similarly trigger reassessment across multiple layers [6], [7], [10].

Adjacent organizational functions connect across this architecture. ERM places AI within enterprise risk appetite and aggregation processes. IT risk addresses technology dependencies and operational resilience. Cybersecurity evaluates adversarial threats, identity and access management, and technical security. Third-party risk management addresses dependencies on providers and concentration. Legal, privacy, and compliance functions interpret applicable obligations. AI governance and steering committees coordinate decisions and escalation. These functions should contribute specialized expertise without obscuring who owns the underlying risk or who provides independent assurance [6], [7], [20], [5].

A Risk-Based and Adaptable Architecture

Not every AI system requires the same level of assurance. A low-impact productivity tool should not necessarily receive the same validation and audit treatment as AI influencing credit, employment, healthcare, security, or autonomous transactions. Assurance intensity should increase with materiality, potential harm, autonomy, sensitive data access, regulatory exposure, architectural complexity, and third-party dependencies [6], [1], [12].

The IAAF therefore establishes functional requirements rather than a universal organizational structure. Organizations can centralize some responsibilities, distribute others, or use hybrid approaches while maintaining four principles: clear ownership, effective risk management, independent challenge, and independent assurance [6], [12], [7].

The central proposition of the framework is consequently straightforward: AI assurance is not a function but an architecture. Effective assurance emerges when governance establishes accountability, risk functions manage AI and model risks, independent validation challenges the evidence supporting system use, and internal audit determines whether the resulting governance and control environment works as intended. The next challenge is translating this architecture into practical implementation across different organizational and AI operating contexts [7], [8], [10], [9].

Assurance Sufficiency Within the Boundary

Operationally, sufficiency begins with material assurance claims. For each claim, the assurance plan should record the intended use and consequences; the evidence required; the evidence available; the evidence unavailable; the tests performed; the limitations; the compensating controls; the monitoring; and the accountable authority for residual uncertainty. Independent validation should challenge whether the evidence is relevant, reliable, sufficiently independent, current, and representative of the organization's actual configuration and use.

The evidence ladder is not a mechanical scoring model. Direct testing may be weak if test cases are unrepresentative, while a scoped independent assessment may be strong for the claim it actually covers. Certifications and provider reports must be evaluated for scope, exclusions, period, methodology, and applicability. Contractual representations create accountability and remedies but do not prove technical performance. Self-attestation may support low-risk uses but is rarely sufficient alone for high-consequence reliance.

A sufficiency decision should expire or be reconsidered when material conditions change. Provider releases, altered topology, new data, expanded autonomy, changed permissions, incidents, performance degradation, regulatory developments, and expired assurance reports can all invalidate the prior decision. Where validation concludes that a material claim lacks sufficient evidence, deployment should be restricted, conditioned, deferred, or rejected unless an authorized risk-acceptance process explicitly addresses the unresolved uncertainty.

OPERATIONALIZING THE FRAMEWORK: MULTIPLE PATHS TO IMPLEMENTATION

Enterprise AI Inventory and Ownership

The inventory should operate as the entry point to assurance rather than as a static model list. Intake should identify the intended use, accountable business and system owners, provider and foundation model where known, topology, data classifications, retrieval sources, integrations, permissions, autonomous capability, affected populations, consequentiality, third-party dependencies, validation status, monitoring, and retirement dependencies. Discovery mechanisms should also identify embedded or employee-acquired AI that bypasses formal procurement.

Completeness should be tested through reconciliation with procurement, architecture repositories, cloud and API usage, software inventories, data-loss-prevention observations, business attestations, and third-party records. Unowned or unclassified AI should not be allowed to progress to material use. Internal audit may later test inventory completeness, but management remains responsible for establishing and maintaining it.

AI Risk Classification and Materiality

Classification should combine financial materiality with consequentiality, autonomy, sensitive-data access, scale, reversibility of harm, regulatory exposure, affected populations, opacity, third-party dependence, and architectural complexity. A use may be financially immaterial while creating significant legal, safety, employment, privacy, or reputational consequences.

The classification should determine minimum governance, validation, approval, monitoring, documentation, and audit expectations. Materiality must be reassessed when scope, data, users, permissions, provider, integrations, or consequences change. Risk and MRM functions may administer classification, but the business owner remains accountable for complete and accurate use-case information.

Architecture/Topology and Third-Party Dependency Assessment

A boundary and dependency assessment should map the model, hosting environment, interfaces, enterprise data, RAG sources, identity, controls, tools, external services, human decisions, and downstream actions. The map should show where data and authority cross organizational or trust boundaries, which components are directly controlled, and which evidence depends on a provider.

The assessment should identify single points of failure, concentration, unobservable change, data residency, exit constraints, and components whose failure could alter consequential outcomes. Architecture, cybersecurity, privacy, TPRM, and validation should contribute specialized challenges while the system owner maintains the authoritative boundary record.

