OISG: Open Intelligent Secure Governed — A Unifying Paradigm for Autonomous AI Systems

Stefano Noferi noze, Pisa, Italy

Version 1.0 — April 2026

Abstract

The rapid deployment of autonomous AI agents in production environments has exposed a structural gap: the frameworks governing openness, intelligence capabilities, security, and compliance operate in silos, each addressing a single dimension of a fundamentally multi-dimensional problem. This paper introduces **OISG** (Open Intelligent Secure Governed) as a unifying paradigm that formalises the interdependence between these four dimensions and provides a conceptual framework for evaluating, designing, and operating AI systems that satisfy all four simultaneously. OISG does not replace existing standards — EU AI Act, NIST AI RMF, ISO/IEC 42001, OWASP Top 10 for Agentic Applications — but establishes a meta-framework that maps their intersections, exposes coverage gaps, and defines a feedback loop where each dimension constrains and enables the others. We ground the paradigm in the current regulatory timeline, the emergent properties of multi-agent systems, and the operational requirements of sovereign AI infrastructure.

Keywords: AI governance, autonomous agents, open source, AI security, EU AI Act, NIST AI RMF, agentic AI, sovereign infrastructure

1. Introduction

1.1 The fragmentation problem

Organisations deploying AI systems in 2026 face an unprecedented convergence of pressures: regulatory obligations (EU AI Act high-risk provisions effective August 2026, NIS2 already operational), security threats specific to autonomous agents (OWASP Top 10 for Agentic

Applications, published December 2025), demands for transparency and auditability (ISO/IEC 42001 certification becoming a procurement requirement), and the operational need for capable, reliable intelligence that respects data sovereignty constraints.

Each of these pressures has produced its own body of standards, toolkits, and best practices. The open source community has developed licensing models, foundation governance (CNCF, LF AI & Data), and interoperability protocols (MCP, OpenTelemetry). The AI research community has produced evaluation frameworks (HELM, Model Cards, benchmark suites). The security community has mapped threat surfaces specific to LLM-based systems and agentic architectures. Regulators have codified compliance requirements into law.

What is missing is a *unifying abstraction* that treats these four dimensions as a single system with feedback loops, rather than as independent checklists to be satisfied in parallel. This absence creates practical problems: security controls that undermine transparency, governance frameworks that ignore capability measurement, open source practices that do not extend to model provenance, and compliance processes disconnected from runtime behaviour.

1.2 Contribution

This paper proposes OISG (Open Intelligent Secure Governed) as a paradigm — not a product, consortium, or specification — that:

1. Defines four interdependent pillars (Open, Intelligent, Secure, Governed) with precise scope and measurable properties.
2. Formalises the feedback relationships between pillars as a directed cycle where each dimension constrains and enables the next.
3. Maps existing standards, regulations, and toolkits onto the OISG structure, identifying coverage and gaps.
4. Provides a decision framework (the OISG Adequacy Test) for evaluating whether a given AI system satisfies all four dimensions simultaneously.

The naming is deliberate: like REST (Representational State Transfer) codified architectural practices that already existed in the early web, OISG names a convergence that is already occurring in practice but lacks a shared vocabulary.

---

2. The Four Pillars

2.1 Open

Definition. An AI system satisfies the Open requirement when its components — models, training methodology, governance infrastructure, communication protocols, and audit logs — are inspectable, reproducible, and interoperable by independent parties.

Scope. Openness in OISG is broader than open source licensing. It encompasses:

- **Model transparency.** Not necessarily open weights, but documented capabilities, limitations, and provenance. The EU AI Act requires transparency obligations for general-purpose AI models (GPAI), including technical documentation, training methodology summaries, and copyright compliance information.
- **Governance infrastructure.** The systems that enforce policy on AI behaviour must themselves be auditable. A proprietary policy engine governing an open model creates a trust displacement, not a trust resolution. Frameworks such as Admina (Apache 2.0) and the Agent Governance Toolkit (MIT) demonstrate that governance infrastructure can be fully open.
- **Protocol interoperability.** Agent-to-agent and agent-to-infrastructure communication must use open standards (Model Context Protocol, Agent-to-Agent Protocol (A2A), OpenTelemetry for AI observability) rather than vendor-locked APIs.
- **Community stewardship.** Code availability is necessary but not sufficient. Sustainable openness requires governance of the open project itself: foundation membership, transparent roadmaps, contribution processes, security disclosure policies.

