



EXECUTIVE PERSPECTIVE

SECURE AI: CONTROL BEFORE AUTONOMY

Risks, Guardrails and Evidence to Protect
Data, Identities and Processes

AUTHORS



Antonio Autorino

Head of AI & Data Sense, **DGS**



Gianluca Cimino

Chief Strategy & Delivery Officer Cybersecurity, **DGS**



Irene Cosimo

Content Marketing Leader, **DGS**



Antonio Di Ronza

Head of Cybersecurity Advisory, **DGS**



Christian Scquizzato

AI & Data Sense Presales Manager, **DGS**

SECURE AI: CONTROL BEFORE AUTONOMY

Risks, Guardrails and Evidence to Protect Data,
Identities and Processes

CONTENTS

Foreword	4
Executive Summary	5
PART I — RISK: A NEW ATTACK SURFACE AND OPERATIONAL CONTROL	
01 Cyber + AI: A Unified Risk Perimeter	8
02 Threat Landscape: Defence Priorities and Accelerating Risk	10
03 AI-Augmented SOC: Autonomy, Guardrails and Control	15
04 GenAI, RAG and Agents: Protecting Data, Identities and Actions	18
05 Secure AI Lifecycle: Gates, Evidence and Residual Risk	23
06 Autonomy and Accountability: The DGS Manifesto	25
PART II — THE DGS RESPONSE: FROM RISK TO OPERATIONAL OVERSIGHT	
07 The Foundations of the DGS Approach	26
08 From Risk Assessment to Continuous Control	29
09 Secure AI Capabilities: Control by Design and Validation	32
10 Client Roadmap: Visibility, Readiness and Controlled Scale-Up	35
CONCLUSIONS	36

FOREWORD

| The New Responsibility of Control

By Salvatore Frosina, Co-CEO DGS

For years, we have protected systems designed to execute instructions. Today, we must govern systems capable of interpreting, generating and turning instructions into actions.

AI accesses data, uses identities and connectors, supports decision-making, and interacts with real-world applications and processes. Risk therefore concerns not only what AI produces, but, above all, what it can access, authorise and execute.

For security leaders, the challenge is to enable AI to create value while retaining control over its actions and their potential consequences. This requires clearly defined boundaries, explicit accountability, verifiable evidence, and the ability to stop or reverse an operation.

This paper brings together DGS's perspective and experience in integrating Cybersecurity, Data & AI, Cloud, identity and operational oversight to support organisations in the secure adoption of AI.

We can entrust AI with an increasing share of our activities. Responsibility for what happens remains in our hands.



Salvatore Frosina
Co-CEO DGS

EXECUTIVE SUMMARY

The Central Argument

The adoption of generative and agentic systems is reshaping the enterprise risk profile. It increases the scale of certain known threats, introduces new assets that require protection, and enables AI systems to interact directly with corporate data, identities and applications. The traditional attack surface, comprising networks, endpoints, infrastructure, applications and identities, now extends to **models, prompts, datasets, vector databases, knowledge bases, connectors, agent memory and actions performed on real-world systems.**

The priority is to retain control as AI becomes embedded in business processes. Security, ownership and evidence must therefore be defined before deployment and maintained throughout the entire lifecycle.

The same agenda must address two inseparable priorities: using AI to enhance prevention, detection and response, and protecting the AI systems adopted by the enterprise. Data, identities, telemetry, incident management and accountability are shared across both areas. Separate governance creates gaps and slows containment.

The most immediate risk stems from the growing scale of phishing, impersonation and reconnaissance activities.

RAG systems and agents expand the risk perimeter through indirect prompt injection, data poisoning, model extraction, tool misuse, excessive privileges and the automated propagation of errors. **Investment priorities should be determined by exposure, impact, detectability and reversibility.**

The **AI Act, GDPR, NIS2** and **DORA** require safeguards aligned with their respective scopes of application. For the CISO, these requirements converge at the operational level: **formalised controls, clearly assigned accountability, explicit residual risk and verifiable evidence.**

Seven Priorities for the CISO: Control, Accountability and Residual Risk

01 Govern a Unified Risk Perimeter

Cybersecurity and AI must share a common governance framework, controls and evidence.

02 Maintain Oversight of High-Impact Decisions

AI can support triage, enrichment, analysis and reporting, while authorisation and accountability remain assigned to clearly identified roles.

03 Extend Control from Outputs to Actions

Identities, privileges, connectors and tools determine which actions an AI system can perform across business processes.

04 Embed Security Throughout the Lifecycle

Threat modelling, data provenance, adversarial testing, release gates, runtime monitoring and incident management must be in place before go-live and remain active throughout operations.

