Compliance Assessment for Agentic Processes: Evidence-Governed Promotion, Control Mapping, and Bounded Compliance Claims

Arslan Brömme
CISSP, CISM, CISA, CAISE
Independent Researcher
arslan.broemme@aviomatik.de

Draft v0.9.0.14 – 20 September 2026

Preprint / working paper. This version is a work in progress and may be updated. Comments are welcome.
DOI: 10.5281/zenodo.22858862

Abstract

Agentic processes increasingly combine AI agents, tools, persistent state, delegated authority, external systems, and human oversight. Their operation can generate large volumes of heterogeneous telemetry, security findings, audit records, testing results, and incident evidence, but neither volume nor technical success alone establishes compliance beyond the specific policy, control, and assessment scope actually evaluated. This paper develops and specifies an evidence-governed assessment model for agentic processes. It addresses three research questions concerning claim-preserving control mapping, bounded compliance assessment, and selective promotion of high-volume observations into active governance workflows. First, the model defines a claim-preserving specialization of established evidence-to-assessment practice for heterogeneous agentic evidence. Second, it defines an evidence-governed promotion chain that separates evidence preservation from workflow activation so that retained telemetry does not automatically become an active GRC object. Third, it places these semantics in an interoperable plan–do–check–act lifecycle that connects controlled testing, runtime monitoring, intrusion response, investigation-derived evidence, remediation, retesting, reassessment, and existing enterprise GRC platforms through a platform-neutral semantic interface. The model uses five typed control-assessment states: Supported, Contradicted, Insufficient evidence, Unresolved, and Not applicable. It also specifies bounded exchange records and a worked approval-before-payment example. The central claim boundary is that security and assurance evidence may support and, where an authorized assessment method defines a closed technical control objective and the required observation conditions are satisfied, may justify a bounded Supported or Contradicted control assessment. Such a result does not by itself establish broader legal, regulatory, or organizational compliance or non-compliance beyond the assessed scope. The model is presented as a research design for subsequent controlled proof of concept evaluation. It does not provide legal advice, prescribe a proprietary GRC implementation, replace mandatory logging or retention obligations, or claim that technical conformance to selected controls establishes legal compliance.

1 Introduction

Agentic systems can execute tools, retain state, delegate tasks and authority, consume external information, and coordinate across organizational or technical boundaries. A single business action may therefore be distributed across planning, retrieval, approval, execution, monitoring, recovery, and human oversight. Compliance assessment in such settings is not only a question of whether a static control exists. It also concerns whether a requirement was applicable at the relevant time, whether the corresponding control was designed and operated as intended, which evidence supports that conclusion, and which limitations remain.

The motivating problem is a recurring claim-boundary error. Runtime monitoring, penetration testing, access-control logs, audit records, or successful recovery can each provide useful evidence. None of these artifacts, in isolation, establishes broader legal compliance. At the same time, sufficiently qualified technical evidence can support a strong bounded assessment of a narrowly specified technical policy or control when an authorized assessment method and its required observation conditions are satisfied. Conversely, missing or inconclusive technical evidence does not automatically establish legal non-compliance. A compliance model for agentic processes should therefore preserve the scope, provenance, time, assumptions, and limitations of its evidence as it moves from technical mechanisms toward governance and assurance decisions.

This problem is operationally relevant because current regulatory and assurance frameworks already require combinations of risk management, logging, monitoring, testing, incident handling, management oversight, and continuous improvement. The EU AI Act treats risk management for high-risk AI systems as a continuous iterative lifecycle process and requires testing before placing such systems on the market or putting them into service. It also requires logging capabilities for high-risk systems and post-market monitoring that supports evaluation of continuous compliance [6]. GDPR Article 32 requires risk-appropriate technical and organizational security measures and includes regular testing, assessment, and evaluation of their effectiveness [7]. NIS2 combines management-body responsibility, cybersecurity risk-management measures, effectiveness assessment, and incident reporting [8]. DORA combines ICT governance and risk management with incident handling, resilience testing, penetration testing, remediation, and advanced threat-led testing for in-scope financial entities [9]. MiCA provides a sector-specific example for crypto-asset service providers through organizational, ICT resilience, risk-assessment, and record-keeping obligations that interact with DORA [10]. These frameworks differ in scope and legal effect, but they share a need for evidence that can be related to controls without overstating what the evidence proves.

This paper addresses three research questions. *Research Question 1 (RQ1) asks how heterogeneous evidence from agentic processes and their security and assurance mechanisms can be mapped to applicable requirements and controls while preserving evidence scope, provenance, time, assumptions, and*

limitations. Research Question 2 (RQ2) asks how mapped evidence can support bounded compliance assessments without treating technical evidence, control observations, or absence of findings as proof of legal compliance or non-compliance. Research Question 3 (RQ3) asks how high-volume observations can be selectively promoted into control, risk, incident, or compliance workflows without conflating telemetry volume, evidentiary relevance, and active governance workflow priority, while preserving evidence and uncertainty for later reassessment.

This paper makes three core contributions. First, it defines a *claim-preserving specialization of evidence-to-assessment practice for heterogeneous agentic evidence*. The contribution is not the existence of assessment records themselves. It is the preservation of source-specific evidentiary properties, assumptions, limitations, scope, time, and assessment authority as technical and organizational records are related to bounded control and compliance claims. Second, it defines *evidence-governed promotion*. This separates evidence preservation from workflow activation and makes workflow eligibility a governed decision rather than an evidentiary strength score. Third, it defines an *interoperable lifecycle and GRC integration boundary* in which existing security tools, third-party or agentic-process-specific evidence producers, bounded investigation-derived evidence, controlled assessment, risk registers, remediation, verification, and existing GRC platforms can exchange structured assessment results without amplifying evidentiary strength. The conceptual records, PDCA placement, deployment profiles, and investigation-feedback interface are architectural elements of these three contributions rather than independent novelty claims.

The two central chains are summarized below. The first describes the evidence-to-assessment path underlying RQ1 and RQ2, while the second summarizes the evidence-governed promotion path underlying RQ3.

Evidence-to-Assessment Chain
Applicability Context → Requirement / Control → Evidence Requirement → Evidence Producer → Evidence Qualification → Control Assessment → Bounded Compliance Claim

Evidence-Governed Promotion Chain
Raw Telemetry → Contextualized Observation → Candidate Evidence → Qualified Evidence → Bounded Claim / Finding → Control or Risk Relevance → Governance Workflow Candidate

Bounded assessment outputs can then be exchanged with existing GRC platforms for control libraries, risk registers, issues, audit workflows, evidence repositories, remediation tracking, and management reporting. That integration is an operational destination for the assessment record, not an additional source of evidentiary authority.

2 Background and Related Work

The literature and state-of-practice scan for this draft includes sources available up to the date of this paper. It is a targeted positioning review rather than a systematic literature review. Primary regulatory, standards, publisher, professional-body, vendor, and original research sources are preferred where available. Commercial and professional-practice sources are used only to characterize current implementation patterns, not as independent scientific or legal validation.

2.1 Control Assessment, Management Systems, and Machine-Readable Assurance

The general problem of assessing controls is well established. NIST SP 800-53A provides procedures for assessing security and privacy controls within a risk management framework and explicitly supports assessments at different phases of the system development lifecycle [15]. NIST OSCAL extends this direction with machine-readable assessment plans and assessment results. Its assessment-results model represents reviewed controls, assessment subjects, observations, findings, risks, evidence, and attestations and supports point-in-time assessment as well as continuous monitoring [16]. OSCAL therefore already provides a rich interchange and assessment structure. The present model does not replace that structure. Its additional focus is the semantics needed when heterogeneous agentic-process evidence enters such structures: claim-specific qualification, preservation of source and property limitations, governed workflow promotion, and bounded claim authority without evidentiary amplification.

NIST AI RMF organizes AI risk-management activity around Govern, Map, Measure, and Manage and treats governance as cross-cutting while emphasizing lifecycle risk management [13, 14]. ISO/IEC 27001 specifies requirements for an information security management system, while ISO/IEC 42001 provides an AI management-system framework [11, 12]. ISO/IEC 23894 adds guidance for integrating AI risk management into organizational activities, ISO/IEC 42005 addresses AI system impact assessment across the lifecycle, and ISO/IEC 42006 addresses bodies that audit and certify AI management systems [18, 19, 20]. These sources are relevant because they separate management, risk, impact assessment, and assurance functions rather than reducing them to one technical compliance signal.

Professional assurance frameworks point in the same direction. COSO's 2026 GenAI guidance places AI risks and controls within an established internal-control framework, ISACA's 2026 lifecycle toolkit emphasizes continuous governance artifacts and oversight, and the IIA AI Auditing Framework addresses governance, management, risk, and internal-audit assessment of AI [35, 36, 37]. The present model does not replace these approaches. It focuses on the evidence semantics and promotion boundary between technical or operational records and downstream control, risk, and compliance assessment.

2.2 Runtime Compliance, Continuous Assurance, and Agentic Governance

Recent work increasingly operationalizes selected governance or regulatory requirements. C-Trace expresses a subset of GDPR requirements as policy predicates over agent execution traces and uses runtime monitoring to reject actions that violate encoded predicates [17]. Audit-as-Code maps governance requirements to versioned policy specifications, machine-checkable evidence bundles, deterministic CI/CD gates, and remediation guidance [24]. These approaches demonstrate that selected requirements can be operationalized and continuously evaluated. The present work addresses a different downstream question: how results from such mechanisms, together with security findings, testing evidence, organizational records, and investigation-derived evidence, can support bounded control and compliance assessment without converting a machine-testable predicate or gate outcome into a universal legal conclusion.

Research on AI auditability similarly emphasizes the need for lifecycle-wide auditability measures and multidisciplinary assurance capability [25]. Santelmann's 2026 review of agentic-AI governance links traceability, auditability, and accountability as an evidence-and-responsibility chain and identifies recurring gaps in audit-grade evidence and responsibility allocation in multi-agent settings [26]. Nannini et al. map AI agents across interacting EU legal regimes and propose a compliance architecture and regulatory-trigger mapping [27]. Kasirzadeh and Gabriel characterize agents through autonomy, efficacy, goal complexity, and generality to support differentiated governance profiles [28]. These works reinforce the need for agent-specific governance while also showing why this paper should not claim a generic first compliance architecture

for agents.

A recent neighboring proposal, AGIL, introduces an adaptive governance layer for discovery, behavioral risk classification, runtime policy enforcement, continuous attestation, and policy adaptation [31]. Runtime Governance for AI Agents formalizes path-dependent compliance policies over agent identity, partial execution path, proposed action, and organizational state [30]. These approaches are closest to runtime governance and enforcement. The present model instead focuses on the downstream semantics of evidence after it has been produced: which bounded claims are supported, how they relate to controls and risks, when they should enter active governance workflow, and what limitations survive downstream integration.