Risk-Based Assurance Requirements

The organization should translate classification and boundary information into an assurance plan. The plan specifies mandatory control domains, evidence, validation depth, approval authority, monitoring frequency, change triggers, retention, and escalation. Higher consequence, autonomy, sensitive-data access, opacity, regulatory exposure, or third-party dependence should increase assurance intensity; reversibility, restricted scope, isolation, and effective human control may reduce it.

Risk-based does not mean discretionary. Governance should establish nonwaivable minimums and the authority required for exceptions. A high-risk use should not receive reduced validation simply because primary evidence is difficult to obtain; difficulty obtaining evidence is itself a factor in deciding whether reliance is defensible.

Validation Requirements and Evidence Standards

Validation scope should be claim- and use-case-based. It may include conceptual soundness, data, intended use, performance, robustness, fairness, explainability, cybersecurity-relevant behavior, human controls, permissions, monitoring, and limitations. The validation plan should identify acceptance criteria, test populations, versions, evidence sources, independence, unresolved limitations, and revalidation triggers.

Evidence should be maintained in an evidence register linking each material claim to supporting artifacts and residual uncertainty. Where direct inspection is unavailable, validation should combine behavioral testing, configuration and data-flow review, scoped independent evidence, provider documentation, contracts, monitoring, and compensating controls. The validator should state whether the combined evidence is sufficient, conditionally sufficient, or insufficient for the intended use.

Approval and Deployment Gates

Deployment gates should require verified ownership, classification, boundary documentation, required controls, completed validation, resolved or accepted findings, monitoring readiness, incident and suspension procedures, and evidence retention. The

approving authority should have the power to approve, condition, restrict, defer, or block deployment.

Gate records should preserve the evidence reviewed, objections, conditions, accountable decision maker, residual-risk rationale, expiration or review date, and triggers for renewed approval. Technical completion alone should not override unresolved legal, risk, security, privacy, or validation concerns.

Continuous Monitoring and KRIs

Monitoring should be derived from the material claims and failure modes established during risk assessment and validation. Measures may include accuracy and severity of errors, subgroup outcomes, hallucination or grounding failures, drift, provider and model changes, source changes, prompt-injection attempts, control failures, data leakage, permissions violations, overrides, complaints, incidents, and high-risk agent actions.

Thresholds should identify warning, escalation, restriction, and suspension conditions. Owners should investigate breaches and document corrective action; validators should receive evidence relevant to continuing reliance; and governance should aggregate systemic trends and concentration. Monitoring that cannot trigger action is observation, not control.

Audit Evidence and Documentation

Evidence should allow an independent party to determine what was known, claimed, tested, approved, changed, monitored, excepted, and remediated. Repositories should link inventory, classification, boundary maps, risk assessments, validation, approvals, changes, monitoring, incidents, exceptions, and retirement records using stable system and version identifiers.

Evidence quality includes integrity, provenance, accessibility, retention, and consistency—not merely document volume. Internal audit should assess whether evidence demonstrates that required processes and controls operate, while management remains responsible for producing and maintaining the records.

Exceptions, Escalation, and Risk Acceptance

Exceptions should identify the unmet requirement, the reason, the affected use, the consequences, compensating controls, the accountable owner, the effective-challenge position, the approving authority, the duration, the review date, and the remediation plan. Validation or specialist-risk objections should remain visible in the record even when management accepts the risk.

Escalation should increase with the severity of consequences and disagreement. No committee should silently convert an unresolved evidence deficiency into acceptance through consensus. Acceptance must be explicit, time-bound, within delegated authority, and subject to monitoring and reconsideration when assumptions change.

Model and AI System Retirement

Retirement should address more than turning off a model endpoint. The organization should revoke agent and service credentials, remove permissions and integrations, decommission retrieval indexes, preserve required evidence, dispose of data appropriately, terminate or modify contracts, redirect downstream dependencies, and verify that shadow copies or fallback workflows do not continue unauthorized use.

Retirement plans should also address provider-initiated model or API withdrawal and emergency replacement. Material migrations require reassessment because a nominally equivalent replacement may change behavior, evidence, limitations, data processing, or contractual protection.

Illustrative Implementation Options by Organizational Context

A regulated financial institution may build on existing MRM, independent validation, compliance, fair lending, and TPRM capabilities while expanding the boundary to include cybersecurity, data, architecture, and agentic controls. A large enterprise consuming external AI services may centralize policy and classification while federating business ownership and relying heavily on security, privacy, architecture, procurement, and vendor evidence. An organization operating private AI may obtain greater observability but assume greater responsibility for infrastructure, model lifecycle, data governance, security, testing, and monitoring.