Adequacy metric. What fraction of the AI system's decision-affecting components can be audited by an independent third party without requiring proprietary access?

2.2 Intelligent

Definition. An AI system satisfies the Intelligent requirement when its capabilities are measured, documented, bounded, and aligned with explicitly stated objectives.

Scope. Intelligent in OISG is not about maximising capability but about *governing* it:

- **Measurable capabilities.** Every model in production must have a capability profile: benchmark results, known failure modes, confidence calibration, domain-specific evaluation suites. Article 13 of the EU AI Act mandates that high-risk AI systems provide sufficient transparency for users to interpret and use output appropriately.

- **Sovereign infrastructure.** The ability to execute models in controlled environments — on-premise, private cloud, air-gapped — is a prerequisite for data sovereignty, particularly in healthcare, public administration, and defence. This requires runtime infrastructure (inference engines such as Ollama, vLLM; vector databases; embedding pipelines) that operates independently of third-party cloud APIs.
- **Traceable retrieval.** Retrieval-Augmented Generation (RAG) pipelines must be auditable: which document, at which version, with which embedding model, informed which response. Without retrieval traceability, the "intelligence" of the system is a black box within a black box.
- **Bounded autonomy.** Autonomous agents must operate within an explicit autonomy taxonomy: task execution (API calls within defined scope), choice (selection among pre-approved alternatives), commitment (actions with organisational impact requiring human approval), self-modification (always requiring human authorisation). This taxonomy must be machine-readable and enforceable at runtime. This four-level taxonomy is proposed as part of the OISG framework.

Adequacy metric. Can the system produce, on demand, a complete explanation of why it gave a specific response — including the data sources consulted, the model version used, and the confidence level — within a defined latency budget?

2.3 Secure

Definition. An AI system satisfies the Secure requirement when it is resilient to adversarial manipulation across all interaction surfaces, at runtime, with measurable detection and response latencies.

Scope. Security for autonomous AI agents differs qualitatively from traditional application security:

- **Bidirectional injection defence.** Prompt injection is not limited to user-to-model attacks. Indirect injection via compromised retrieval documents, tool outputs, or inter-agent messages represents an equally critical attack surface. Defence must operate on both request and response paths, for every interaction. Pattern-based detection should operate at microsecond latency to avoid becoming a bottleneck.
- **Cryptographic agent identity.** In multi-agent systems, every agent must possess a verifiable identity (Decentralised Identifiers, Ed25519 key pairs). Inter-agent trust must be quantifiable and dynamic, not assumed. Protocols such as the Inter-Agent Trust Protocol (IATP), implemented within the Agent Governance Toolkit's AgentMesh module, demonstrate how inter-agent trust can be formalised through cryptographic identity verification and dynamic trust scoring.

- **Transactional kill switch.** Emergency termination of an autonomous agent must preserve forensic state, enable transactional rollback, and guarantee system consistency. "Turning it off" is not a kill switch; it is an uncontrolled halt. A kill switch is an architectural component with defined pre-conditions, state preservation guarantees, and recovery procedures.
- **Model supply chain integrity.** Model fingerprinting — identifying a model through its observable behaviour rather than its declared identity — is necessary to verify that the model executing in production is the model that was evaluated and approved. Software supply chain practices (SLSA, SBOM, cryptographic provenance) must extend to model weights, adapter layers, and fine-tuning datasets.
- **Data boundary enforcement.** Personally identifiable information (PII) must be redacted before reaching any model endpoint not explicitly authorised for PII processing. Redaction must be performed at the infrastructure level, not delegated to the model itself, and must operate at latencies that do not degrade user experience.

Adequacy metric. If an agent is compromised at 03:00, what is the mean time to detection, mean time to containment, and mean time to forensic-quality state recovery?

2.4 Governed

Definition. An AI system satisfies the Governed requirement when its compliance with applicable regulations, organisational policies, and ethical constraints is verified automatically, continuously, and with immutable evidence.