TAIP addresses a closely neighboring assurance-scaling problem. It normalizes heterogeneous evidence into reusable AI Assurance Objects and recursively aggregates evidence into posture across claim, system, organizational, and ecosystem levels [29]. The present model is complementary but narrower in a different direction: it focuses on evidentiary meaning, bounded claim authority, and the promotion boundary that determines whether a qualified claim should enter an active governance workflow. In the approval-before-payment example below, an assurance object can retain or contribute approval evidence, while the present model separately asks whether that evidence supports the exact transaction amount and scope, which control-assessment state follows under the declared method, and whether the resulting bounded claim is eligible for workflow activation. The distinction is therefore between posture aggregation and claim-specific qualification, assessment, and workflow eligibility. The present model does not attempt to compute an overall trustworthiness posture.

AcquireBound addresses a narrower upstream authorization problem that arises when an agent acquires compute, credentials, accounts, services, or other resources and the acquired output could introduce new authority into a task [32]. Its activation model separates acquisition or transaction success from authorization to make the resulting resource usable. For the present paper, such activation and enforcement records can become evidence inputs, but successful acquisition, payment, or delivery does not itself establish authorization to activate the resulting capability. A recent structured review of authorization architectures for tool-using AI agents similarly separates identity and credential lifecycle, delegation, runtime enforcement, prompt-injection bypass, and auditability or provenance [33]. These distinctions support the separation used here between identity evidence, authorization context, observed execution, and downstream compliance assessment.

2.3 Evidence Triage, Alert Prioritization, and Logging

In this subsection, *triage* is used in the general sense of preliminary sorting, prioritization, and routing of alerts, observations, or evidence for further handling. It is distinct from *incident triage*, which evaluates and prioritizes suspected or confirmed incidents for investigation and response. Evidence-governed promotion goes beyond such preliminary triage by requiring claim-specific evidence qualification and governed relevance assessment before a bounded claim enters an active governance workflow.

High-volume security telemetry and alert fatigue are established operational problems. NIST SP 800-92 already distinguishes filtering and prioritization of log analysis from alteration of the original logs and recommends organizational prioritization and correlation to focus analyst attention [22]. More recent work on risk-based alerting reformulates alert handling as a continuous prioritization problem rather than a binary threshold decision and shows that such prioritization can materially reduce analyst review burden [34]. In the financial sector, the DORA regulatory technical standards require entities to identify events to be logged, define retention, and align log detail with purpose and use, with retention informed by business, information-security, and ICT-risk considerations [23].

The contribution here is therefore not generic log filtering, SIEM correlation, or risk-based alerting. It is evidence-governed promotion: a claim-preserving transition from observation to active governance workflow. The model asks whether available records can support a bounded claim, whether that claim is sufficiently supported for a defined assessment purpose, whether it is relevant to a control, risk, incident, or compliance obligation, and only then whether it should be promoted into an active governance workflow. A conforming promotion process must preserve the evidence references, policy version, decision context, and limitations associated with hold or defer outcomes. It must also permit later policy, applicability, or evidence changes to trigger re-evaluation without requiring the original evidence to be reconstructed from discarded telemetry. Preservation and workflow activation remain separate decisions.

2.4 Enterprise GRC and AI-Governance Practice

Current enterprise practice confirms that AI governance is increasingly integrated into existing GRC and control environments rather than implemented as an isolated parallel system. EY describes an integrated governance approach combining information-security and AI-management structures around ISO/IEC 27001, ISO/IEC 42001, and the EU AI Act [38]. KPMG describes AI governance and compliance implemented through existing platforms such as ServiceNow [39]. PwC's 2026 Regulation & Controls Lifecycle uses agentic AI for process mapping and for tracing regulatory obligations to policies, processes, and controls [40]. Deloitte describes lifecycle controls, risk appetite, monitoring, remediation, and audit-ready evidence and has also introduced AI-assisted control-testing and evidence-analysis capabilities [41, 42].

The platform layer shows a similar pattern. ServiceNow combines AI Control Tower with AI Risk and Compliance across an AI governance lifecycle [43]. IBM documents direct integration between watsonx.governance and OpenPages, while watsonx.governance also exposes continuous policy, obligation, and compliance-evidence functions [44, 45]. OneTrust provides AI inventory, risk scoring, approvals, runtime controls, policy evaluation, and audit-ready evidence within an AI-governance platform [46]. These examples are not treated as validation of the proposed model. They demonstrate that the destination objects assumed here, including inventories, controls, risks, evidence, findings, issues, remediation, and assessment results, already exist in major enterprise environments. The proposed contribution is intended to be semantically mapped into such environments rather than to replace them.

2.5 Prior Evidence and Security Layers

This work continues an evidence-oriented research line for agentic processes. The black-box architecture preserves selected records through capture, canonicalization, hashing, anchoring, certification, and verification [1]. The subsequent evidence model separates evidentiary properties from mechanisms, assumptions, limitations, and threat scope [2]. Those papers motivate the terminology used here, but the compliance architecture does not require their implementations. Evidence-governed promotion consumes bounded evidence claims from any producer that satisfies the interface contract below and uses retained properties, assumptions, and limitations to assess suitability for a downstream control, risk, or compliance purpose.

The A-IDS model produces bounded runtime findings while separating visibility from matching and unresolved evidence from confirmed deviation [3]. The A-IPS model separates

evidence sufficiency, admissibility, authorization, enforcement, effect verification, and recovery validity [4]. Penetration testing adds controlled adversarial evidence and retesting while maintaining that a demonstrated weakness remains bounded by the tested conditions [5]. The compliance layer does not treat these prior components as independent validation of one another. It treats them as conceptually distinct evidence producers whose outputs must retain their original limits. Conventional security infrastructure can enter the same model as complementary producers, while GRC platforms can provide governed assessment context and serve as integration and workflow environments.

The Black Box, Evidence Model, A-IDS, A-IPS, and agentic-process penetration testing are architecturally additive components. They complement existing logging, monitoring, security, testing, enforcement, audit, and GRC capabilities rather than replacing them. The Black Box supports preservation and provenance. The Evidence Model supports bounded interpretation and qualification. A-IDS supports bounded detection and visibility assessment. A-IPS supports governed intervention and authorization. Penetration testing supports controlled adversarial assessment. Their outputs can enter existing governance and compliance processes, but each remains subject to claim-specific qualification and does not, by itself, establish compliance beyond the scope and evidentiary property its output actually supports.

In the reference architecture, A-IDS and A-IPS denote the detection and governed-intervention roles defined in this research series. Functionally comparable third-party mechanisms may instantiate these roles where they provide equivalent bounded outputs and satisfy the same evidence, decision-boundary, and interface requirements. The acronyms therefore remain the reference labels used in this paper, while the compliance architecture remains open to interoperable mechanisms that fulfill the corresponding functions.

Machine identity and authority governance remain separate control functions. They determine which identities, credentials, delegated permissions, and execution rights are valid within an agentic process [33]. The evidence architecture does not replace those controls. It preserves and qualifies evidence about their operation so that downstream assessments can distinguish identity, authorization, observed execution, and evidentiary support. Authenticated identity evidence does not by itself establish authorization. Authorization evidence can establish an authorization-specific control state when the governing policy and assessment scope make that property dispositive, but it does not by itself establish broader policy conformance or compliance beyond that scope. Acquired capabilities introduce the same boundary because possession or successful acquisition of a resource does not itself establish authority to activate or use it [32].

At the interface level, the compliance layer assumes a generic producer contract. An evidence producer must expose the record or bounded finding being asserted, the subject and covered period, the evidentiary property or observation status, source identity and provenance information available for that property, relevant scope and observation path, and material assumptions and limitations. Where the producer makes a derived finding, it should also expose the source references and decision or rule version needed to interpret that finding. A compact interface can be represented as

$$\begin{aligned}\mathsf{EvidenceProducerOutput} = (&\mathsf{record/finding}, \mathsf{subject},\\ &\mathsf{period}, \mathsf{property/status}, \mathsf{source},\\ &\mathsf{provenance}, \mathsf{scope},\\ &\mathsf{observation\ path}, \mathsf{assumptions},\\ &\mathsf{limitations}, \mathsf{rule\ refs}).\end{aligned}$$

The compliance model does not require a specific detector, anchoring mechanism, response engine, penetration-test framework, or implementation from the earlier series. Conventional logging, third-party detectors, external assurance records, human decisions, or other evidence producers can satisfy the same contract. The earlier series is therefore an example source ecosystem rather than a prerequisite for the present model.

3 Compliance Context and Claim Boundary

The paper uses *compliance assessment* as a bounded evaluation of evidence against identified requirements, controls, policies, or assurance objectives. It does not treat compliance as a property inferred directly from one detector, one penetration test, one audit record, or one technical control.

3.1 Applicability Before Assessment

The model begins with applicability rather than with evidence. A requirement may apply only to a particular legal role, system category, sector, processing activity, jurisdiction, contractual relationship, risk class, or point in time. The EU AI Act, GDPR, NIS2, and DORA illustrate this dependence on role and scope in different ways [6, 7, 8, 9]. Technical or operational evidence about system behavior does not by itself establish that a legal or regulatory requirement applies. Applicability can instead be supported by governed legal, contractual, organizational, classification, or other authoritative context and should be recorded as an assessment input with its own source, decision authority, effective period, assumptions, and limitations.

A simplified applicability record can be represented as:

$$\begin{aligned}\mathsf{ApplicabilityContext} = (&\mathsf{source}, \mathsf{decision\ authority}, \mathsf{role},\\ &\mathsf{jurisdiction}, \mathsf{scope}, \mathsf{effective\ period},\\ &\mathsf{basis}, \mathsf{assumptions}, \mathsf{limitations}).\end{aligned}$$

This is a conceptual record schema, not an implementation serialization or automated legal-reasoning system. Applicability is an authoritative governance, legal, contractual, or organizational input to the technical assessment path. A technical assessment or promotion layer may consume an applicability decision, but it does not create, validate, or replace the legal, contractual, or organizational interpretation underlying that decision.

3.2 Claim Boundaries

Three distinctions are maintained throughout the model:

Insufficient evidence $\neq$ non-compliance.
No detected deviation $\neq$ compliance.
Control satisfaction $\neq$ legal compliance.

The first distinction prevents missing visibility from being converted into an adverse legal conclusion. The second prevents negative monitoring or test results from being interpreted as proof of security or compliance. The third preserves the separation between technical or organizational control assessment and the wider legal, regulatory, contractual, or certification decision.

