SAFEGUARD

Kingfisher Advisory

AI Implementation Assessment Methodology

Independent, NIST-informed review of connected AI workflows

November 2026

Public methodology | Version 1.1

KINGFISHER ADVISORY LLC | BOISE, IDAHO

Contents

Public methodology and document control.

Purpose and scope 3

Framework foundation 4

Framework crosswalk 5

Governance and access 6

Data and action boundaries 7

Platform and operations 8

Assessment and evidence 9

Assessment lifecycle 10

Risk classification and decisions 11

Handoff, validation and references 12

Revision history

VERSION	DATE	CHANGE
1.0	November 2026	Initial methodology draft.
1.1	November 2026	First public release: contents, framework crosswalk, assessment lifecycle and revision history.

Purpose and scope

An independent review of a connected business workflow.

The question SafeGuard answers

If an AI-related identity, input, tool or workflow behaves unexpectedly, what could the system access or do, and what would that mean for the business? SafeGuard examines the connected workflow rather than relying on the model or vendor product alone.

What is in scope

A defined integration, including triggers, people, service identities, AI services, applications, connectors, data stores, destinations, decisions, automated actions and recovery paths. The assessment applies to systems already operating, proposed workflows before go-live, and material changes to an existing integration.

Independent review

Kingfisher advises on behalf of the business. The provider remains responsible for its solution and technical changes. The business owner authorizes testing and retains responsibility for deployment decisions and residual risk. We do not assume a provider is deficient.

Proportionate depth

Decision-stage risk screening identifies questions and approval conditions. A separately scoped SafeGuard assessment examines evidence and may include selected authorized technical checks. Review depth is determined by the sensitivity of data, consequences of actions, access and evidence available.

Boundaries

This is not an enterprise-wide security audit, penetration test, legal opinion, certification, continuous monitoring or implementation engagement. Findings apply to the agreed workflow and point in time. No guarantee of security or compliance is implied.

Framework foundation

NIST guidance informs the assessment; it does not certify the result.

NIST AI RMF 1.0 (AI 100-1)

GOVERN establishes ownership and risk oversight. MAP identifies intended use, dependencies, data and consequences. MEASURE examines evidence and checks relevant behaviors. MANAGE prioritizes treatment, residual risk and ongoing review. Governance runs throughout the work.

NIST Cybersecurity Framework 2.0 (CSWP 29)

GOVERN and IDENTIFY inform accountability, asset knowledge and supplier oversight. PROTECT informs access, data and platform safeguards. DETECT, RESPOND and RECOVER inform telemetry, escalation, shutdown and recovery. These outcomes are applied proportionately to the workflow.

NIST Generative AI Profile (AI 600-1)

The profile adds generative-AI-specific risk considerations, including confabulation, information security, data privacy, and human-AI configuration. SafeGuard considers these risks when selecting relevant evidence and misuse scenarios.

How guidance becomes a review

Kingfisher translates relevant outcomes into six assessment domains. The internal register records scope, criteria, evidence and verification methods. Mappings are practitioner interpretations, not an official NIST crosswalk or an assertion that all NIST outcomes have been assessed.

Guidance under development

NIST AI RMF 1.0 is being updated. NIST IR 8596, the Cyber AI Profile, is an initial preliminary draft as of November 2026. Kingfisher tracks it as emerging guidance, not as a finalized requirement or an assurance claim.

Framework crosswalk

Primary NIST functions informing each SafeGuard assessment domain.

SAFEGUARD DOMAIN	AI RMF 1.0	CSF 2.0	GEN AI PROFILE FOCUS
Governance and inventory	GOVERN, MAP	GOVERN, IDENTIFY	Human oversight; supplier and component dependencies.
Identity and permissions	MEASURE, MANAGE	PROTECT	Information security; authorized access.
Data boundaries	GOVERN, MAP, MEASURE	GOVERN, IDENTIFY, PROTECT	Data privacy; information security.
Actions and human oversight	MAP, MEASURE, MANAGE	IDENTIFY, PROTECT	Human-AI configuration; confabulation; information security.
Platform and integration security	GOVERN, MEASURE	GOVERN, IDENTIFY, PROTECT	Information security; value chain and component integration.
Monitoring and recovery	MEASURE, MANAGE	IDENTIFY, PROTECT, DETECT, RESPOND, RECOVER	Information security; human-AI configuration.