Scope. Governed in OISG is an operational discipline, not a periodic audit:

- **Runtime compliance.** EU AI Act (Articles 6–15), NIS2, GDPR, ISO/IEC 42001 requirements must be enforced at the system level, not satisfied through annual reviews. Every decision by an autonomous agent must be automatically classified against the applicable risk tier and logged with the corresponding compliance evidence.
- **Forensic black box.** By analogy with aviation flight data recorders, every high-risk AI system must maintain an immutable, hash-chained (SHA-256) log of all interactions, decisions, human interventions, and system state changes. This log must be tamper-evident, retention-policy-compliant, and retrievable within defined SLAs.
- **Proportional risk classification.** Not all AI systems carry equal risk. A FAQ chatbot is not a credit scoring system. Governance controls must be proportional to the classified risk level, and classification must be automated, auditable, and updatable as system capabilities evolve.
- **Architectural human-in-the-loop.** Human oversight must be defined as an architectural component, not a policy aspiration. This means specifying *which decisions* require human review, *what information* the reviewer receives, *what response time* is required, and *what*

happens if the human does not respond. Absent these specifications, "human-in-the-loop" is a compliance fiction.

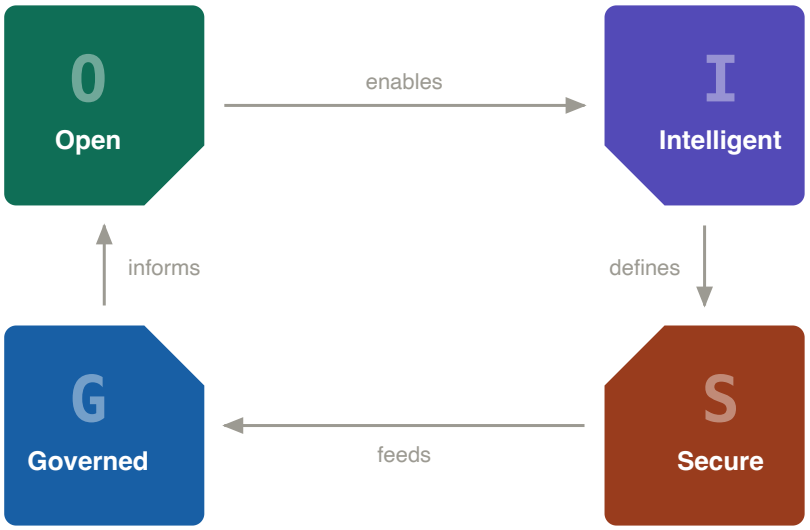
- **End-to-end observability.** Distributed tracing (OpenTelemetry), governance dashboards, SLOs, error budgets, and circuit breakers — operational reliability practices established in cloud-native engineering — must extend to AI systems. AI agents are services; they deserve the same observability rigour as any production microservice.

Adequacy metric. If a supervisory authority requests evidence of compliance for a specific AI system, how many hours does it take to produce the required documentation?

---

3. The Feedback Cycle

The four pillars of OISG are not independent variables. They form a directed cycle with the following relationships:



Open enables Intelligent. Inspectable models and open data pipelines permit capability measurement, independent evaluation, and continuous improvement. Without openness, intelligence is asserted, not verified.

Intelligent defines Secure requirements. The capabilities and autonomy level of the AI system determine its attack surface. A more capable agent with broader tool access presents a larger threat model. Security controls must scale with intelligence.

Secure feeds Governed. Security controls produce the telemetry — logs, traces, alerts, forensic snapshots — on which governance operates. Without security instrumentation, governance has no data to govern.

Governed informs Open. Governance policies determine what can be open, to whom, under what constraints. Data classification, export controls, and intellectual property policies shape the boundaries of openness.

Breaking the cycle produces pathology. Governance without openness produces institutional opacity. Intelligence without security is negligence. Security without governance is unaccountable control. Openness without governance is uncontrolled exposure. The four pillars must operate as a system.