These distinctions do not imply that technical evidence is incapable of justifying a strong bounded control assessment. In this paper, a technical control objective is treated as *closed for the assessment* only when an authorized assessment method identifies the bounded subject, control or policy version, assessment period, observable predicate, required observation paths and coverage, exception handling, and decision rule without requiring resolution of an open legal qualifier such as whether a measure is appropriate or proportionate. For example, if a governed technical policy states that an agent must not initiate Internet access and these assessment conditions are established, qualified evidence may justify a *Supported* or *Contradicted* assessment for that specific policy. The claim

boundary concerns widening that result into broader compliance conclusions without the additional applicability context, scope, authority, and evidence required for those conclusions.

4 Evidence-to-Control Model

4.1 Requirement and Evidence Requirement

A compliance source may contain obligations, principles, controls, outcomes, or management-system requirements. The model therefore normalizes a selected item into a *requirement or control objective* without claiming that every legal provision is reducible to one control.

A conceptual requirement record is:

$$\mathsf{Requirement} = (\mathsf{id}, \mathsf{source}, \mathsf{text\ reference}, \mathsf{applicability}, \mathsf{control\ objective}, \mathsf{version}).$$

The corresponding *EvidenceRequirement* states what evidence would be relevant to assessing the selected control objective. It can identify expected event classes, control-operation records, approvals, test results, monitoring periods, sampling rules, or management decisions. Evidence requirements should also declare when absence of evidence is meaningful and when it merely reflects an observation limit.

4.2 Evidence Binding and Qualification

Evidence from heterogeneous producers should not be flattened into a single unqualified "proof" field. The compliance layer should retain at least the producer, evidence type, subject, covered period, source identity, provenance or integrity properties, observation path, relevant assumptions, and limitations.

We use the following conceptual schema:

$$\mathsf{EvidenceBinding} = (\mathsf{producer}, \mathsf{subject}, \mathsf{property}, \mathsf{period}, \mathsf{source}, \mathsf{assumptions}, \mathsf{limitations}, \mathsf{reference}).$$

Qualification evaluates whether an evidence item is suitable for the specific evidence requirement. Qualification is claim- and requirement-relative rather than a global trust score. A qualification decision should therefore bind the evidence property being relied upon, the declared requirement, the observation path and coverage, the relevant source and provenance properties, and material assumptions and limitations. The same record may be strong evidence for temporal existence and weak evidence for capture completeness, semantic validity, or authorization. This preserves the evidence-property distinction introduced in the earlier evidence model [2].

4.3 Control Assessment

A control assessment binds a control objective to the qualified evidence actually reviewed under a governed assessment method. The method should identify the assessed scope, period, observation and sampling conditions, decision rules, exception handling, required authority, and the evidence conditions needed for the requested assessment. Qualified evidence therefore does not naturally yield a control status by itself. Conceptually, the assessment step combines qualified evidence, the control objective, an authorized assessment method, and applicability context. This structure is compatible with the general assessment logic used by NIST SP 800-53A and OSCAL while remaining focused on the additional claim-boundary problem created by heterogeneous agentic-process evidence [15, 16]. Table 1 summarizes the conceptual records and the claim boundary preserved by each record.

Sections 3–4 provide the design-synthesis answer to RQ1: heterogeneous records are not mapped directly to compliance. They pass through governed applicability, explicit evidence requirements, source-specific evidence binding and qualification, and a control assessment that retains scope, provenance, time, assumptions, and limitations.

5 Assessment States and Bounded Claims

Binary states are often too coarse for evidence-driven assessment. The model therefore distinguishes at least five assessment states.

- **Supported**: the qualified evidence supports the evaluated control objective for the declared scope and period.
- **Contradicted**: qualified evidence demonstrates a relevant deviation from the evaluated control objective within the assessed scope.
- **Insufficient evidence**: evidence required for the requested control assessment is absent, lacks the governed minimum coverage, or is otherwise inadequate to reach the requested assessment.
- **Unresolved**: relevant evidence is present, but it is conflicting, ambiguous, or dependent on assumptions or relationships that remain open, so the assessment cannot yet be resolved.
- **Not applicable**: a documented applicability decision states that the requirement or control is outside the declared assessment context.

These states describe the evidentiary status of a control assessment. They are not legal verdicts. For example, a contradicted technical control may be highly relevant to a compliance decision but may still require an applicability analysis, compensating controls, materiality analysis, temporal context, and an authorized human or institutional determination. Similarly, supported controls do not establish that all applicable requirements were identified or that the evidence is complete across the organization.

The central assessment operation can be represented schematically as

$$\mathrm{assess}_{P_v}(E, c, m, a) \in \{\mathrm{Supported}, \mathrm{Contradicted}, \mathrm{Insufficient\ evidence}, \mathrm{Unresolved}, \mathrm{Not\ applicable}\}.$$

where E is the qualified evidence set, c is the control objective, m is the governed assessment method, a is the authoritative applicability context, and P_v is the applicable versioned assessment policy. The method m supplies the decision rule that relates evidence to the control objective. The function is policy- and method-relative and does not compute an objective truth about the system or its legal compliance. An authoritative not-applicable decision in a yields *Not applicable*. Missing or inadequate required evidence can yield *Insufficient evidence*. Conflicting or materially ambiguous evidence can yield *Unresolved*. A *Supported* or *Contradicted* result requires sufficient evidence and a method that determines whether the bounded control predicate is satisfied or deviated from within the declared scope. The notation is conceptual and does not eliminate assessor judgment where the governing method requires it.

The status types are deliberately distinct:

$$\mathsf{SufficiencyStatus} \neq \mathsf{ControlAssessmentStatus},$$
$$\mathsf{ControlAssessmentStatus} \neq \mathsf{PromotionDecision}.$$

A sufficiency decision asks whether evidence is adequate for a declared claim or purpose. A control-assessment state applies the governed assessment method to a control objective. A promotion decision determines workflow eligibility. None of these types is interchangeable with a risk rating or legal conclusion.

A bounded compliance claim can therefore be represented

Conceptual record	Core fields	Primary boundary
ApplicabilityContext	source, decision authority, role, jurisdiction, scope, effective period, basis, assumptions, limitations	Evidence does not determine legal applicability
Requirement	identifier, source, text reference, applicability link, control objective, version	Control abstraction does not replace source text
EvidenceRequirement	expected evidence, observation period, sufficiency conditions, absence semantics	Requested evidence does not guarantee available visibility
EvidenceBinding	producer, subject, property, period, source, assumptions, limitations, reference	Evidence supports only bounded properties
PromotionDecision	evidence and claim references, relevance basis, promotion policy version, status, reassessment trigger, limitations	Promotion creates workflow eligibility, not evidentiary truth, risk treatment, authorization, or a compliance conclusion
IncidentEvidenceClaim	claim, scope/resolution, support, source relations, causal status, sufficiency, assumptions, limitations, time	Incident-derived claims are not incident ground truth, and claim granularity cannot exceed evidentiary support
ControlAssessment	control, assessed scope, qualified evidence, method, authority, status, unresolved issues, limitations	Control result is not organization-wide compliance
AssessmentExchange	source assessment, control and requirement references, evidence references, status, scope, period, assumptions, limitations, authority, destination reference	GRC ingestion does not strengthen the underlying evidence or widen the claim
ComplianceClaim	applicability decision reference, assessed entity, control and version, status, evidence basis, assessment method, authority, period, exceptions, reassessment trigger, residual limits	Bounded assessment record is not a legal or regulatory verdict

Table 1. Conceptual records for evidence-governed compliance assessment. The schemas separate applicability, requirements, evidence, promotion, assessment, and claim boundaries.

as:

$$\begin{aligned}\mathsf{ComplianceClaim} = (&\mathsf{applicability\ ref}, \mathsf{entity},\\ &\mathsf{controls/versions}, \mathsf{status}, \mathsf{evidence\ basis},\\ &\mathsf{method}, \mathsf{authority}, \mathsf{period}, \mathsf{exceptions},\\ &\mathsf{reassessment\ trigger}, \mathsf{limitations}).\end{aligned}$$

In this paper, *ComplianceClaim* denotes a bounded assessment record, not a legal or regulatory verdict. The *authority* field is deliberate. A technical component may produce a policy-assessment record, but the authority to make a formal organizational, certification, supervisory, or legal determination may remain with a human role, auditor, conformity-assessment body, regulator, or other designated institution. Exceptions can include documented exclusions or compensating controls where the governed assessment method permits them.

5.1 Incident-Derived Evidence, Claim Granularity, and Sufficiency

Evidence-based investigation can become an additional evidence producer, but its output should not be represented as one undifferentiated incident truth record. It may include incident analysis and reconstruction and combine preserved agentic-process records, A-IDS findings, A-IPS intervention and recovery records, conventional security telemetry, penetration-test and retest evidence, human observations, and external-system records. Each source can support different properties and can share common-mode dependencies with other sources. Correlation and aggregation therefore do not by themselves amplify evidentiary strength.

The relevant question is claim-specific: what can the currently available evidence justify for a particular statement, at a particular level of detail? A coarse claim such as "an authenticated call reached endpoint X during the incident window" may be sufficiently supported while a stronger claim about which agent caused the call, why it acted, or whether another agent compromised it remains unresolved. The model therefore applies a *claim granularity boundary*: the scope, resolution, and causal specificity of an incident-derived claim should not exceed what the underlying evidence can support.

For a qualified evidence set E, candidate claim c, and governed, versioned assessment or evidence policy P_v, claim-specific sufficiency can be represented as

$$\mathrm{sufficient}_{P_v}(E, c) \in \{\mathrm{sufficient}, \mathrm{insufficient}, \mathrm{indeterminate}\}.$$

Here, *indeterminate* is reserved for claim-specific sufficiency when available evidence cannot be evaluated decisively under the current policy because material evidence relationships or assumptions remain open. It is distinct from the *Unresolved* control-assessment state defined above. This status is not the probability that an incident hypothesis is true. It is a policy-relative judgment about whether the available evidence is adequate for the declared claim and purpose. Relevant dimensions can include support coverage, source separation or independence, temporal resolution, identity binding, causal support, reproduction or corroboration support, visibility, scope, and freshness. These dimensions form an evidentiary adequacy profile rather than one universal confidence score.

A conceptual IncidentEvidenceClaim therefore retains the claim, its scope or resolution, supporting evidence references, source relationships, causal status, sufficiency status, assumptions, limitations, and assessment time. These fields form the minimum interface contract required by the compliance layer. A broader evidence-based investigation method, including incident analysis and reconstruction, may add richer hypothesis, propagation, or provenance structures without changing this boundary. Re-evaluation can strengthen, weaken, refine, or supersede an earlier incident claim when new observations, forensic artifacts, controlled replays, penetration tests, or external telemetry become available. Historical claims and the evidence basis on which they were issued should remain traceable rather than being overwritten.