05 Treat Agents and Connectors as Operational Identities

Least privilege, temporary credentials, segregation of duties, approval and revocation must also apply to non-human identities.

06 Grant Autonomy Incrementally

Automation should increase only when telemetry, playbooks, observability, rollback mechanisms and kill switches are operational and have been validated.

07 Measure Effectiveness and Residual Risk

KPIs, KRIs, overrides, blocked actions, scope violations and rollback tests must inform release and review decisions.

| Priorities for the Next Twelve Months

01 | PROTECT NOW

SCOPE

GenAI and RAG systems already in use; identities, data and connectors; shadow AI; fraud prevention and incident management processes.

ACTION

Inventory, risk classification, access control, provenance, logging and playbooks.

02 | BUILD NOW

SCOPE

Assessment methodologies, a skills map, two accelerators, AI-assisted SOC pilots and an evidence pack.

ACTION

Owners, backlog, seed funding, pilot customers and stage gates.

03 | AVOID PREMATURE COMMITMENTS

SCOPE

A fully autonomous SOC, mature Managed AI Security services, and proprietary AI-native platforms without supporting evidence.

ACTION

Present the roadmap, limitations and next steps, backed by evidence.

PART I

Risk: A New Attack Surface and Operational Control

CYBER + AI: A UNIFIED RISK PERIMETER

D1

AI brings multiple risk profiles together within a single system: it enhances defensive capabilities, increases the effectiveness of certain offensive techniques, and introduces new assets with access to data, identities and applications. With AI agents, risk extends beyond outputs to actions.

With agents, risk extends from output to action.

This requires a unified governance model spanning **application security, data governance, IAM, SOC and compliance.**

AI FOR CYBERSECURITY

- 1 | Triage and Prioritisation
- 2 | Enrichment
- 3 | Threat intelligence
- 4 | Detection engineering
- 5 | Vulnerability prioritization

CYBERSECURITY FOR AI

- 1 | Prompt and model security
- 2 | RAG security
- 3 | Agent identity
- 4 | AI Infrastructure and Supply Chain
- 5 | Tool governance

OPERATIONAL CONVERGENCE

| A **unified model** for **people, processes, controls** and **evidence.**

| Four Dimensions of Cyber + AI Risk

01 AI as an Attack Tool

Generative models reduce the cost of producing credible content, increase attack volumes, and accelerate reconnaissance, personalisation and support for malicious coding.

02 AI as a Defence Tool

AI delivers increasing value through contextual correlation, alert prioritisation, incident enrichment, and support for detection, reporting and response.

03 AI as an Asset to Protect

Models, data, prompts, knowledge bases, vector databases, pipelines, connectors and dependencies become assets that must be inventoried, classified and monitored.

04 Agents as a Qualitatively Different Risk

Agents can use identities, memory, tools and enterprise systems. The level of risk depends on their privileges, autonomy, oversight and the reversibility of their actions.

Act now: Establish a cross-functional steering committee involving the **CISO, Data & AI Leadership, Enterprise Architecture, Risk and Compliance**. The first expected outcome is a **shared, up-to-date view** of use cases, assets, risks, owners and controls.

PART I

Risk: A New Attack Surface and Operational Control

THREAT LANDSCAPE

02

Threat Landscape: Defence Priorities and Accelerating Risk

The priority is not to pursue every emerging scenario, but to distinguish between the threats posing a risk today and those that require preparation.

AI industrialises established techniques while introducing **direct attacks** on **models**, **retrieval** systems and **agents**.

This distinction enables management to allocate budgets and controls based on evidence, exposure and potential impact.



| Indicative Maturity of Key Cyber + AI Threats



Defence Priorities: Protect, Operationalise, Experiment

01 | PROTECT NOW

SCOPE

Phishing, impersonation, credentials, shadow AI and information disclosure.

ACTION

MFA, identity controls, DLP, usage policies, logging and security awareness.

02 | OPERATIONALISE NOW

SCOPE

GenAI and RAG systems in production, supply chain risks and prompt injection.

ACTION

Inventory, threat modelling, provenance, gateways, testing and observability.

03 | EXPERIMENT SECURELY

SCOPE

Agents, tool use, multi-agent systems and autonomous remediation.

ACTION

Sandboxes, limited scope, approval workflows, temporary credentials, rollback and kill switches.

| Three Criteria for Risk Prioritisation

01 Exposure

The same technique can have different consequences depending on system access, data value, connector privileges and process dependency. Risk changes significantly between an isolated sandboxed agent and one connected to ERP systems and identity providers.