A smaller organization without dedicated MRM can preserve the framework's functions without recreating a large-bank bureaucracy. It may combine first- and second-line activities, use external validation for material systems, establish explicit executive risk acceptance, and obtain periodic independent audit or assurance. Adaptation is legitimate only if ownership, expertise, challenge, evidence, escalation, and assurance remain identifiable. Table 6 summarizes the operational decision, minimum evidence, accountable role, and independent challenge mechanism for each implementation step described above (Sections 9.1–9.11).

Table 6: Section 9 Operationalization Matrix

Section	Operational Decision	Minimum Evidence / Artifact	Primary Accountable Role	Independent Challenge / Assurance
9.1	Enterprise AI Inventory and Ownership	AI inventory + owner + boundary map	AI System / Business Owner	MRM/Risk; Audit tests completeness
9.2	AI Risk Classification and Materiality	Risk/materiality classification	Risk Owner / MRM	Independent validation; compliance where applicable
9.3	Architecture/Topology and Third-Party Dependency Assessment	Architecture/data-flow/dependency map	Architecture + System Owner	Cyber/Privacy/TPRM; validation
9.4	Risk-Based Assurance Requirements	Assurance plan and required evidence	Risk Owner / MRM	Independent validation
9.5	Validation Requirements and Evidence Standards	Validation report/evidence register	Model/AI Risk Owner	Independent validation; Audit assesses process

9.6	Approval and Deployment Gates	Gate decision/approval record	Named Deployment Authority	Validation/risk objections; Audit later assesses governance
9.7	Continuous Monitoring and KRIs	Monitoring/KRI dashboard + incidents	System Owner	Validation/risk monitoring; Audit samples effectiveness
9.8	Audit Evidence and Documentation	Evidence repository/audit trail	System/Risk Owner	Internal Audit
9.9	Exceptions, Escalation, and Risk Acceptance	Exception/risk-acceptance record	Designated Risk-Acceptance Authority	Effective challenge + Internal Audit
9.10	Model and AI System Retirement	Retirement/decommission record	System Owner	Risk/Security/Privacy; Audit
9.11	Illustrative Implementation Options by Organizational Context	Context-specific operating model	Executive / Governance Authority	Internal Audit assesses governance design

Table 7 ranks categories of evidence by relative strength and describes how each may be used where observability is incomplete.

Table 7: Evidence Sufficiency Ladder

Evidence Tier	Example	Relative Strength	Use in Incomplete Observability
1. Direct primary evidence	Independent testing, raw logs, reproducible results, configuration/version records	Highest when relevant and reliable	Preferred for material claims within organizational control
2. Independent external evidence	Third-party assessment, audit/assurance report, certification with relevant scope	High to moderate	Useful where direct inspection is unavailable; scope/recency must be evaluated
3. Provider technical evidence	Model/system cards, evaluation reports, security/AI documentation, change notices	Moderate	Can support assurance but should be challenged against intended use
4. Contractual representation	Warranties, commitments, audit rights, incident/change notification	Moderate to low for technical truth; strong for accountability/remedy	Compensating governance evidence, not a substitute for behavioral validation
5. Self-attestation	Vendor questionnaire or management assertion	Lowest alone	May be acceptable for low-risk use; weak basis for high-consequence reliance without corroboration

Table 8 illustrates how assurance expectations scale with consequence, autonomy, and opacity across five representative use cases.

Table 8: Illustrative Assurance Intensity by Use Case

Use Case	Consequence	Autonomy	Opacity	Illustrative Assurance Expectation
Internal meeting summarization	Low	Low	Potentially high	Behavioral testing, confidentiality controls, human review, monitoring; limited provider transparency may be tolerable.
Customer-facing advisory assistant	Moderate	Low/Moderate	Medium/High	Stronger testing, content controls, escalation, logging, monitoring, provider evidence, defined limitations.
Employment screening recommendation	High	Moderate	Medium/High	Validated job-related criteria, fairness/proxy testing, decision traceability, substantive human review, version/evidence retention.
Credit underwriting/adverse action	High / regulated	Moderate/High	Medium/High	High evidence threshold; validation, lineage, fair-lending testing, specific adverse-action rationale, monitoring, reconstructability.
Autonomous agent with transaction authority	Potentially critical	High	Variable	Identity/least privilege, transaction limits, segregation, containment, logging, intervention/kill mechanisms, rigorous validation and monitoring.

APPLYING THE FRAMEWORK TO HIGH-RISK AI USE CASES

AI-Assisted Credit and Lending Decisions

Fair lending is a useful stress test because credit decisions combine predictive performance with nondiscrimination, adverse-action reasoning, documentation, validation, monitoring, and accountability. Machine learning may improve underwriting, while nonlinear relationships and large feature sets create interpretability and proxy-variable challenges. Hall et al. emphasize governance, human review, and collaboration among legal, compliance, audit, risk, and data-science functions [30].