For integration with an existing GRC platform, the model uses a conceptual *AssessmentExchange* record rather than prescribing an API or serialization format. It carries the control and requirement references, source evidence references, assessment status, scope, covered period, assumptions, limitations, assessment authority, and a destination-side reference. The record is intentionally semantically conservative. Moving a bounded finding into a GRC workflow does not make the finding stronger, more complete, or legally dispositive.

Section 5 provides the design-synthesis answer to RQ2: mapped evidence can support explicit evidentiary assessment states and bounded compliance claims, while assessment authority, unresolved uncertainty, and legal applicability remain separate from technical evidence and control observations.

6 Evidence-Governed Promotion and Scalable Governance Workflows

In this paper, *telemetry* denotes technically generated runtime and observation data, including logs, events, tool-call records, network observations, and identity or access telemetry. The assessment model is broader than telemetry. Approvals, audit records, management decisions, test artifacts, and external assurance records can enter the same evidence-qualification process as other source records. The promotion chain is written from the high-volume runtime telemetry path addressed by RQ3, while non-telemetry source records can enter the

qualification and assessment process at the appropriate stage.

Agentic processes and their surrounding infrastructure can generate far more telemetry than can or should become individual GRC objects. The model therefore distinguishes three quantities that are often conflated: *telemetry volume*, *evidentiary relevance*, and *active governance workflow priority*. In this paper, *active governance workflow* is the umbrella term for creation or update of an assessment, risk-reassessment, issue, incident, compliance-review, audit, or remediation workflow. It is distinct from *risk treatment*, which retains its established risk-management meaning. A large number of records may be necessary for later reconstruction, while only a small subset may currently support a claim that is eligible for an active governance workflow.

Raw telemetry $\neq$ qualified evidence.
Qualified evidence $\neq$ claim requiring active governance workflow.
Governance workflow relevance $\neq$ risk rating or legal conclusion.

6.1 Preservation Posture Versus Governance Workflow Posture

The first separation is between *evidence preservation posture* and *governance workflow posture*. Preservation determines what is captured and retained, for how long, and under which integrity and access protections. It is constrained first by applicable legal, regulatory, contractual, audit, and sectoral requirements. System criticality, threat model, organizational risk appetite, and operational capacity can then shape discretionary capture and retention choices where those obligations permit. The EU AI Act, for example, requires high-risk AI systems to support automatic recording of events relevant to specified risk, monitoring, and operational purposes, while DORA's technical standards require financial entities to identify logged events, retention periods, and purpose-aligned levels of detail [6, 23].

Governance workflow activation determines which preserved or newly observed material is actively promoted into control assessment, risk reassessment, issue management, incident handling, compliance review, audit, or remediation. Broad preservation therefore does not imply broad workflow activation. A critical-infrastructure or highly regulated organization can retain extensive telemetry and still use selective evidence promotion designed to reduce analyst and GRC workload. Conversely, a lower-criticality organization can adopt a more risk-adaptive preservation posture, provided that mandatory logging and retention obligations are not weakened.

This distinction also avoids a false optimization target. The primary efficiency gain of evidence-governed promotion is not necessarily lower storage volume. Even where raw-data retention remains broad, the model can reduce the volume of material that must be normalized into findings, assessed as controls, converted into GRC objects, reviewed by humans, or assigned remediation owners. Risk-adapted preservation can additionally reduce retained data only where applicable obligations and the governed risk posture allow it.

6.2 Evidence Promotion Chain

The evidence-governed promotion chain summarized in the Introduction is operationalized here. The transition from candidate to qualified evidence relies on the evidence semantics defined in the earlier evidence model rather than on an opaque importance score [2]. Qualification asks which evidentiary property the record can support, through which mechanism, under which assumptions and limitations. The next transition asks whether the resulting bounded claim is relevant to an applicable requirement, control objective, risk, incident, or mandatory escalation rule. Within this promotion path, only then does it become a Governance Workflow Candidate. Table 2 states the question and non-implication boundary at each promotion stage.

A compact promotion decision can be represented as

$$\text{promote}_{P_v}(E, c, g) \in \{\text{promote}, \text{hold}, \text{defer}\},$$

where E is the qualified evidence set, c is the bounded claim, g is the declared governance, control, risk, incident, or compliance relevance, and P_v is the governed and versioned promotion policy. The notation is a conceptual policy filter rather than a numerical importance score. P_v denotes versioned routing rules whose implementation may be deterministic, rule-based, human-reviewed, or AI-assisted. The rules can consider evidentiary sufficiency, applicability, control or risk relevance, materiality, workflow state, mandatory triggers, and designated review authority. *Promote* means that the claim is suitable for creation or update of an active assessment, issue, risk-reassessment, incident, compliance-review, audit, or remediation workflow. *Hold* means that the current evidence and ordinary promotion policy do not justify workflow activation at that time. It does not delete or invalidate the underlying evidence. *Defer* means that the promotion decision itself cannot yet be completed because evidentiary support, relevance, applicability, or required governance context remains open. This avoids overloading the control-assessment state *Unresolved*.

Three invariants constrain this policy. First, promote, hold, and defer outcomes retain the evidence references, policy version, decision context, and limitations needed for later review. Second, later evidence, policy, or applicability changes can trigger re-evaluation without silently rewriting the historical decision. Third, mandatory preservation, notification, reporting, or escalation obligations take precedence over an ordinary hold wherever those obligations apply. Promotion is therefore a governed routing decision over a bounded claim, not a measure of truth and not a substitute for mandatory duties. A *promote* outcome establishes workflow eligibility under the declared policy. It does not itself execute or authorize the downstream workflow.

As an illustrative policy instance, a payment-control claim under P_v may be promotion-eligible when the applicability context is active, the required evidence is qualified for the assessed transaction, the control assessment is *Contradicted* or materially *Unresolved*, and the declared workflow policy requires review for that control and transaction class. Missing required applicability or evidence context instead supports *defer*, while a policy-defined non-material case can remain on *hold*. Mandatory notification or escalation rules override that ordinary routing logic where applicable. This example illustrates one versioned policy instance rather than a universal promotion rule.

6.3 Deployment Profiles and AI Assistance

The same semantic model can support different operating profiles. A *comprehensive evidence posture* can use broad capture and retention, extensive security instrumentation, and enterprise GRC platforms while still promoting only treatment-relevant bounded claims. A *risk-adaptive evidence posture* can prioritize capture, retention, and assessment effort according to applicable obligations, system criticality, risk appetite, and organizational capability. The second profile is not a weaker definition of evidence. It is a different resource and risk posture around the same qualification and claim boundaries.

AI can assist both profiles by correlating high-volume records, identifying candidate evidence, mapping candidate claims to control or risk objects, detecting conflicts or missing support, preparing evidence packages, and proposing treatment options. Current practice already shows agentic or generative AI being applied to regulation-and-controls mapping, control testing, evidence analysis, and GRC workflows [40, 42, 39]. Such assistance should not be treated as an autonomous compliance authority. In deployed or evaluated operational workflows,

Stage	Question	Does not imply
Raw telemetry	What did a source record or emit?	that the record is relevant, trustworthy, or complete
Contextualized observation	Who or what, when, where, and through which observation path?	that the observation supports a governed claim
Candidate evidence	Could the observation support a defined evidentiary property or evidence requirement?	that it is adequate for the intended assessment
Qualified evidence	Is the evidence suitable for this requirement under stated assumptions and limits?	that a control is satisfied or violated
Bounded claim / finding	What specific statement is supported, contradicted, or unresolved?	a risk rating, incident truth, malicious intent, or legal conclusion
Control / risk relevance	Does the claim map to an applicable control, risk, obligation, incident, or mandatory trigger?	that active governance workflow is required, authorized, or proportionate
Governance workflow candidate	Does the claim justify creation or update of an active GRC, incident, audit, or remediation workflow under policy?	that the claim is true beyond its evidence boundary or that downstream treatment is authorized, implemented, or effective

Table 2. Evidence-promotion stages. The chain is designed to reduce active assessment volume without erasing preservation, uncertainty, or claim boundaries.

an AI-generated preliminary triage, mapping, or promotion recommendation is itself a bounded claim and should retain, at minimum, source-evidence references, model and version, relevant configuration or context, target claim or control, assessment time, assumptions, limitations, and the subsequent human or governance disposition where one exists. Where practical for a material decision, the retained context should also identify the input set and the disposition of material contrary evidence. Authorization of downstream interventions, risk treatment and acceptance, legal conclusions, and verification of mitigation effectiveness remain separately governed decisions.

This paper stops short of defining the implementation architecture for a proportionate GRC deployment.

Section 6 provides the design-synthesis answer to RQ3: high-volume observations are separated from active governance workload by preserving evidence independently from promotion, qualifying candidate evidence before claim formation, mapping only bounded claims to governed relevance, and applying a versioned promotion policy that can promote, hold, or defer without deleting preserved evidence or suppressing uncertainty.

7 Operational PDCA Lifecycle and GRC Integration

The architecture is organized around a plan–do–check–act lifecycle for an agentic system rather than around the publication order of the underlying research components. Figure 1 presents the simplified lifecycle view used in this paper. PDCA is used as an operational management frame because the relevant standards and regulations emphasize continuous risk management, monitoring, testing, review, and improvement [6, 9, 11, 12, 14]. The research contribution is not PDCA itself. The contribution is the placement of evidence-producing and evidence-assessing mechanisms within that lifecycle while preserving claim boundaries.

Figure 1 intentionally shows two distinct information paths into existing GRC platforms. CHECK provides bounded findings and assessment results, while the Evidence Foundation provides qualified evidence with preserved provenance and limitations. A separate outbound context edge represents governed information returned from the GRC platform to other governance objects, including policy state, control state, risk context, and assessment scope. These exchanges can enter existing governance, risk management, compliance, audit, issue, remediation, and incident processes without requiring a new parallel workflow. When an incident or suspected incident requires deeper reconstruction, the same evidence streams can be combined within evidence-based investigation, including incident analysis and reconstruction, to produce bounded incident claims and reassessment inputs that return to those existing processes. Evidence-based investigation is therefore conditional rather than a mandatory gateway for ordinary control or compliance evidence.

7.1 Plan

The planning phase identifies applicable requirements, risks, control objectives, evidence requirements, testing strategy, and architectural constraints. Governance defines roles, responsibilities, decision rights, policies, oversight, segregation of duties, and risk-acceptance authority. Risk management connects agentic threat and failure scenarios to control selection and treatment. Compliance and assurance planning determine which evidence will later be required to assess selected controls.