---

4. Mapping Existing Frameworks

Pillar	Standards / Frameworks	Coverage	Gap (addressed by OISG integration)
O	Apache 2.0, OSI Definition, CNCF/LF AI governance, MCP, A2A	Software licensing, project governance, protocol specs	Extension to model provenance, governance infrastructure auditing, embedded transparency for GPAI
I	NIST AI RMF, HELM, Model Cards, ISO/IEC 42001 Annex A	Model evaluation, risk categorisation, management systems	Integration with runtime security telemetry, sovereign infra requirements, RAG traceability
S	OWASP Top 10 Agentic AI 2026, CoSAI, MITRE ATLAS, NIS2	Threat taxonomies, vulnerability databases, network security	Agent identity, transactional kill switch, model supply chain, bidirectional injection defence
G	EU AI Act, GDPR, ISO/IEC 42001, NIST AI RMF (Govern function)	Regulatory requirements, management systems, risk frameworks	Runtime compliance automation, forensic black box, architectural HITL, proportional automation

The table reveals a pattern: each existing framework covers one pillar deeply but references the others only peripherally. NIST AI RMF includes a "Govern" function but does not specify runtime compliance mechanisms. The EU AI Act mandates transparency but does not prescribe open source practices. OWASP maps agentic threats but does not integrate governance telemetry. OISG provides the connective tissue.

---

5. The OISG Adequacy Test

For any AI system under evaluation, the following test determines OISG adequacy. This paper proposes a quantitative scoring system: each pillar is assessed against five criteria, scored 0–5. The total score ranges from 0 to 100.

O — Is it Open?

1. Model documentation (capabilities, limitations, provenance) is available to independent auditors.
2. Governance infrastructure (policy engines, decision logic) is open and auditable.
3. Communication protocols use open standards (MCP, OpenTelemetry, A2A).
4. Open projects have community stewardship (contribution process, security disclosure, governance).
5. Model provenance and training methodology are documented and reproducible.

I — Is it Intelligent (governably)?

1. Model capabilities are measured with benchmark results, known failure modes, and confidence calibration.
2. Infrastructure supports sovereign execution (on-premise, private cloud, air-gapped) where required.
3. RAG pipelines are traceable (document version, embedding model, retrieval path).
4. Agent autonomy scope is explicit, machine-readable, and enforced at runtime.
5. System can produce on demand a complete explanation of why it gave a specific response.

S — Is it Secure?

1. Bidirectional injection defence operates on both request and response paths.
2. Agent identities are cryptographically verifiable (DIDs, Ed25519 key pairs).
3. Transactional kill switch preserves forensic state and enables rollback.
4. PII redaction is enforced at infrastructure level before model endpoints.
5. Model supply chain integrity is verified (fingerprinting, SBOM, cryptographic provenance).

G — Is it Governed?

1. Compliance is verified automatically at runtime, not through periodic audits.
2. Immutable forensic log (hash-chained) records all interactions and decisions.
3. Human oversight is architecturally defined (which decisions, what info, what timeout).
4. End-to-end observability is in place (distributed tracing, SLOs, dashboards).

5. Risk classification is proportional, automated, and auditable as capabilities evolve.

Score levels: 0–24 Critical gaps, 25–49 Partial coverage, 50–79 Good coverage, 80–100 OISG adequate. A gap may be acceptable depending on the system's risk classification, but it must be *conscious and documented* — not an oversight. The full interactive test is available at oisg.ai/test.

---

6. Implementation Considerations

6.1 Incremental adoption

OISG does not require simultaneous implementation of all four pillars at maximum depth. A pragmatic adoption path:

1. **Assess.** Map the existing AI system against the four pillars. Identify which pillar has the largest gap relative to the system's risk classification.
2. **Instrument.** Begin with observability (Governed pillar) — it produces the data needed to evaluate gaps in the other three pillars.
3. **Secure.** Add runtime security controls (anti-injection, PII redaction, agent identity) — these are the most operationally urgent and produce the telemetry that governance consumes.
4. **Open.** Ensure governance infrastructure is auditable and model documentation is current — this enables independent verification of the other three pillars.
5. **Iterate.** The feedback cycle means improvements in one pillar reveal requirements in the next. This is by design, not by defect.

6.2 Tooling landscape

The OISG paradigm is tool-agnostic but benefits from integrated implementations that cover multiple pillars within a single operational surface. The open source ecosystem is converging toward this integration: governance frameworks that embed security controls (policy engines with anti-injection), security toolkits that produce governance-grade telemetry (OWASP-aligned logging with compliance mapping), and infrastructure platforms that combine model serving with data sovereignty enforcement.

6.3 Regulatory alignment

OISG maps directly to the regulatory timeline:

- **Already active:** NIS2 (network and information security for essential entities), GDPR (data protection, including AI-specific implications under EDPB guidance).
- **August 2025:** EU AI Act obligations for GPAI models (transparency, copyright compliance, systemic risk assessment).
- **August 2026:** EU AI Act obligations for high-risk AI systems (conformity assessment, post-market monitoring, fundamental rights impact assessment).
- **Ongoing:** ISO/IEC 42001 certification cycles, NIST AI RMF voluntary adoption, sector-specific guidance (EBA for financial services, MDR for medical devices).