An allegation that a denial resulted from race, sex, age, national origin, ZIP code, or another prohibited or proxy factor does not, in itself, prove discrimination. The institution nevertheless requires sufficient evidence to demonstrate which factors actually drove the decision and whether the process complied with approved policy, fair-lending requirements, and validated model behavior.

MRM contributes institutional defensibility by establishing inventory, ownership, intended use, independent validation, variable/proxy review, outcome testing, limitations, monitoring, change control, overrides, and effective challenge [3]. Within the broader AI Assurance Boundary, upstream data transformations, third-party services, policy controls, human interventions, and downstream adverse-action explanations must also be considered.

Sound MRM cannot guarantee that a lender will never be challenged and does not create a legal safe harbor. It can materially improve the institution's ability to reconstruct

a challenged decision, identify error, demonstrate applied controls, and show appropriate governance and challenge [17].

AI-Assisted Employment Decisions

Research on algorithmic hiring cautions that vendors' claims of objectivity or bias mitigation cannot substitute for examining how targets, features, training data, validation, and organizational use interact. Raghavan et al. found substantial variation in disclosed development and bias-mitigation practices among algorithmic assessment vendors, reinforcing the need for contextual, technical, and legal challenge rather than reliance on marketing claims [31].

Under Title VII and the Uniform Guidelines, an AI-supported screening procedure should be evaluated as an employment selection procedure. Assurance should address job-relatedness, validity evidence, adverse impact, accommodation, recordkeeping, and whether the employer can explain and reconstruct how the procedure influenced an employment decision. EEOC materials also identify disability risks where tools screen out individuals because of inaccessible interfaces or disability-related characteristics unrelated to effective job performance [16].

Applied through the IAAF, governance defines permissible employment uses and decision rights; risk and compliance functions identify discrimination, privacy, and vendor risks; independent validation challenges job-relatedness, data, proxy variables, subgroup outcomes, human review, and reconstructability; and internal audit assesses whether inventories, approvals, validation, monitoring, complaints, exceptions, and remediation operate consistently.

AI-assisted hiring and candidate screening create a parallel assurance problem. Apparently neutral variables can reproduce historical patterns, act as proxies, or interact with other controls to create differential outcomes. Fairness therefore cannot be established merely by omitting protected characteristics from the primary model.

If similarly situated applicants receive materially different outcomes, the difference does not establish unlawful discrimination. It becomes an assurance concern when the organization cannot reconcile the result with validated job-related criteria or reconstruct the decision pathway. Relevant questions include: who established the screening criteria; which data and derived attributes were used; whether proxies and disparate outcomes were tested; which model/control versions participated; whether human review was substantive; and whether a rejected outcome can be reconstructed.

Agentic AI in Enterprise Operations

Consider a bounded enterprise scenario in which an AI agent can read support tickets, query customer records, generate remediation steps, and execute approved changes through administrative tools. The business objective may be operational efficiency, but the Assurance Boundary includes the foundation model, ticket content, retrieved knowledge, identity provider, credentials, tools, target systems, transaction limits, approval workflow, logs, human operators, and external provider dependencies.

The system should be validated against ordinary tasks and failure conditions: malicious instructions embedded in tickets or retrieved documents; ambiguous authorization; attempts to access unrelated customers; chained tool calls; actions exceeding transaction or change limits; provider or model changes; unavailable human approvers; incomplete rollback; and emergency suspension. Evidence should show which identity acted, what information and instructions it used, which permissions and controls were evaluated, what action occurred, whether intervention was possible, and how the event was reconstructed [1], [25], [15].

The scenario demonstrates functional integration. The system owner manages operation; cybersecurity establishes identity, permission, containment, and incident controls; operational risk addresses resilience and recovery; TPRM evaluates external dependencies; independent validation challenges whether the complete system is fit for the approved autonomy; governance accepts or rejects residual consequence risk; and internal audit assesses whether the architecture of ownership, testing, monitoring, escalation, and remediation operates as designed.

Agentic AI provides the clearest example of the transition from output risk to consequence risk. Assurance should address autonomous actions, permissions, identity, segregation of duties, transaction limits, human intervention, logging, escalation, containment, and emergency suspension. A single agentic event may simultaneously implicate model behavior, cybersecurity, operational risk, third-party risk, governance, validation, and independent assurance [1], [25], [15].

SYNTHESIS: WHAT CHANGES—AND WHAT MUST REMAIN

What Traditional MRM Still Gets Right

The analysis confirms that these practices are not artifacts of a particular generation of statistical modeling. They address the enduring organizational problem of justified reliance on systems whose outputs can be wrong, misused, changed, or misunderstood. Their value increases—not decreases—when contemporary AI creates greater uncertainty, provided their application expands to the material system and use rather than stopping at an inaccessible model.