7.2 Do

The implementation and operation phase develops and configures agents, workflows, tools, access paths, security controls, monitoring, and evidence capture. Pre-deployment and change-driven penetration testing should preferentially occur in a controlled assessment environment when adversarial exposure may alter persistent or adaptive state. This is consistent with the broader idea that testing can occur during development and before deployment. The EU AI Act explicitly requires testing of high-risk systems during development as appropriate and before placing them on the market or putting them into service [6]. DORA likewise requires resilience testing and, for specified systems, vulnerability assessment before deployment or redeployment [9].

7.3 Check

The checking phase combines runtime monitoring, A-IDS findings, existing security-tool telemetry, audit evidence, penetration testing, investigation-derived evidence, control assessment, and compliance assessment. When an incident or suspected incident is investigated, the checking function can acquire and correlate available observations, evaluate candidate claims, and rate their claim-specific evidentiary adequacy. This paper uses only the bounded output interface needed for compliance assessment. Penetration testing is not limited to one late lifecycle step. It can be periodic, risk-driven, change-triggered, or initiated after a material incident or remediation. GDPR explicitly includes regular testing, assessment, and evaluation of security measures within Article 32, NIS2 requires policies and procedures to assess effectiveness of cybersecurity risk-management measures, and DORA requires recurring digital operational resilience testing [7, 8, 9]. Findings remain bounded by the tested system, configuration, state, visibility, and test conditions.

7.4 Act

The action phase covers authorized response and recovery, remediation, policy and configuration changes, lessons learned, and retesting. A-IPS supports governed intervention and recovery evidence. Incident-derived claims can support updates to agentic-process-related risks, control gaps, treatment options, and verification requirements in an existing GRC risk register. The evidentiary rating of a claim should inform, but remain distinct from, the rating of the associated risk. Likelihood, impact, risk appetite, and treatment authority remain risk-management decisions. Evidence-based investi-

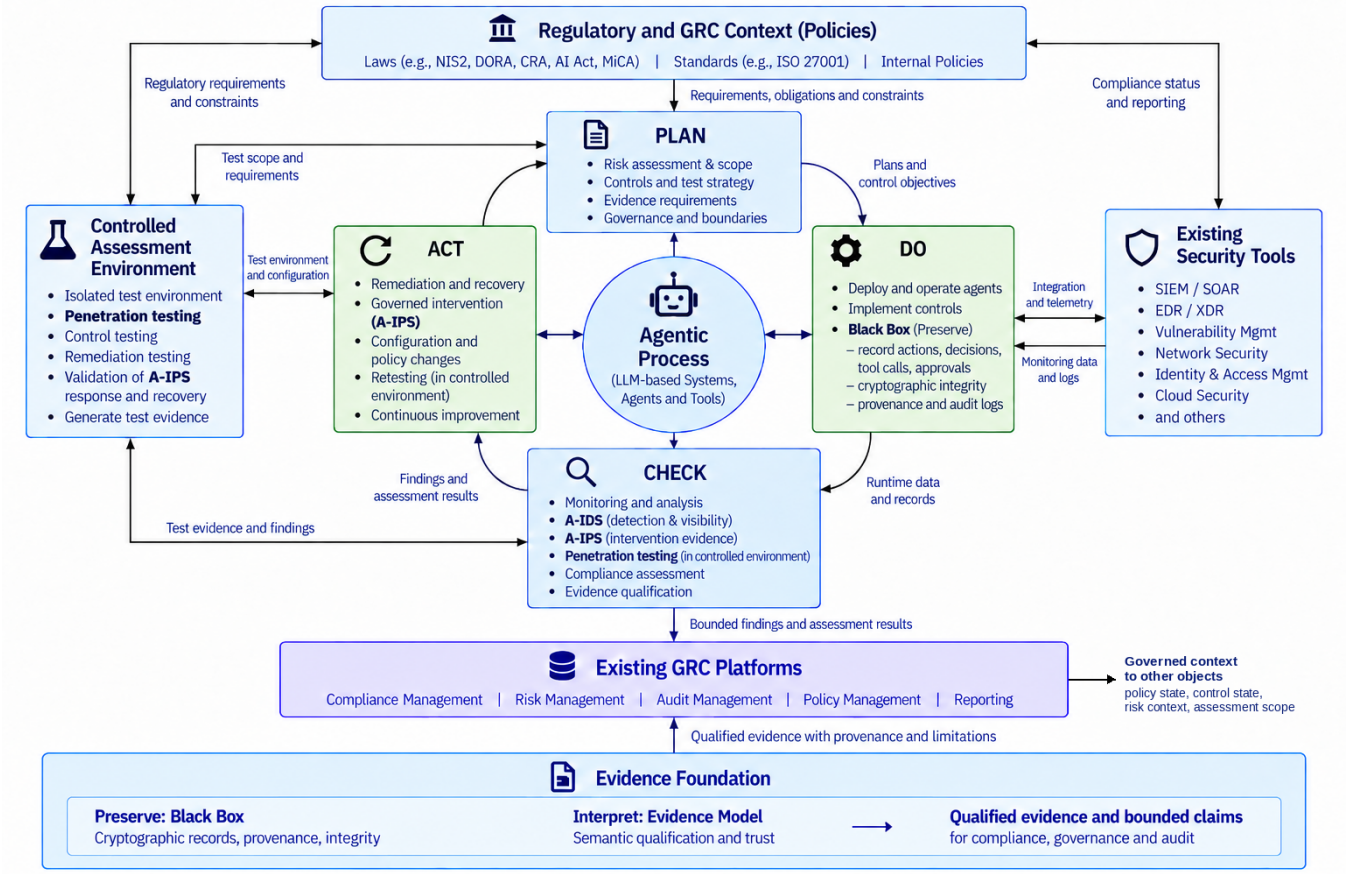

Figure 1. Operational lifecycle and evidence governed GRC integration for secure and compliant agentic processes. The two inbound paths to existing GRC platforms distinguish bounded findings and assessment results from qualified evidence with provenance and limitations. The outbound context edge indicates that existing GRC platforms can also provide governed context, including policy state, control state, risk context, and assessment scope, to other governed objects. These exchanges support existing governance and assurance processes without increasing evidentiary strength.

gation, including incident analysis and reconstruction, can use retained evidence to determine what can be supported about what occurred, which controls were affected, and which changes should feed back into planning, monitoring rules, test cases, and risk assessment. DORA explicitly requires its ICT risk management framework to be reviewed following major ICT incidents and relevant testing or audit conclusions and to be continuously improved from lessons learned [9].

8 GRC and Evidence Governance

Governance, Risk Management, and Compliance (GRC), together with evidence governance, are cross cutting functions rather than isolated lifecycle steps. Figure 1 separates the Regulatory and GRC Context, Existing GRC Platforms, and the Evidence Foundation while showing how they interact with the operational lifecycle. The assessment model provides the structured path from external and internal requirements to bounded assessment without collapsing these distinct functions into a single control plane.

Governance establishes accountable roles, decision rights, policies, segregation of duties, oversight, exception authority, and risk-acceptance authority. This aligns with management-body responsibilities visible in NIS2 and DORA and with broader management-system approaches in ISO/IEC 27001 and ISO/IEC 42001 [8, 9, 11, 12].

Risk management identifies risks, selects controls, evaluates residual risk, and determines when changes, incidents, or new evidence require reassessment. Investigation-derived evidence can update the factual basis of a risk entry, but an evidence rating is not itself a risk rating. Strong evidence that a control failed does not by itself determine likelihood or impact, while an incompletely reconstructed incident can still justify consideration of risk treatment or mitigation when a bounded weakness claim is sufficiently supported. The EU AI Act's high-risk risk-management process is continuous and iterative, while NIST AI RMF similarly treats risk management as continuous across the lifecycle [6, 14].

Compliance and assurance determine applicability, map requirements to control objectives, define evidence requirements, qualify received evidence, perform control assessment, and issue bounded assessment claims under an identified authority. The model does not replace external auditors, legal functions, certification bodies, supervisory authorities, or GRC platforms. It provides a structured evidence interface to them.

8.1 Integration with Existing GRC Platforms

The proposed model is intended to complement rather than replace existing GRC platforms. Its integration boundary is bidirectional, but it is deliberately selective: not every captured record, log line, or detector output should become a GRC object. Evidence promotion first determines which bounded claims are eligible for active governance workflows under the declared policy. Existing GRC systems can supply governed context such as control identifiers, control ownership, risk-register references, policy versions, assessment scope, issue status, remediation state, and management decisions. The agentic-process assessment layer can return qualified evidence references, control-assessment states, bounded compliance claims, incident-derived claims, unresolved issues, reassessment triggers, and links to remediation or audit workflows. For risk-register use, a record can reference the affected process or asset, relevant control gap, evidentiary basis and status, residual uncertainty, proposed treatment, and required verification while keeping likelihood and impact assessment separate.

This interface is a semantic design objective rather than a demonstrated product integration. A deployment may map the conceptual records to a commercial GRC platform, an internal control system, an audit repository, a ticketing workflow, or a machine-readable interchange format such as OSCAL where appropriate [16]. The model does not prescribe one vendor, database, API, or serialization. Any later implementation should preserve the assessment boundary instead of flattening a qualified result into an unqualified "compliant" or "non-compliant" flag.

The conceptual exchange records are serialization-neutral. In an implementation, they may be represented as structured JSON objects and exchanged through APIs, message buses, event streams, or file based interfaces. JSON is one practical option for connecting the proposed components with existing security, monitoring, incident response, and GRC systems, but it is not an architectural requirement. A JSON representation can preserve fields such as evidence references, policy and control identifiers, scope, covered period, status, assumptions, limitations, assessment authority, and destination references. The serialization must preserve the original claim boundaries and evidentiary limits.

The integration therefore follows a simple rule: *GRC ingestion does not amplify evidence.* A bounded technical finding remains bounded after ingestion. A control assessment remains limited to its declared scope, time, evidence set, assumptions, and assessment authority. Existing GRC workflows can use the record for risk treatment, issue management, audit planning, management reporting, or remediation tracking, but those downstream actions do not retroactively increase the evidentiary strength of the source record.

9 Evidence Producers and Existing Security Tools

The compliance model is mechanism-neutral. Evidence can come from the agentic-process research layers, conventional security infrastructure, organizational processes, or third-party assurance activities. Raw logs and telemetry are inputs to this process, not automatically qualified evidence. Table 3 illustrates the intended relationship after source-specific qualification.

All producer outputs remain bounded by their declared source properties, scope, assumptions, limitations, and observation conditions. No producer row implies broader compliance, complete visibility, or independent corroboration by itself.

The same control may require evidence from more than one producer. An access-control objective might require IAM records, agentic-process provenance, evidence that approval was valid at execution time, and adversarial testing that attempts to bypass the control. Combining these artifacts can strengthen an assessment only if their scopes and dependencies remain explicit. Multiple records from the same compromised observation path do not automatically constitute independent corroboration.

