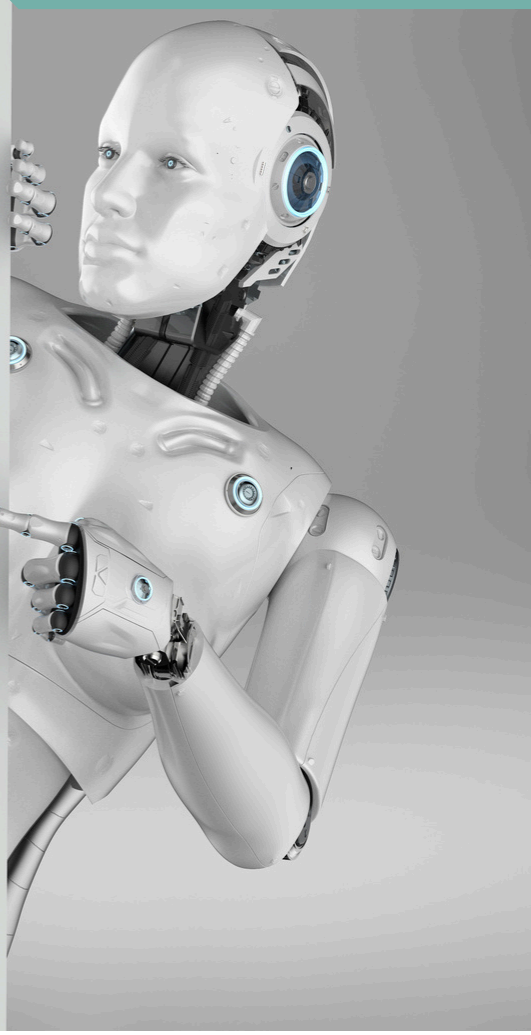
WHITEPAPER

RDG-AX™

A Governance Architecture for the Evaluation & Certification of Agentic AI Systems based on XRSI RDG™



CONCEPTUAL FRAMEWORK OVERVIEW: VERSION 1.0

BY KAVYA PEARLMAN, XRSI

Abstract

Agentic AI systems introduce forms of operational autonomy that extend beyond conventional model-centric artificial intelligence. These systems invoke tools, maintain memory, execute multi-step workflows, and adapt over time within dynamic environments. While such capabilities enable enterprise transformation, they simultaneously create new governance, reproducibility, and control challenges that are not adequately addressed by existing benchmarking or security-only approaches.

RDG-AX™ establishes a structured governance and evaluation architecture for Agentic AI systems. Built upon the RDG™ data lifecycle governance standard, the framework introduces a stage-gated evaluation process and six behavioral domains that collectively assess autonomy, stability, controllability, regulatory alignment, societal impact, and internal model integrity. Certification is issued by Cautelare following structured evidence review and controlled sandbox analysis.
This document presents the architectural structure and methodological foundations of RDG-AX™ without disclosing proprietary scoring algorithms or evaluation heuristics.

Executive Summary

Agentic AI systems operate with autonomy, tool execution, and adaptive behavior that traditional model-centric evaluation frameworks cannot adequately assess. RDG-AX™ introduces a layered governance and certification architecture specifically designed for enterprises deploying autonomous AI systems.

- Structured Autonomy Governance: RDG-AX™ evaluates agents across six behavioral domains (Capability, Reliability, Controllability, Compliance, Impact, and Model Integrity), ensuring autonomy remains bounded, observable, and enforceable across the system lifecycle.
- Stage-Gated Certification: Five structured evaluation gates produce reproducible, auditable evidence of controlled autonomy, with runtime action security assessment validating agent behavior within defined policy boundaries.
- Enterprise-Ready and Vendor-Neutral: Independently evaluated by Cautelare, RDG-AX™ operates across cloud platforms and orchestration stacks, signaling governance maturity with documented autonomy discipline.

Trust in agentic AI is not assumed. It is earned through structured evidence, independent evaluation, and verifiable governance. RDG-AX™ provides the architecture to make that trust possible.

1. Introduction

The evolution from static predictive models to agentic systems marks a structural shift in artificial intelligence deployment. Traditional evaluation paradigms focus on model accuracy, robustness to adversarial inputs, or isolated task performance. Agentic systems, however, exhibit compound behaviors that unfold over time and interact with external systems.