Traditional MRM remains highly relevant because inventory, ownership, materiality, documentation, independent validation, monitoring, limitations, change management, governance, effective challenge, and retirement address enduring problems of model reliance [3].

Which Legacy MRM Assumptions Require Adaptation

The necessary adaptation is evidentiary and architectural. Validation may no longer begin with complete access to model construction, and monitoring may no longer track a stable model alone. Assurance must instead follow dependencies, observable behavior, controls, change, human decisions, and consequences while identifying unavailable evidence as residual uncertainty.

What changes are assumptions about boundaries, transparency, controllability, change, third-party dependence, evidence availability, multidisciplinary expertise, and autonomy. Contemporary AI may be distributed, externally controlled, dynamically updated, context-sensitive, and partially opaque.

How AI Risk Management Extends the MRM Boundary

This extension does not dissolve MRM into general risk management. MRM retains specialized responsibility for model identification, risk classification, validation standards, limitations, monitoring, and effective challenge. Broader AI risk management connects those capabilities to risks arising from data, infrastructure, security, privacy, law, third parties, human use, and autonomy that cannot be evaluated solely by model performance.

The AI Assurance Boundary provides the conceptual bridge: it preserves MRM's specialized role while requiring evidence from cybersecurity, privacy, architecture, data governance, TPRM, legal/compliance, business owners, engineering, and human oversight where those dependencies materially affect reliance.

Why Independent Validation Remains Essential

Opacity changes the method of validation, not the need for it. Independent validators may rely more heavily on intended-use testing, behavioral evaluation, system configuration, data flows, controls, provider evidence, monitoring, and explicit limitations. Their essential contribution remains an objective determination of whether the evidence supports reliance and whether unresolved uncertainty has been presented to the proper authority.

As systems become more opaque and controls become more layered, independent validation becomes more, rather than less, important. The object and method of validation may change, but objective effective challenge remains necessary.

Why AI Audit Cannot Substitute for Risk Ownership or Validation

The synthesis also clarifies that independence exists for different purposes. Validation provides effective challenge of a system and its intended use, while audit assures the organization's governance and control environment. Combining those purposes can leave management without clear ownership, validators without sufficient independence, or audit responsible for activities it must later assess.

Internal audit independently assesses whether governance, risk management, validation, and controls are appropriately designed and operating; it should not become the owner or operator of those processes [14], [8], [10].

How Assurance Changes for Black-Box and Third-Party AI

The critical finding is that external control does not move a material dependency outside the Assurance Boundary. It changes the evidence available and may require compensating controls, contractual rights, narrower use, stronger monitoring, or rejection. The

organization remains accountable for deciding whether reliance is justified even when the provider controls the underlying model.

Black-box conditions shift validation toward observable behavior, intended-use testing, interfaces, provider evidence, system configuration, contractual controls, monitoring, and compensating controls. This adaptation should not allow opacity to become an excuse for lack of accountability.

From Compliance Alignment to Meaningful AI Assurance

Taken together, the analysis answers the research questions by replacing a search for one universal AI owner or perfect model transparency with a coordinated architecture of sufficient evidence. Governance assigns accountability and decision rights; risk functions manage relevant exposures; validation challenges the evidence supporting use; audit assesses whether the architecture operates; and the Assurance Boundary ensures that material dependencies are neither ignored nor hidden by organizational ownership.

Compliance alignment and assurance are related but not identical. An organization may satisfy a documentation requirement without evidence that a control works or possess strong technical evidence while failing a legal disclosure obligation. Meaningful assurance requires both applicable compliance and sufficient evidence for justified reliance.

The synthesis therefore resolves the central issue as one of proportional sufficiency rather than perfection: MRM should not be interpreted as requiring traceability and explanation of every byte, bit, parameter, or internal operation. It should require evidence sufficient for the consequence, materiality, legal obligation, and risk of the intended use. Where sufficient evidence cannot be obtained, the resulting uncertainty must be treated as risk rather than ignored.

IMPLICATIONS FOR PRACTICE

Implementation should begin with a minimum viable assurance architecture: an enterprise definition and inventory of AI; named business and system owners; risk and materiality classification; an Assurance Boundary record; minimum evidence and validation requirements; approval and risk-acceptance authority; continuous monitoring; incident and suspension procedures; and independent audit coverage. Organizations can mature these capabilities without waiting to establish a single centralized AI department.

Boards and executive committees should receive aggregated information about material AI uses, unresolved validation limitations, accepted uncertainty, incidents, concentration, and remediation—not merely policy adoption or project counts. Risk leaders should define nonwaivable assurance requirements and their escalation procedures. Technology and business leaders should demonstrate control operation and evidence. Validators should state the limits of what can be concluded. Internal audit should determine whether the combined system produces reliable challenge, decisions, monitoring, and correction.