Existing security tools can also act as enforcement points rather than only as evidence producers. A-IPS is architecturally additive and does not replace preventive, containment, access-control, network-security, isolation, or response mechanisms. Instead, it provides an evidence-governed decision and authorization layer that can invoke or coordinate available enforcement capabilities. Firewalls, egress controls, IAM mechanisms, gateways, sandboxing, process controllers, and other technical controls can therefore execute an authorized intervention. A-IPS preserves evidence about the decision, authorization, execution, observed effect, and recovery for subsequent governance and compliance assessment. These records can support downstream assessment and, for narrowly defined response controls, may be sufficient for a bounded control assessment. They do not by themselves establish broader compliance beyond the declared scope, policy, and evidence conditions.

GRC platforms are intentionally not treated as another security detector in Table 3. They are integration and workflow environments that can both provide governed context and consume bounded assessment outputs. This separation mirrors the distinction between security evidence production and compliance authority, i.e., a GRC system can orchestrate or store an assessment without becoming the source of truth for the technical event that the assessment references.

10 Controlled Assessment Environment and Penetration Testing

The controlled assessment environment is a supporting environment rather than a separate compliance authority. It can host pre-deployment tests, controlled replicas, penetration testing, red-team scenarios, incident reproduction, change validation, and regression testing. This separation is particularly important when the target retains memory, adaptive state, credentials, external side effects, or attacker-supplied artifacts.

Three penetration-test triggers are relevant to the lifecycle. *Planned testing* occurs before deployment or on a defined periodic schedule. *Change-triggered testing* follows material model, tool, memory, policy, authority, or architectural changes. *Incident- or remediation-triggered testing* attempts to reproduce a failure or validate a mitigation after an incident or identified weakness. These categories are operational triggers rather than exhaustive regulatory classifications.

The sandbox does not imply that all production-relevant testing can or should be isolated from production. Some controls may require observation in their operational environment for a sufficiently representative assessment. Production testing must additionally account for test-induced state changes, since adversarial exposure can alter persistent or operational state. The governance question is therefore which tests are safe and meaningful in production, which require a production-equivalent clone or staged environment, and which evidence is lost when the environment is not equivalent. A failed or successful test in a materially different environment should not be treated as equivalent to production evidence without qualification.

11 Regulatory and Assurance Framework Mapping

The framework is deliberately regulation-neutral at the architecture level, while its operational assessment value depends on instantiation against concrete requirements. Table 4 shows selected examples. The table is illustrative and does not constitute a complete legal mapping.

This mapping also explains why one universal "compliance score" would be misleading. Different frameworks ask different questions, apply to different actors, and use different notions of control, risk, conformity, responsibility, and evidence. The common layer is not a universal verdict. It is a structured evidence-to-assessment interface.

12 Worked Example: Approval Before Payment

Consider an agentic process with a planning agent, an approval agent, and a payment agent. The planning agent prepares a payment instruction. Organizational policy requires independently validated approval before payment execution. Depending on context, external requirements may also create obligations around authorization, oversight, security, logging, or accountability. The example intentionally begins with the internal control objective because legal applicability must be assessed separately.

Evidence producer	Representative evidence	Compliance-assessment use and boundary
Preserve / Black Box	committed records, provenance, integrity, approvals, tool-call and decision artifacts	supports traceability and later verification, but not semantic truth, completeness, or legal sufficiency by itself
Interpret / Evidence Model	evidence properties, assumptions, limitations, threat scope	qualifies what a record can support and supports explicit separation between mechanism and claim
A-IDS	runtime observations, visibility state, bounded security findings	supports monitoring and deviation assessment, but absence of findings does not establish compliance
A-IPS	authorization, intervention, enforcement invocation, execution, effect, and recovery records	supports bounded assessment of authorization, invocation, observed effect, and recovery under declared conditions
Penetration testing	adversarial findings, test governance, retest claims, test conditions	supports control-effectiveness assessment under tested conditions, but successful retest does not establish permanent security
Evidence-based investigation (future layer)	source-bounded incident claims, reconstruction evidence, control failures, impact and recovery observations, re-evaluation history	supports control and risk reassessment, treatment and verification planning, but claim sufficiency is granularity-dependent and does not create incident ground truth
Existing security tools	SIEM/XDR, IAM, vulnerability, DLP, network, cloud, SOAR, threat intelligence and related telemetry	complementary evidence whose relevance and trust properties depend on source, scope, configuration, and observation path

Table 3. Illustrative evidence producers. The architecture integrates existing tools rather than replacing them.

Framework	Selected requirement areas	Relevant evidence / lifecycle links	Assessment boundary
EU AI Act	risk management, testing, logging, human oversight, deployer monitoring, post-market monitoring	PLAN risk/control design, DO testing and deployment, CHECK logs/monitoring, ACT corrective action and reassessment	applicability depends on system and role, technical evidence does not alone establish conformity
GDPR	accountability, data protection by design, security of processing, regular testing and evaluation	access and processing evidence, security controls, monitoring, testing, incident and recovery evidence	security-control evidence addresses only selected obligations and processing contexts
NIS2	management oversight, cybersecurity risk management, incident handling, effectiveness assessment, reporting	governance records, monitoring, A-IDS/A-IPS, test evidence, incident analysis, management response	entity scope, national implementation, materiality, and reporting duties require separate assessment
DORA	ICT governance, ICT risk management, incident management, resilience testing, penetration testing, TLPT	full PDCA lifecycle, controlled testing, existing security tools, incident/recovery evidence, remediation and retest	applies to defined financial entities with proportionality and specific testing requirements
MiCA	organizational requirements, ICT continuity and security, risk assessment, service and transaction records	governance records, identity and authorization evidence, transaction and service records, DORA-linked ICT evidence, monitoring and corrective action	sector-specific applicability depends on actor and service, and record-keeping or technical-control evidence does not establish overall MiCA compliance
ISO/IEC 27001	information-security management system, risk management, continual improvement	governance, risk assessment, controls, audit evidence, monitoring, corrective action	certification or conformity is determined against the standard, not by one evidence producer
ISO/IEC 42001	AI management system, responsible AI governance, continual improvement	AI governance, lifecycle evidence, risk and control assessment, monitoring and improvement	management-system evidence does not equal compliance with separate laws
ISO/IEC 23894 / 42005	AI risk management and AI system impact assessment	risk identification, impact assessment, lifecycle reassessment, treatment and evidence updates	guidance and impact assessment support risk decisions rather than a universal compliance verdict
ISO/IEC 42006	audit and certification of AI management systems	assurance authority, audit competence, audit evidence and certification context	certification-body assessment remains distinct from technical evidence production
NIST AI RMF	Govern, Map, Measure, Manage	cross-cutting governance plus lifecycle risk mapping, measurement, management and reassessment	voluntary risk-management framework rather than a legal compliance determination
NIST 800-53A / OSCAL	control assessment, assessment plans/results, observations, findings, evidence, risks	structured control-assessment and machine-readable evidence integration	assessment structure does not determine external legal applicability

Table 4. Illustrative framework mapping. The same evidence architecture can support different assurance contexts while preserving framework-specific applicability and authority.

Plan. The organization records the applicable policy, identifies the control objective "payment execution requires valid independent approval," defines the responsible roles, and specifies evidence requirements. Required evidence may include the approval request, approver identity, approval scope, policy version, execution time, payment parameters, and an observation that the payment action remained within the approved scope.

Do. The workflow implements independent approval and preserves the relevant records. IAM provides authenticated identity evidence. The black-box layer preserves approval and execution artifacts with integrity and provenance properties. A controlled sandbox is used to test whether the payment agent can be induced to act on an unsupported peer claim or stale approval.

Check. A-IDS can compare observed approval and execution events with the governed expectation that approval is due before the payment action. Existing security telemetry can contribute identity, access, network, or transaction evidence. Penetration testing can attempt to bypass the approval path through prompt injection, delegated authority confusion, replay, shared-state manipulation, or other controlled adversarial conditions. A control assessor then qualifies the evidence against the declared evidence requirements.

Suppose the approval record is present and integrity-verifiable, but the evidence shows that the approval applied to a different payment amount. A naive pipeline could treat the authentic approval artifact as sufficient to mark the control satisfied. The present model does not permit that widening. The evidence binding retains the approval scope and amount, while the governed assessment method compares those properties with the executed transaction. The control assessment can therefore be *Contradicted* for the assessed transaction even though the approval record itself is authentic. This near miss illustrates why integrity evidence, authorization evidence, and policy-conformance evidence are distinct and why evidence qualification alone does not determine the control state.

Act. If the runtime finding reached the required evidence, admissibility, and authorization thresholds, an authorized A-IPS intervention could have invoked or coordinated an available enforcement mechanism to block, isolate, restrict, or contain the action. A firewall, gateway, IAM control, sandbox, process controller, or payment-specific control would remain the technical enforcement point. The assessment does not rely on an agent's self-report of execution or enforcement. Security-relevant state changes require a separately governed observation path whose trust assumptions and visibility are explicit. That path may be deterministic or otherwise indepen-

dently controlled according to the deployment. If execution occurred, evidence-based investigation can reconstruct the approval path, policy state, tool action, and response. Remediation may change approval binding, freshness, authority validation, or monitoring rules. A retest in the controlled environment then provides new evidence. The compliance assessment is updated rather than assumed to remain valid indefinitely.

The resulting compliance claim remains bounded. It may state that qualified evidence supports or contradicts the selected approval control for the identified transaction, policy version, period, and assessment scope. It does not establish that all payment controls, all applicable regulations, or the organization as a whole are compliant. Through an *AssessmentExchange* record, that result can be linked to the corresponding control, risk, issue, evidence, and remediation objects in an existing GRC platform. The destination system may then track ownership, due dates, management response, or reassessment without changing the original claim boundary.

13 Continuous and Event-Driven Reassessment

Compliance assessment for agentic processes should be treated as a maintained state rather than a one-time label. Reassessment can be triggered by time, changes, incidents, tests, new evidence, or changed requirements. This follows the lifecycle logic already visible in regulatory and management-system sources. The EU AI Act links post-market monitoring to continuous compliance evaluation for high-risk systems, DORA requires review after major ICT incidents and testing or audit conclusions, and ISO management-system standards emphasize continual improvement [6, 9, 11, 12, 18].

Material triggers can include model or prompt changes, new tools or permissions, new external services, changes to agent roles or delegated authority, memory or retrieval changes, changed security controls, new vulnerabilities, incidents, penetration-test findings, changed laws or standards, and changes to organizational scope. A previous assessment should not silently survive a change that invalidates its assumptions.