How to read the crosswalk

The functions summarize the primary mappings in Kingfisher's internal control register. Generative AI Profile topics are illustrative considerations, not a control-by-control mapping. Governance applies throughout every engagement. This is a practitioner crosswalk, not an official NIST mapping or a claim of complete coverage.

Governance and access

The organization must know what the workflow is for and what it can access.

Governance and inventory

Establish the accountable business owner, intended purpose, authorized decisions and actions, connected systems, vendors and review triggers. Identify who operates the integration and who can approve changes.

Evidence and checks

Review ownership records, workflow and system inventories, data-flow diagrams, acceptable-use rules and supplier approvals. Trace the mapped dependencies against the deployed configuration and compare the actual workflow with the approved purpose.

Identity and permissions

Review dedicated integration identities, role assignments, API scopes and access to records or functions. Compare actual privileges with what the task requires. Establish that user and tenant authorization is enforced by the connected service, not by a prompt.

Evidence and checks

Examine redacted role exports, service account inventories, access matrices and credential storage or rotation arrangements. Where authorized, use test identities to confirm permitted access and denial of unrelated records or actions.

Business consequence

A workflow that can reach unrelated customer files or privileged functions may disclose or change information far beyond its intended purpose. A model refusing a request is not evidence that the underlying service would deny it.

Data and action boundaries

The application must constrain what information crosses boundaries and what AI can cause.

Data boundaries

Identify personal, financial, regulated or otherwise sensitive information. Determine what enters prompts, retrieval, payloads, logs and downstream destinations. Assess minimization, segregation, retention, sharing and model-improvement terms or settings.

Evidence and checks

Review data classification, request and response mappings, vendor terms, configuration and access policies. Use synthetic or redacted test records to observe whether unnecessary fields or unrelated client records can leave an approved boundary.

Actions and human oversight

Identify callable tools and permitted operations. Review meaningful approval before consequential actions, safe handling of model outputs, untrusted inputs, limits, retries, manual override and failure behavior. Prompts and retrieved content must not grant authority.

Evidence and checks

Review tool inventories, permissions, approval rules and run records. Where permitted, use non-destructive scenarios to test denial, approval bypass resistance, prompt injection, unintended changes, replay and safe fallback. Verify outputs before they trigger a consequential action.

Business consequence

A client document can contain instructions that manipulate a model. The key question is whether those instructions can disclose records, trigger a call or send a message without valid authorization.

Platform and operations

Supporting systems must protect the workflow and make failures manageable.

Platform and integration security

Review how secrets are stored, APIs and webhooks authenticate and validate requests, replay is handled, exposed components and relevant dependencies are configured, and responsibility is divided between the provider and the business.

Evidence and checks

Examine relevant configuration exports, architecture, secret-management settings, dependency inventories and shared-responsibility records. Review authorized API or webhook checks without extracting real data or exposing credentials.

Monitoring and recovery

Determine whether meaningful actions and access decisions can be attributed, failures or permission changes detected, and the workflow disabled promptly. Review escalation, access revocation, rollback and reassessment procedures.

Evidence and checks

Inspect sanitized audit records, alerts, retention rules and response contacts. Demonstrate representative events or a safe notification check where authorized. Review emergency-stop and revocation procedures and walk through a recovery scenario with responsible owners.

Business consequence

An integration may fail silently, send duplicate messages or continue operating after an access change. The business needs a reliable way to recognize the problem, stop the workflow and recover safely.

Assessment and evidence

A defensible conclusion requires evidence, not assurances alone.

Authorize and scope

Agree the sponsor, operator, systems, environments, test identities, permitted operations, time window, exclusions, data handling, abort conditions and escalation contacts. Obtain written authorization and any necessary provider consent before technical testing.