Organizations with limited internal expertise should distinguish resource constraints from independence. External specialists may provide technical testing, legal analysis,

validation, or audit support, but accountable decisions remain with the organization. Smaller organizations should narrow autonomy and use, demand stronger provider evidence, or decline high-consequence deployments when they cannot obtain the expertise and evidence necessary for defensible reliance. Table 9 presents a RACI+ accountability matrix that maps each lifecycle activity to the responsible, accountable, consulted, effective-challenge, and independent-assurance parties.

Table 9: RACI+ Accountability Matrix

Lifecycle Activity	Responsible	Accountable	Consulted	Effective Challenge	Independent Assurance
AI use-case proposal	Business/ Product Owner	Business Executive	AI Governance, Legal, Architecture	Risk/MRM	Internal Audit (later assurance)
Risk/ materiality classification	Risk/MRM	Risk Executive / designated authority	Business, Legal/ Compliance, Security, Privacy	Independent Validation where material	Internal Audit
Architecture/data-flow review	Architecture/Engineering	System Owner	Security, Privacy, TPRM, MRM	Specialist risk functions	Internal Audit
Third-party AI due diligence	TPRM	Vendor/ Business Owner	Security, Privacy, Legal, MRM, Architecture	MRM/Validation for model evidence	Internal Audit
Model/AI validation	Independent Validation	Validation Executive / designated authority	MRM, Security, Privacy, domain experts	Validation itself provides effective challenge	Internal Audit
Deployment approval	System/Business Owner	Named Deployment Authority	Risk, Validation, Security, Legal/Compliance	Risk/Validation objections documented	Internal Audit
Continuous monitoring	System Owner / Operations	Business/System Owner	MRM, Security, TPRM, Compliance	Validation/ Risk	Internal Audit
Exception/risk acceptance	Control/Risk Owner	Designated Risk-Acceptance Authority	Affected specialists	Independent risk/validation challenge	Internal Audit
Retirement	System Owner	Business Owner	Security, Privacy, TPRM, Records, MRM	Risk functions	Internal Audit

RESEARCH LIMITATIONS AND FUTURE RESEARCH

Finally, the AI Assurance Boundary and Assurance Sufficiency Principle are proposed conceptual devices rather than empirically validated standards. Their usefulness depends on whether organizations can apply them consistently, whether reviewers reach comparable sufficiency judgments, and whether the resulting decisions reduce control gaps without producing unmanageable assurance burden. Empirical case studies, comparative organizational research, and development of claim-to-evidence methods are therefore needed.

This article has several limitations. First, AI governance and assurance requirements are evolving rapidly. Regulatory guidance, international standards, professional practices, and organizational approaches will continue to change as generative and agentic AI mature. Consequently, the Integrated AI Assurance Framework should be viewed as an adaptable architecture rather than a static compliance model [3], [1], [29].

Second, organizations differ substantially in regulatory exposure, risk-management maturity, technical capabilities, AI architecture, and reliance on third-party providers. The framework therefore cannot prescribe a single organizational structure or assurance approach applicable to every environment. Its implementation requires adaptation to

organizational context, materiality, data sensitivity, potential harm, autonomy, and third-party dependency [6], [12].

Third, empirical evidence concerning organizational assurance of generative AI remains limited, and evidence regarding agentic AI is even less mature. The framework developed here is primarily conceptual and applied; future empirical research should examine how organizations allocate AI governance, MRM, validation, cybersecurity, compliance, and audit responsibilities in practice and whether different structures produce measurable differences in risk and assurance outcomes.

Several issues warrant focused follow-on research. These include defining the boundary between AI MRM and AI audit; developing practical methods for independently validating third-party foundation models; examining assurance requirements for high-impact uses such as lending and employment; identifying appropriate controls and audit approaches for autonomous, agentic AI; and developing an AI assurance maturity model. Such research can test, refine, and extend the proposed framework as organizational practices and AI technologies continue to evolve [4], [15], [9].

CONCLUSION

Effective assurance for generative and agentic AI cannot be achieved by AI governance, MRM, AI risk management, validation, AI auditing, cybersecurity, ERM, TPRM, or compliance acting independently. Traditional MRM remains an essential foundation, but contemporary AI increasingly exceeds the organizational, technical, and evidentiary boundaries within which conventional MRM has historically operated.

The analysis further identifies the principal challenges to contemporary AI assurance as shifting system boundaries, incomplete observability, third-party dependence, dynamic change, increasing autonomy, fragmented organizational responsibilities, and limited evidence available for independent validation. These conditions do not make traditional MRM obsolete; rather, they require its core principles to operate within a broader system-level assurance architecture.