These systems may plan, reason, retrieve, execute tool calls, and update internal state. Their behavior cannot be reduced to single-turn inference quality. Consequently, evaluation must shift from model benchmarking toward autonomy governance.

RDG-AX™ addresses this need by formalizing a governance-aligned evaluation architecture designed specifically for agentic systems operating in enterprise and regulated environments.

2. Governance Foundation

RDG-AX™ is structurally anchored in the RDG™ data lifecycle governance standard. Before behavioral evaluation is conducted, each agent must be mapped to a governance backbone that establishes traceability and accountability across the data lifecycle.

This governance foundation requires formal articulation of data provenance, transformation pathways, storage boundaries, access controls, retention logic, role accountability, incident response structures, and change management protocols. The purpose of this backbone is not to certify regulatory compliance in isolation, but to ensure that autonomy operates within defined structural boundaries.

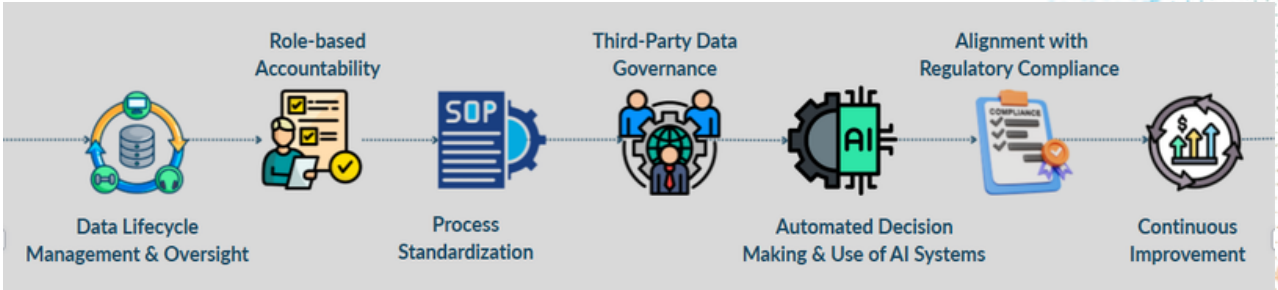
Evaluation without governance preconditions produces incomplete risk visibility. Accordingly, RDG-AX™ requires governance mapping prior to sandbox assessment.

3. Architectural Structure of RDG-AX™

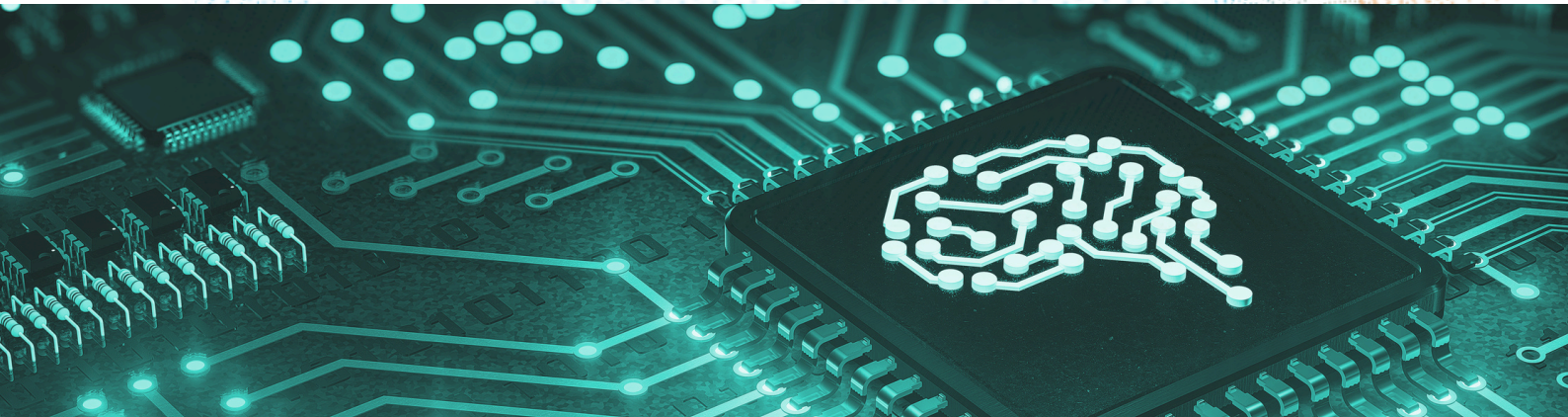
RDG-AX™ operates across three interrelated layers.