This motivates an explicit relationship between current standing and historical evidence. Evidence can remain historically valid while no longer being sufficient for the current system state. A valid past assessment is therefore not automatically a current assessment. The assessment record should declare its covered configuration, effective period, and reassessment triggers.

14 Proposed Evaluation

The present work deliberately focuses on the conceptual specification of the assessment architecture, its evidence semantics, claim boundaries, promotion logic, and integration points. The next research step is a controlled proof of concept evaluation in a sandboxed agentic environment. Such an evaluation should operationalize the proposed roles and decision stages, generate reproducible execution traces, introduce predefined normal, incomplete, conflicting, and adversarial conditions, and test whether the resulting evidence can support the intended bounded assessment states and promotion decisions. This controlled setting is intended to examine whether the proposed architecture behaves as specified before broader evaluation across different agent configurations, policies, models, and operational environments.

A useful experiment would replay normal and adversarial multi-agent traces through alternative promotion policies while keeping the underlying source records available for later reassessment. The trace corpus should include routine business flows, policy-relevant deviations, incomplete visibility, contradictory evidence, and changes that invalidate prior assessments. Synthetic or benchmark-derived agent traces may be used, but benchmark attacks should be adapted to the specific control and evidence questions being evaluated and must not be interpreted as evidence of legal or regulatory compliance. The experimental environment may combine task agents with separately controlled evidence preservation, detection, intervention, testing, and compliance-assessment functions so that individual stages can first be evaluated independently and later as an integrated evidence and governance chain. At least one proof-of-concept scenario should directly attack the observation or promotion path itself, for example through poisoned evidence summaries, prompt injection through log content, source suppression, policy downgrade, or unauthorized GRC-object creation, so that the threat model is tested rather than merely stated.

The principal measures are retained source records, contextualized observations, candidate and qualified evidence items, bounded claims produced, claims promoted, held, or deferred, human review volume, downstream GRC objects, false escalations, and missed claims that should have entered an active workflow under the declared scenario policy. The evaluation manifest should separate four reference layers rather than collapsing them into one ground-truth label: the injected scenario state, the expected technical control state, the evidence-availability state, and the policy-defined workflow relevance. Legal or regulatory applicability should be supplied separately as an authoritative scenario input rather than inferred from the injected technical condition. Where a judgment is not deterministic, the proof of concept should use either a declared policy oracle or independent blinded adjudication and retain disagreement rather than silently converting it into ground truth. Additional analysis should measure how often evidence is insufficient or indeterminate, how sensitive results are to policy versions, and whether later evidence correctly triggers reassessment of held or deferred records. The evaluation should compare at least a comprehensive evidence posture and a risk-adaptive evidence posture over the same governed agentic-process scenarios. The purpose is not to optimize one universal promotion threshold, but to test whether evidence-governed promotion preserves claim boundaries and can reduce unnecessary governance work without increasing missed workflow-relevant claims.

15 Evidence-Based Investigation and Feedback

Incidents create a particularly important feedback path. Runtime detection and response do not end the assurance cycle. NIST SP 800-61 Rev. 3 similarly embeds incident response across cybersecurity risk-management activities rather than treating it as an isolated terminal process [21]. An incident can reveal that a control was incorrectly designed, bypassed, not observed, not enforced, or based on an invalid assumption. Evidence-based investigation should therefore draw on preserved process evidence, A-IDS findings, A-IPS decisions and effects, security-tool telemetry, test evidence, and human or organizational records.

The output of evidence-based investigation should feed risk reassessment, control redesign, monitoring changes, new test scenarios, remediation, verification, and compliance reassessment. Investigation can include incident analysis and reconstruction and can organize currently available material into claim-specific evidence assessments that remain open to later re-evaluation. The present model uses only the bounded interface required for compliance and governance feedback. It does not define a full investigation method for evidence acquisition, source relationships, reconstruction, causal uncertainty, propagation, intervention history, recovery, or evidence composition.

For the present model, the required interface is narrower. Investigation-derived evidence must retain source, scope, res-

olution, time, source relationships, causal status, sufficiency, assumptions, and limitations before it influences a control, risk, or compliance claim. The resulting evidentiary rating can support a GRC risk-register update, but it must remain distinct from likelihood and impact. Likewise, an investigation-derived weakness claim can justify consideration of mitigation even when a deeper root-cause or compromise hypothesis remains unresolved, provided that the weakness itself is sufficiently supported for the proposed risk-treatment or mitigation decision.

The downstream feedback path is summarized below.

Observation → Qualified Evidence → Bounded Incident Claim → Risk Reassessment → Risk Treatment Decision → Implementation → Verification → Re-Evaluation

Risk-treatment decisions do not authorize themselves, and implementation does not establish effectiveness. Under the proposed model, an authorized governance or response function must approve the action, and later testing, monitoring, or separately controlled verification must support any bounded effectiveness claim required by the governing policy. In particular, mitigation derived from an incident is not equivalent to demonstrated effective mitigation.

16 Limitations and Future Work

The present version specifies a conceptual assessment model and does not report empirical validation. It does not determine legal applicability, replace qualified legal or audit judgment, guarantee complete evidence capture, or establish that every requirement can be reduced to a machine-testable control. Evidence producers may be incomplete, compromised, mutually dependent, or based on different observation windows. Requirements can change over time and may depend on facts not visible to the technical system. Control objectives can be correctly assessed while the underlying legal interpretation remains disputed or incomplete.

The model also does not solve independence by aggregation. Multiple evidence items can share the same compromised infrastructure, credentials, clock, telemetry path, policy source, or semantic interpreter. Evidence qualification must therefore record common-mode dependencies where they materially affect the assessment. Observation and promotion paths are themselves part of the threat model. Forged or suppressed records, poisoned summaries, prompt injection through log content, duplicate amplification, manipulated source relationships, policy downgrade, or unauthorized GRC-object creation can distort downstream assessment. The conceptual model does not guarantee resistance to these attacks. Implementations must protect source integrity, policy and version integrity, provenance, access control, and independently governed observation where required. If the relevant observation or enforcement path cannot be trusted or sufficiently observed, the assessment should remain *Insufficient evidence* or *Unresolved* rather than infer successful control operation.

The PDCA lifecycle is an organizing view rather than a mandatory implementation sequence. Runtime monitoring, evidence capture, governance, and risk management are continuous activities that can span phases. Some production testing may be necessary or materially more informative, and some controls may not be meaningfully validated solely in a sandbox. The model therefore requires environment differences to be recorded rather than assuming test equivalence.

Future work should first implement the specified architecture in a controlled test environment and exercise its evidence, detection, intervention, testing, and assessment boundaries under reproducible conditions. This staged evaluation should test individual functions before an integrated end-to-end run so that observed failures can be attributed to evidence capture, qualification, detection, authorization, enforcement, or assessment rather than being hidden inside one composite agent workflow. Subsequent work should formalize evidence-requirement profiles, conflict handling, evidence sufficiency, promotion-policy calibration, assessment expiry, reassessment triggers, and interoperable interfaces to external audit and GRC platforms, including mappings to established machine-readable assurance formats where appropriate. The proposed evaluation above should compare comprehensive and risk-adaptive evidence postures in terms of retained data, analyst workload, false escalation, missed claims that should have entered an active governance workflow under the declared policy, and downstream GRC effort. Future work may extend the present model toward evidence-based investigation around the research direction *Investigate → Evaluate → Rate Evidence → Re-Evaluate*. Incident analysis and reconstruction are possible methods within that broader investigation layer. Such work would examine how heterogeneous observations can be acquired, related, evaluated, and composed into incident-level claims whose granularity, causal support, sufficiency, and re-evaluation history remain explicit. Future work may also examine how AI-assisted evidence triage, evidence promotion, and compact structured registers can support proportionate GRC operating models while preserving the same mandatory floors and evidentiary boundaries. AI-generated mappings, summaries, evidence packages, and treatment proposals will remain bounded claims rather than autonomous compliance decisions. Behavioral fingerprinting of individual agents and agentic processes can subsequently enter the same architecture as another evidence producer, provided that behavioral divergence is not treated as proof of compromise and similarity is not treated as proof of state equivalence.

17 Conclusion

Compliance assessment for agentic processes should not collapse technical evidence into binary legal conclusions or convert telemetry volume into governance workload. The proposed model separates applicability, requirements, evidence needs, evidence producers, evidence qualification, control assessment, bounded compliance claims, and their exchange with existing GRC workflows. It adds an evidence-governed promotion chain that distinguishes evidence preservation from workflow activation and therefore allows extensive technical records to remain available without requiring every record to become a finding, risk, issue, or remediation object.

The central conclusion is deliberately limited. Under the proposed model, security and assurance evidence strengthens compliance assessment to the extent that the conditions and limitations governing what it establishes remain explicit. Evidence-governed promotion should therefore move only bounded claims that satisfy the declared workflow policy rather than merely high-volume or high-severity telemetry. A comprehensive evidence posture and a risk-adaptive posture can share the same evidence semantics while differing in capture, retention, tooling, and resource intensity, subject to mandatory regulatory and contractual floors. The architecture is intended to support integration with enterprise GRC platforms and to provide a semantic basis for possible proportionate, AI-assisted GRC approaches. These are design objectives to be tested in the proposed controlled evaluation rather than demonstrated operational benefits of the present conceptual version.

Evidence-based investigation can additionally support risk and control reassessment when source relationships, causal status, claim granularity, and evidence sufficiency remain explicit. In this paper, evidence-based investigation is the broader future layer, with incident analysis and reconstruction as possible methods within it. Evidence rating must remain distinct from risk rating, and a risk-treatment or mitigation decision must remain distinct from demonstrated effectiveness. Existing GRC platforms can provide governed context and operationalize

bounded assessment outputs through control libraries, risk and issue workflows, audit processes, remediation tracking, and management reporting, but integration must not amplify source evidence or erase its limitations. The resulting architecture is therefore not a replacement for logging, security tooling, audit, risk management, or enterprise GRC. It is a claim-preserving assessment and exchange interface that structures what heterogeneous evidence can support, which claims are eligible for governed workflow activation, and how those claims can move through governance without becoming stronger merely because they were aggregated, scored, or ingested into a larger system.