The proposed Integrated AI Assurance Framework responds by preserving accountable ownership, risk management, effective challenge, and independent assurance while expanding the scope of evidence through the AI Assurance Boundary. The framework does not presume that complete model transparency is attainable. Instead, it requires organizations to determine whether available evidence is sufficient for justified reliance on a specific use, to characterize unavailable evidence as uncertainty, to apply compensating controls where feasible, and to accept or reject the resulting residual risk.

Accordingly, opacity does not eliminate accountability; it increases uncertainty. The appropriate response is neither to abandon MRM nor to impose technically unrealistic expectations of perfect traceability and explainability. It is to establish proportionate assurance requirements that reflect materiality, consequentiality, autonomy, regulatory obligation, architecture, third-party dependence, and observability. For engineering and computing practice, this means that assurance must follow the operational AI system across model, data, retrieval, infrastructure, identity, permissions, interfaces, controls, and autonomous actions rather than stop at the nominal model boundary.

In concise terms: governance establishes accountability and decision rights; AI risk management and MRM manage relevant risks; validation provides effective challenge; audit independently assesses whether the assurance framework and its controls are working; and the organizational arrangement must adapt to context while preserving sufficient evidence for defensible reliance.

DECLARATIONS

Ethics and Consent

Not applicable. This conceptual and applied framework-development study did not involve human participants, identifiable personal data, animals, or an intervention requiring ethics approval.

Data and Code Availability

Not applicable. No original empirical dataset, software code, or trained model was generated or analyzed for this conceptual and applied framework-development study. The public guidance, standards, professional materials, and scholarly sources used in the analysis are identified in the reference list.

Conflicts of Interest

The authors declare no conflicts of interest.

Funding

No specific funding was received for this research.

Author Contributions

T.G. and J.B.: Conceptualization, methodology, investigation, writing—original draft, and writing—review and editing. Both authors reviewed and approved the final manuscript and accept responsibility for its content.

Use of Generative AI

OpenAI ChatGPT (GPT-5.6 Sol) was used as an editorial assistance tool for language refinement, organization, and manuscript preparation. The authors reviewed, revised, and verified the resulting text, sources, arguments, and conclusions and remain fully responsible for the accuracy, originality, and integrity of the manuscript.

Acknowledgements

None.

Preprint / Prior Version

No preprint or prior published version of this manuscript is declared.

REFERENCES

[1] C. Autio et al., Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1. Gaithersburg, MD, USA: National Institute of Standards and Technology, 2024, doi: 10.6028/NIST.AI.600-1.

[2] R. Bommasani et al., On the Opportunities and Risks of Foundation Models. Stanford, CA, USA: Stanford Center for Research on Foundation Models, 2021. [Online]. Available: https://crfm.stanford.edu/report.html

[3] Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation, and Office of the Comptroller of the Currency, Revised Guidance on Model Risk Management, SR 26-2, Apr. 17, 2026. [Online]. Available: https://www.federalreserve.gov/supervisionreg/srletters/SR2602.htm

[4] A. Bhattacharyya, Y. Yu, H. Yang, R. Singh, T. Joshi, J. Chen, and K. Yalavarthy, "Model risk management for generative AI in financial institutions," *arXiv preprint arXiv:2503.15668*, 2025, doi: 10.48550/arXiv.2503.15668.

[5] Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation, and Office of the Comptroller of the Currency, Interagency Guidance on Third-Party Relationships: Risk Management, 2023. [Online]. Available: https://www.federalreserve.gov/supervisionreg/srletters/sr2304.htm

[6] E. Tabassi, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1. Gaithersburg, MD, USA: National Institute of Standards and Technology, 2023, doi: 10.6028/NIST.AI.100-1.

[7] U.S. Government Accountability Office, Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities, GAO-21-519SP, Washington, DC, USA, 2021. [Online]. Available: https://www.gao.gov/products/gao-21-519sp

[8] The Institute of Internal Auditors, The IIA's Three Lines Model: An Update of the Three Lines of Defense. Lake Mary, FL, USA, 2020. [Online]. Available: https://www.theiia.org/en/content/position-papers/2020/the-iias-three-lines-model-an-update-of-the-three-lines-of-defense/

[9] D. S. Schiff, S. Kelley, and J. Camacho Ibanez, "The emergence of artificial intelligence ethics auditing," Big Data & Society, 2024, doi: 10.1177/20539517241299732.

[10] I. D. Raji et al., "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in Proc. 2020 Conf. Fairness, Accountability, and Transparency, 2020, pp. 33–44, doi: 10.1145/3351095.3372873.