The first layer consists of the RDG™ governance backbone, which defines structural data controls and organizational accountability. The second layer introduces behavioral evaluation domains through which agent performance and autonomy characteristics are assessed. The third layer formalizes certification logic and trust signaling.

This layered architecture ensures that evaluation is not limited to task success or system uptime, but extends to autonomy boundaries, societal impact, and internal representational stability.



Ref: XRSI Responsible Data Governance Framework (XRSI RDG™)



4. Stage-Gated Evaluation Model

Evaluation under RDG-AX™ proceeds through a sequential stage-gated process designed to maintain structural rigor and reproducibility.

Gate 1: Intake & Scoping: The initial gate establishes intake and scoping. During this phase, the intended use, declared autonomy level, operational environment, data boundaries, tool interfaces, and regional deployment context are documented. Risk classification is performed at this stage.

The declared autonomy tier informs the depth and nature of sandbox evaluation, including the applicability of runtime action security approaches such as those defined in the AARM (Autonomous Action Runtime Management) Specification v1.0, particularly for systems with external action capabilities and multi-step execution behavior.

Gate 2: Governance Alignment: The second gate evaluates governance alignment. Data lifecycle discipline, lawful basis, minimization constraints, retention boundaries, and accessibility considerations are reviewed. Identified gaps are documented for remediation.

Gate 3: Sandbox Analysis: The third gate introduces controlled sandbox analysis. The agent is evaluated in a deterministic environment using structured scenarios. Telemetry capture is enforced. Behavioral observations are documented across defined evaluation domains. Sandbox evaluation incorporates runtime action security assessment for systems with action-taking capabilities, using approaches aligned with the AARM Specification v1.0, which treats the action layer as a primary control boundary for evaluating agent behavior.

Gate 4: Deployment Readiness: The fourth gate confirms deployment readiness. Remediation is validated, autonomy tiers are confirmed, and control mechanisms such as rollback and human intervention pathways are verified. Evidence generated through runtime action security evaluation, including action mediation, policy enforcement, and audit traceability, is included as part of deployment readiness validation for systems operating beyond assistive autonomy.

Gate 5: Post-Deployment Monitoring: The fifth gate establishes post-deployment monitoring requirements. Drift detection, change documentation, and recertification cadence are defined. Post-deployment monitoring extends to runtime action visibility, including traceability of agent-initiated actions, context continuity, and alignment between system behavior and declared intent, consistent with principles defined in the AARM Specification v1.0.

This stage-gated structure ensures that certification reflects structured progression rather than isolated testing events. The integration of AARM-aligned evaluation within sandbox analysis ensures that runtime action execution is assessed as a first-class governance dimension, reflecting the shift from model-centric evaluation to action-centric risk validation.

5. Behavioral Evaluation Domains

RDG-AX™ evaluates agentic systems across **six domains** that collectively define responsible autonomy.

1. **Capability** assesses whether the agent achieves its declared objectives across representative tasks while maintaining performance stability across adjacent scenarios.

2. **Reliability** evaluates robustness under environmental perturbations, tool variability, and operational stress conditions. Deterministic behavior and graceful degradation are central considerations.

3. **Controllability** examines whether human supervision and intervention rights remain structurally preserved. This includes autonomy classification discipline, boundary enforcement, and auditability of agent decisions.

4. **Compliance** evaluates adherence to defined data lifecycle controls and jurisdictional constraints. The framework does not replace regulatory compliance programs but verifies that autonomy operates within governance-defined limits.

5. **Impact** examines accessibility, fairness, transparency, and organizational incentive alignment. The objective is to assess whether deployment conditions promote responsible human outcomes.

6. **Model Integrity** evaluates representational stability, simulation coherence, boundary enforcement, and adaptive consistency across internal predictive or learned state architectures. This includes systems employing world models, reflective planning loops, latent embedding prediction, or self-supervised learning mechanisms. The domain addresses governance risks introduced by experience-based learning, internal state evolution, and bounded autonomy over time.

The methodological details of scoring and weighting across these domains remain proprietary.

6. Sandbox Evaluation Methodology

Sandbox analysis under RDG-AX™ is designed to produce reproducible and auditable evidence. The environment is controlled to ensure deterministic starting states and structured scenario execution. Telemetry capture is enforced to preserve traceability of tool invocations, memory interactions, and behavioral transitions.