Map the workflow

Record triggers, identities, data classes, trust boundaries, service dependencies, approved actions and expected behavior. Identify the consequences of an unauthorized disclosure, change, decision or outage.

Examine and verify

Compare the approved workflow with actual configuration. Request redacted records and verify selected boundaries with authorized identities or synthetic data. Record what was observed directly, what the provider reported and what was not tested. Do not treat a policy or prompt as proof an enforcement mechanism works.

Keep evidence traceable

Record the source, system, date, reviewer, test conditions, expected and actual outcomes, relevant limitations and evidence references. Use only necessary data, avoid collecting secrets, and follow agreed retention and deletion arrangements.

Handle limits explicitly

A missing artifact, inaccessible vendor setting or prohibited test is not a verified result. Record the gap and its decision relevance. Avoid destructive tests, production data extraction or unnecessary exposure. Refer specialized or regulated questions for separate qualified review as needed.

Assessment lifecycle

An evidence-led process with management accountability throughout.

GOVERNANCE THROUGHOUT: accountable ownership and authorization

Authorize and scope

Written authorization, defined workflow, evidence and testing boundaries.

Map the workflow

Identify identities, connected systems, data flows and permitted actions.

Examine and verify

Compare expected behavior with available evidence and authorized checks.

Classify risk and decide

Prioritize business consequences, uncertainty and decision conditions.

Hand off and reassess

Assign owners, verify fixes and revisit material changes or incidents.

REASSESS ON MATERIAL CHANGE

The designated provider remains responsible for remediation. Management retains deployment authorization and risk acceptance; SafeGuard documents the evidence supporting those decisions.

Risk classification and decisions

Prioritize consequences, not a composite score.

Control outcomes

Verified: evidence supports the scoped criterion. Needs remediation: a gap is observed. Not verified: available evidence is insufficient. Not applicable: exclusion is justified.
Accepted risk: an authorized owner documents a residual-risk decision; it is not a passing control.

High - action before use or expansion

A material exposure or unauthorized action with serious business consequences, or a missing safeguard essential to safe operation. Recommend holding the affected activity until corrected and verified. Urgent or critical conditions are flagged explicitly.

Medium - planned corrective action

A material weakness with more limited exposure or consequence, or a gap in oversight or resilience. Address before relevant expansion or by an agreed date based on business impact.

Low - routine improvement

A limited exposure or improvement opportunity. Record accountable ownership and schedule proportionate follow-up. Severity considers business impact, exploitability and exposure, with uncertainty stated separately.

Decision record

Recommend proceed, revise or hold the affected workflow based on evidence and material unresolved risks. Do not average individual findings into a SafeGuard score. Management retains the decision and documents accepted residual risks, conditions and review dates.

Handoff, validation and references

The assessment ends with decisions and accountable follow-up.

What the business receives

Documented scope and data flows; a record of controls examined and evidence limitations; findings with business consequences; prioritized actions and proposed owners; and a leadership walkthrough with the provider or technical team as appropriate.

How fixes are verified

For each issue, identify the observed condition, affected boundary, accountable owner, recommended action and evidence or retest needed to show it has been fixed. Record the new date, result and remaining exceptions. Implementation is the responsibility of the designated provider.

When to revisit

Reassess after a material connector, permission, model, vendor, data-class or high-consequence action change; a new public entry point; or a relevant incident. Review the methodology annually and when material guidance or threat patterns change.

Reference editions

NIST AI RMF 1.0 (2023): https://doi.org/10.6028/NIST.AI.100-1
NIST CSF 2.0 (2024): https://doi.org/10.6028/NIST.CSWP.29
NIST Generative AI Profile (2024): https://doi.org/10.6028/NIST.AI.600-1
NIST Cyber AI Profile (draft): https://csrc.nist.gov/pubs/ir/8596/iprd

Document control

Public methodology v1.1 | November 2026 | Kingfisher Advisory LLC. Detailed control criteria, test procedures and internal mappings remain separate. This is a methodology, not a client assessment or NIST certification.