OISG does not add regulatory burden. It provides a structure for addressing multiple regulatory requirements through a unified operational framework rather than through parallel, disconnected compliance programmes.

---

7. Limitations and Future Work

OISG is a paradigm, not a specification. It intentionally does not prescribe implementation details, specific technologies, or quantitative thresholds for adequacy metrics. These decisions are context-dependent and should be made by the implementing organisation based on its risk profile, regulatory obligations, and operational constraints.

Future work should address:

- **Refinement of the scoring model** proposed in Section 5, including weighted criteria, sector-specific thresholds, and automated assessment tooling.
- **Sector-specific instantiations** of OISG for healthcare (MDR, clinical AI), financial services (DORA, EBA guidelines), and public administration (national AI strategies, digital sovereignty requirements).
- **Formal verification** of the feedback cycle properties, particularly the conditions under which breaking the cycle produces measurable degradation in system trustworthiness.
- **Benchmarking suites** that test all four pillars simultaneously, rather than in isolation.

---

8. Conclusion

The AI systems of 2026 are not the AI systems of 2020. They act autonomously, coordinate with peers, access tools, and make decisions with organisational impact. The governance, security, and transparency frameworks developed for previous generations of AI are necessary but not sufficient.

OISG — Open Intelligent Secure Governed — provides a unifying paradigm for this new reality. It does not invent new requirements; it formalises the interdependence of existing ones. The four pillars and their feedback cycle offer a shared vocabulary for technologists, regulators, security professionals, and open source communities to reason about AI systems as integrated wholes rather than as collections of independent concerns.

Disclosure

The author is the creator and maintainer of Admina, an open source governed AI development framework cited in this paper as an example of auditable governance infrastructure (Section 2.1). Admina is referenced as an illustrative implementation of the patterns described by the OISG paradigm and is planned to incorporate OISG adequacy metrics in future releases. Admina is independently available under the Apache 2.0 licence and is not a commercial product. The inclusion of Admina in this paper reflects its relevance as an open source reference implementation, not an endorsement of a specific toolchain. The OISG paradigm is tool-agnostic by design.

References

1. European Parliament and Council. *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act)*. Official Journal of the European Union, 2024.
2. OWASP. *Top 10 for Agentic Applications for 2026*. Open Worldwide Application Security Project, December 2025.
3. NIST. *AI Risk Management Framework (AI RMF 1.0)*. National Institute of Standards and Technology, January 2023.
4. ISO/IEC. *ISO/IEC 42001:2023 — Artificial intelligence — Management system*. International Organization for Standardization, 2023.

5. European Parliament and Council. *Directive (EU) 2022/2555 on measures for a high common level of cybersecurity across the Union (NIS2)*. Official Journal of the European Union, 2022.
6. Fielding, R.T. *Architectural Styles and the Design of Network-Based Software Architectures*. Doctoral dissertation, University of California, Irvine, 2000.
7. OECD. *Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449, revised May 2024)*. Organisation for Economic Co-operation and Development, 2019/2024.
8. Coalition for Secure AI (CoSAI). *Principles for Secure-by-Design Agentic Systems*. OASIS Open Project, 2025.
9. LF AI & Data Foundation. *Principles for Trusted AI*. Trusted AI Committee, Linux Foundation, February 2021.
10. UNESCO. *Recommendation on the Ethics of Artificial Intelligence*. United Nations Educational, Scientific and Cultural Organization, 2021.
11. Noferi, S. *Admina — Governed AI Development Framework*. Apache 2.0. https://admina.org, 2025–2026.
12. Microsoft. *Agent Governance Toolkit — Policy Enforcement, Zero-Trust Identity, Execution Sandboxing, and Reliability Engineering for Autonomous AI Agents*. MIT License. https://github.com/microsoft/agent-governance-toolkit, April 2026.
13. Google. *Agent-to-Agent Protocol (A2A)*. https://github.com/google/A2A, April 2025.

Licence

DOI: 10.5281/zenodo.19605659 Preprint: oisg.ai/paper/