Evaluation emphasizes behavioral integrity rather than isolated output correctness. Where applicable, human-in-the-loop checkpoints are incorporated to verify escalation and override pathways.

Exact evaluation matrices and stress-test libraries are maintained by Cautelare to preserve certification integrity. For systems with external action capabilities, sandbox evaluation includes runtime action security validation aligned with the AARM Specification v1.0.

Evaluation scenarios include:

- Verification of action interception prior to execution, ensuring the presence of a mediation boundary
- Assessment of context accumulation across multi-step execution chains
- Evaluation of policy-aligned decision-making under varying contextual conditions
- Validation of enforcement outcomes, including allow, deny, modification, deferral, and escalation pathways
- Verification of tamper-evident logging and traceability of actions and decisions

These evaluation elements align with the AARM model, which defines runtime security as the interception, contextual evaluation, enforcement, and recording of AI-driven actions at the point of execution.

Within RDG-AX™, this evaluation extends sandbox analysis to the action layer, enabling structured assessment of how agentic systems translate reasoning into real-world effects under governed conditions.

7. Certification Logic

Upon completion of the stage-gated evaluation process, Cautelare issues a structured trust signal. The certification reflects maturity within the declared deployment scope and is supported by documented evidence and registry entry. The trust signal is categorical rather than continuous and is accompanied by an executive summary describing intended use, autonomy classification, and identified mitigation conditions. Machine-readable artifacts may be generated for enterprise integration.

Certification does not constitute legal compliance guarantee but signals structured governance maturity and controlled autonomy.

8. Enterprise Integration Considerations

RDG-AX™ may be deployed as a pre-onboarding stage gate, a release authorization checkpoint, a Center of Excellence evaluation function, or a recurring assurance program.

The framework is vendor-neutral and architecture-agnostic. It is designed to operate independently of specific model providers, orchestration stacks, or cloud platforms.

Evaluation is conducted independently by Cautelare to maintain certification integrity.

9. Strategic Context & Conclusion

As agentic AI systems evolve toward self-directed learning, world-model reasoning, and multi-agent coordination, evaluation paradigms must evolve correspondingly. Reward-based testing and static benchmarking are insufficient for systems that learn through accumulated experience, internal representation refinement, and real-time interaction with external systems. This shift requires evaluation approaches that extend beyond model outputs to include the governance of actions at execution time, as reflected in emerging runtime security specifications such as AARM, which define the action layer as a critical control boundary.

RDG-AX™ provides a structural governance architecture aligned with this trajectory. By integrating governance mapping, behavioral evaluation, and certification signaling, the framework enables enterprises to deploy agentic systems with documented autonomy discipline, while incorporating evaluation approaches that assess both decision-making processes and their real-world effects.

Agentic AI systems require structured autonomy governance beyond traditional model evaluation. RDG-AX™ establishes a layered, stage-gated architecture for evaluating and certifying agentic systems operating in enterprise environments, ensuring that autonomy is bounded, observable, and enforceable across the system lifecycle.

By combining governance backbone discipline with behavioral domain analysis, runtime action-aware evaluation perspectives, and independent certification, RDG-AX™ operationalizes responsible autonomy without exposing proprietary evaluation mechanics.

References

X Reality Safety Intelligence. (2025). The responsible data governance standard (2025v1). XRSI. https://xrsi.org/wp-content/uploads/2026/02/XRSI-RDG-Core-Standard-2025v1.pdf

Errico, H. (2025). Autonomous Action Runtime Management (AARM): A system specification for securing AI-driven actions at runtime. arXiv. https://arxiv.org/abs/2602.09433

DISCLAIMER: THE FOREGOING GUIDANCE IS PROVIDED "AS IS" WITHOUT ANY EXPRESS OR IMPLIED WARRANTY OF ANY KIND INCLUDING WARRANTIES OF MERCHANTABILITY, NON INFRINGEMENT OF INTELLECTUAL PROPERTY, OR FITNESS FOR ANY PARTICULAR PURPOSE. XRSI CONTINUES TO INVESTIGATE THESE AND OTHER TECHNIQUES AND MAY MODIFY OR UPDATE THE INFORMATION HEREIN WITHOUT NOTICE.

XRSI™

Human Intelligence In The Loop

www.xrsi.org/rdg | info@xrsi.org