[11] ISO/IEC 42001:2023, Information Technology—Artificial Intelligence—Management System, International Organization for Standardization, Geneva, Switzerland, 2023. [Online]. Available: https://www.iso.org/standard/42001.html

[12] ISO/IEC 23894:2023, Information Technology—Artificial Intelligence—Guidance on Risk Management, International Organization for Standardization, Geneva, Switzerland, 2023. [Online]. Available: https://www.iso.org/standard/77304.html

[13] The Institute of Internal Auditors, The IIA's Artificial Intelligence Auditing Framework. Lake Mary, FL, USA, 2024. [Online]. Available:

https://www.theiia.org/en/content/tools/professional/2023/the-iias-updated-ai-auditing-framework/

[14] The Institute of Internal Auditors, Global Internal Audit Standards, 2024 ed. Lake Mary, FL, USA, 2024. [Online]. Available: https://www.theiia.org/en/standards/2024-standards/global-internal-audit-standards/

[15] E. Debenedetti et al., "AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents," in Advances in Neural Information Processing Systems 37, 2024. [Online]. Available: https://arxiv.org/abs/2406.13352

[16] U.S. Equal Employment Opportunity Commission, "Artificial intelligence and algorithmic fairness," and Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. pt. 1607. [Online]. Available: https://www.eeoc.gov/ai. [Accessed: Sep. 1, 2026].

[17] Equal Credit Opportunity Act, 15 U.S.C. § 1691 et seq.; Regulation B, 12 C.F.R. pt. 1002; Consumer Financial Protection Bureau, Circular 2022-03, Adverse Action Notification Requirements in Connection with Credit Decisions Based on Complex Algorithms, May 26, 2022. [Online]. Available: https://files.consumerfinance.gov/f/documents/cfpb_2022-03_circular_2022-05.pdf

[18] European Parliament and Council of the European Union, Regulation (EU) 2024/1689, Artificial Intelligence Act, Off. J. Eur. Union, 2024. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

[19] New York City Department of Consumer and Worker Protection, Automated Employment Decision Tools, Local Law 144 Implementation Guidance, New York, NY, USA. [Online]. Available: https://home4.nyc.gov/site/dca/about/automated-employment-decision-tools.page

[20] C. Pascoe, S. Quinn, and K. Scarfone, The NIST Cybersecurity Framework (CSF) 2.0, NIST CSWP 29. Gaithersburg, MD, USA: National Institute of Standards and Technology, 2024, doi: 10.6028/NIST.CSWP.29.

[21] M. Mitchell et al., "Model cards for model reporting," in Proc. Conf. Fairness, Accountability, and Transparency, 2019, pp. 220–229, doi: 10.1145/3287560.3287596.

[22] T. Gebru et al., "Datasheets for datasets," Commun. ACM, vol. 64, no. 12, pp. 86–92, 2021, doi: 10.1145/3458723.

[23] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," Proc. Mach. Learn. Res., vol. 81, pp. 77–91, 2018. [Online]. Available: https://proceedings.mlr.press/v81/buolamwini18a.html

[24] L. Huang et al., "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," ACM Trans. Inf. Syst., vol. 43, no. 2, Art. no. 42, 2025, doi: 10.1145/3703155.

[25] A. Vassilev et al., Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations, NIST AI 100-2e2025. Gaithersburg, MD, USA: National Institute of Standards and Technology, 2025, doi: 10.6028/NIST.AI.100-2e2025.

[26] J. Su, J. P. Zhou, Z. Zhang, P. Nakov, and C. Cardie, "Towards more robust retrieval-augmented generation: Evaluating RAG under adversarial poisoning attacks," *arXiv preprint arXiv:2412.16708*, 2024, doi: 10.48550/arXiv.2412.16708.

[27] R. Parasuraman and D. H. Manzey, "Complacency and bias in human use of automation: An attentional integration," Human Factors, vol. 52, no. 3, pp. 381–410, 2010, doi: 10.1177/0018720810376055.

[28] K. Goddard, A. Roudsari, and J. C. Wyatt, “Automation bias: A systematic review of frequency, effect mediators, and mitigators,” J. Amer. Med. Inform. Assoc., 2012, doi: 10.1136/amiajnl-2011-000089.

[29] G. Romeo and D. Conti, “Exploring automation bias in human-AI collaboration: A review and implications for explainable AI,” AI & Society, 2026, doi: 10.1007/s00146-025-02422-7.

[30] P. Hall et al., “A United States fair lending perspective on machine learning,” Front. Artif. Intell., vol. 4, Art. no. 695301, 2021, doi: 10.3389/frai.2021.695301.

[31] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, “Mitigating bias in algorithmic hiring: Evaluating claims and practices,” in Proc. 2020 Conf. Fairness, Accountability, and Transparency, 2020, pp. 469–481, doi: 10.1145/3351095.3372828.