References

[1] A. Brömme, "A Black Box for Agentic Processes: Blockchain-Anchored Evidence for AI Agent Communication, Human Oversight, and GRC Audits," arXiv:2609.04017 [cs.CR], 2026. doi:10.48550/arXiv.2609.04017. https://arxiv.org/abs/2609.04017

[2] A. Brömme, "An Evidence Model for Agentic Processes: Evidence Claims, Trust Assumptions, and Policy Assessment," arXiv:2609.08481 [cs.CR], 2026. doi:10.48550/arXiv.2609.08481. https://arxiv.org/abs/2609.08481

[3] A. Brömme, "Intrusion Detection for Agentic Processes: Evidence-Based Runtime Monitoring," Zenodo, Version v0.9.1.8, 2026. doi:10.5281/zenodo.22764610. https://doi.org/10.5281/zenodo.22764610

[4] A. Brömme, "Intrusion Prevention for Agentic Processes: Evidence-Governed Runtime Response and Enforcement," Zenodo, Version v0.9.1.4, 2026. doi:10.5281/zenodo.22765147. https://doi.org/10.5281/zenodo.22765147

[5] A. Brömme, "Penetration Testing for Agentic Processes: Evidence-Governed Adversarial Assessment and Retesting," Zenodo, Version v0.9.0.10, 2026. doi:10.5281/zenodo.22766559. https://doi.org/10.5281/zenodo.22766559

[6] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," 2024, consolidated version 27 July 2026. https://eur-lex.europa.eu/eli/reg/2024/1689/2026-07-27/eng

[7] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 (General Data Protection Regulation)," 2016. https://eur-lex.europa.eu/eli/reg/2016/679/oj

[8] European Parliament and Council of the European Union, "Directive (EU) 2022/2555 on measures for a high common level of cybersecurity across the Union (NIS 2 Directive)," 2022. https://eur-lex.europa.eu/eli/dir/2022/2555/oj

[9] European Parliament and Council of the European Union, "Regulation (EU) 2022/2554 on digital operational resilience for the financial sector (DORA)," 2022. https://eur-lex.europa.eu/eli/reg/2022/2554/oj

[10] European Parliament and Council of the European Union, "Regulation (EU) 2023/1114 on markets in crypto-assets (MiCA)," 2023. https://eur-lex.europa.eu/eli/reg/2023/1114

[11] ISO, "ISO/IEC 27001:2022, Information security, cybersecurity and privacy protection – Information security management systems – Requirements," 2022. https://www.iso.org/standard/27001

[12] ISO, "ISO/IEC 42001:2023, Information technology – Artificial intelligence – Management system," 2023. https://www.iso.org/standard/42001

[13] E. Tabassi, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, National Institute of Standards and Technology, 2023. doi:10.6028/NIST.AI.100-1. https://doi.org/10.6028/NIST.AI.100-1

[14] National Institute of Standards and Technology, "AI RMF Core," AI Resource Center, accessed Sept. 2026. https://airc.nist.gov/airmf-resources/airmf/5-sec-core/

[15] Joint Task Force, "Assessing Security and Privacy Controls in Information Systems and Organizations," NIST Special Publication 800-53A Rev. 5, 2022, Release 5.2.0 updated 2025. doi:10.6028/NIST.SP.800-53Ar5. https://csrc.nist.gov/pubs/sp/800/53/a/r5/final

[16] National Institute of Standards and Technology, "OSCAL Assessment Results Model," accessed Sept. 2026. https://pages.nist.gov/OSCAL/learn/concepts/layer/assessment/assessment-results/

[17] N. Kahani, M. Barati, and D. Addae, "Runtime Compliance Verification for AI Agents," arXiv:2606.19242, 2026. https://arxiv.org/abs/2606.19242

[18] ISO, "ISO/IEC 23894:2023, Information technology – Artificial intelligence – Guidance on risk management," 2023. https://www.iso.org/standard/77304.html

[19] ISO, "ISO/IEC 42005:2025, Information technology – Artificial intelligence (AI) – AI system impact assessment," 2025. https://www.iso.org/standard/42005

[20] ISO, "ISO/IEC 42006:2025, Information technology – Artificial intelligence – Requirements for bodies providing audit and certification of artificial intelligence management systems," 2025. https://www.iso.org/standard/42006

[21] A. Nelson, S. Rekhi, M. Souppaya, and K. Scarfone, "Incident Response Recommendations and Considerations for Cybersecurity Risk Management: A CSF 2.0 Community Profile," NIST Special Publication 800-61 Rev. 3, 2025. doi:10.6028/NIST.SP.800-61r3. https://csrc.nist.gov/pubs/sp/800/61/r3/final

[22] K. Kent and M. Souppaya, "Guide to Computer Security Log Management," NIST Special Publication 800-92, 2006. doi:10.6028/NIST.SP.800-92. https://csrc.nist.gov/pubs/sp/800/92/final

[23] European Commission, "Commission Delegated Regulation (EU) 2024/1774 supplementing Regulation (EU) 2022/2554 with regulatory technical standards specifying ICT risk management tools, methods, processes and policies," 2024, incl. 2025 corrigendum. https://eur-lex.europa.eu/eli/reg_del/2024/1774/oj

[24] A. E. Muhammad, K.-C. Yow, and S. Alsenan, "Audit-as-code: a policy-as-code framework for continuous AI assurance," *Frontiers in Artificial Intelligence*, vol. 9, 2026. doi:10.3389/frai.2026.1759211. https://doi.org/10.3389/frai.2026.1759211

[25] Y. Li and S. Goel, "Artificial intelligence auditability and auditor readiness for auditing artificial intelligence systems," *International Journal of Accounting Information Systems*, vol. 56, art. 100739, 2025. doi:10.1016/j.accinf.2025.100739. https://doi.org/10.1016/j.accinf.2025.100739

[26] B. Santelmann, "Governance für KI-Agenten: Ein risikobasierte Perspektive für Transparenz und Nachvollziehbarkeit," *HMD Praxis der Wirtschaftsinformatik*, 2026. doi:10.1365/s40702-026-01311-x. https://doi.org/10.1365/s40702-026-01311-x

[27] L. Nannini, A. L. Smith, M. J. Maggini, E. Panai, S. Feliciano, A. Tiulkanov, E. Maran, J. Gealy, and P. Bisconti, "AI Agents Under EU Law," arXiv:2604.04604, 2026. https://arxiv.org/abs/2604.04604

[28] A. Kasirzadeh and I. Gabriel, "Agentic profiles for effective AI governance," *Nature*, vol. 656, pp. 320–328, 2026. doi:10.1038/s41586-026-10805-z. https://doi.org/10.1038/s41586-026-10805-z

[29] G. Lupo, B. Q. Vo, and N. Locke, "Trustworthy AI Posture (TAIP): A Framework for Continuous AI Assurance of Agentic Systems at Horizontal and Vertical Scale," arXiv:2603.03340, 2026. https://arxiv.org/abs/2603.03340

[30] M. Kaptein, V.-J. Khan, and A. Podstavnychy, "Runtime Governance for AI Agents: Policies on Paths," arXiv:2603.16586, 2026. https://arxiv.org/abs/2603.16586

[31] S. Bokkasam and B. Durgalakshmi, "Governing at Machine Speed: An Adaptive Intelligence Architecture for Real-Time AI Policy Enforcement," arXiv:2609.13466, 2026. https://arxiv.org/abs/2609.13466

[32] G. Zhu, "AcquireBound: Runtime Authorization for Resources Acquired by AI Agents," arXiv:2609.14744, 2026. https://arxiv.org/abs/2609.14744

[33] R. K. Surapani, P. K. D. Kakitapelli, A. Morampudi, and P. Padi, "Authorization Architectures for Tool-Using AI Agents," arXiv:2609.15906, 2026. https://arxiv.org/abs/2609.15906

[34] R. Uetz, P. Bönninghausen, L. Hackländer-Jansen, and M. Henze, "Can Risk-Based Alerting Mitigate Cybersecurity Alert Fatigue?" arXiv:2609.02465, 2026. https://arxiv.org/abs/2609.02465

[35] Committee of Sponsoring Organizations of the Treadway Commission, "Achieving Effective Internal Control Over Generative AI," 2026. https://www.coso.org/generative-ai

[36] ISACA, "Governing AI Across its Lifecycle: A Framework for Risk Practitioners," Apr. 15, 2026. https://www.isaca.org/resources/toolkits/governing-ai-across-its-lifecycle-a-framework-for-risk-practitioners

[37] The Institute of Internal Auditors, "The IIA's Artificial Intelligence Auditing Framework," 2nd ed., Sept. 13, 2024. https://www.theiia.org/en/content/tools/professional/2023/the-iias-updated-ai-auditing-framework

[38] EY, "Integrierte Governance für Informationssicherheit und KI: Ein praxisorientierter Leitfaden zur Verzahnung von ISMS und KI-Governance," May 2026. https://www.ey.com/de_de/functional/forms/download/2026/05/integrierte-governance-fuer-informationssicherheit-und-ki

[39] KPMG, “Governance und Compliance für KI – zentral gesteuert mit dem ServiceNow AI Control Tower,” 2026. https://kpmg.com/de/de/themen/digital-transformation/kuenstliche-intelligenz/governance-und-compliance-fuer-ki.html

[40] PwC, “Smarter risk and controls, powered by agentic AI,” June 30, 2026. https://www.pwc.com/gx/en/services/risk/regulation-controls-lifecycle.html

[41] Deloitte, “AI risk management and governance program,” accessed Sept. 20, 2026. https://www.deloitte.com/us/en/services/consulting/services/ai-risk-governance-program.html

[42] Deloitte, “ControlCatalyst.AI: AI for Internal Audit and Controls,” accessed Sept. 16, 2026. https://www.deloitte.com/us/en/services/audit-assurance/services/controlcatalyst-ai.html

[43] ServiceNow, “AI governance life cycle,” updated Mar. 12, 2026. https://www.servicenow.com/docs/r/de-DE/governance-risk-compliance/ai-risk-management/ai-gov-lifecycle.html

[44] IBM, “Integrating with watsonx.governance,” IBM OpenPages 9.2.x documentation, accessed Sept. 16, 2026. https://www.ibm.com/docs/en/openpages/9.2.x?topic=governance-integrating-watsonxgovernance

[45] IBM, “IBM watsonx.governance,” accessed Sept. 20, 2026. https://www.ibm.com/products/watsonx-governance

[46] OneTrust, “Ship AI Faster With Risk Control Where AI Runs,” accessed Sept. 20, 2026. https://www.onetrust.com/solutions/ai-governance/

Declaration on the Use of AI Tools. AI language models were used as research and writing support during preparation of this paper. The research direction, core ideas, conceptual constructs, and substantive argumentative choices were defined and repeatedly steered in detail by the author. OpenAI GPT-5.6 supported literature-oriented analysis, drafting, critical examination, iterative refinement, and LaTeX/PDF artifact generation. Claude, Gemini, Grok, and Perplexity were used as independent review systems for critical manuscript assessment, literature checking, and identification of conceptual or methodological weaknesses. Their feedback was evaluated by the author and selectively incorporated. The author remains solely responsible for the manuscript, its claims, references, and conclusions.