02 Propagation Speed

Priority increases when an error or compromise can spread rapidly across users, applications, agents and systems, resulting in technical, operational or financial impacts.

03 Detectability and Reversibility

The control threshold must increase when behaviour is difficult to observe, attribute, contain or reverse.

Operational Thresholds: Assist, Approve, Automate

MODE	ACCOUNTABILITY	EXAMPLES	PREREQUISITES
ASSIST	AI provides context and options; humans decide.	Triage, enrichment, queries, incident summaries and reports.	Data quality, source attribution and human validation.
PREPARE + APPROVE	The agent prepares the task; a human authorises the action.	Tickets, priority changes, containment and privilege revocation.	Dedicated identity, least privilege, approval workflow and audit trail.
AUTOMATE WITHIN SCOPE	The system acts within predefined, reversible boundaries.	Tagging, enrichment, isolated actions and low-impact remediation.	Defined thresholds, rollback, kill switch, logging and measurement.

Proof before autonomy: the Gates That Authorise Action

GATE	PASS CONDITION	MINIMUM EVIDENCE
TELEMETRY	Events, context and assets are available at a measured level of quality.	Coverage map and data quality baseline.
PLAYBOOK	The process is formalised and tested, with clearly defined exceptions and thresholds.	Version-controlled playbook with tests and an assigned owner.
IDENTITY	Agents, tools and connectors use least privilege and temporary credentials.	Access matrix, audit records and automated revocation.
OBSERVABILITY	Prompts, retrieval, decisions, tool use and overrides are fully traceable.	End-to-end audit trail and anomaly alerts.
REVERSIBILITY	The action can be reversed, offset or contained without causing widespread harm.	Tested rollback, kill switch and escalation criteria.

Progression criterion: A scenario passes a gate only when the prerequisite is operational and the required evidence is available.

Autonomy increases with the ability to **demonstrate, observe** and **reverse actions**.

PART I

Risk: A New Attack Surface and Operational Control

AI-AUGMENTED SOC

03

| Al-augmented SOC: autonomy, guardrails and control

SOAR, SIEM correlation and automated playbooks have already brought automation into Security Operations.

Generative models add **language** and **contextual** capabilities: they summarise incidents, correlate evidence, explain queries and support the **investigative workflow**.

Expected benefits must be measured alongside the risks of overreliance, data leakage and unsupported decisions.



1 | Traditional SOC

Predominantly manual triage and investigation.

Backlogs, inconsistency and limited scalability.

MTTA, MTTD, backlog.

2 | Data-driven SOC

Consolidated telemetry, defined use cases and structured analytics.

Silos, incomplete coverage and data quality issues.

Log coverage and alert accuracy.

3 | AI-assisted SOC

Copilots for triage, enrichment, summarisation and reporting.

Overreliance, hallucinations and data leakage.

Productivity, accuracy and overrides.

4 | Agentic

Tool-enabled agents performing scoped tasks subject to approval.

Privilege misuse, tool misuse and error propagation.

Task approval, execution and rollback.

5 | Autonomous scope

Autonomous remediation in high-confidence scenarios.

False remediation and business impact.

Task approval, execution and rollback.

| Non-Negotiable Guardrails

01

Autonomy thresholds defined before deployment: which actions are recommended, which require approval and which may be performed automatically.

02

Rollback and containment: every automated action has a tested reversal or compensation mechanism.

03

End-to-end auditability: inputs, context, model, tools invoked, approvals, outputs and overrides remain fully traceable.

04

Quality measurement: triage accuracy, false positives, acceptance rates, time saved and introduced incidents are monitored together.

Operational implication: the **DGS X-SOC** provides the **operational environment** for developing **AI-augmented Cyber Defence** capabilities. Moving from **claims** to **demonstrable capabilities** requires an **internal catalogue** of operational use cases, metrics, authorised datasets, controls, limitations and publishable customer cases.

PART I

Risk: A New Attack Surface and Operational Control

GEN AI, RAG AND AGENTS

D4

| GenAI, RAG and agents: Protecting Data and Actions

The CISO must govern the entire system, not just the model. Risk spans data, prompts, orchestration, retrieval, vector databases, APIs, connectors, identities, memory, tools, outputs and human processes. Controls restricting access and actions must remain external to the model and enforceable even when it produces an incorrect or manipulated output.

LLM Security: Governing Inputs, Access and Outputs

Prompt injection, data leakage, model abuse and model extraction target different surfaces. The response requires access controls, rate limiting, a model gateway, separation of data and instructions, logging and output management. Effectiveness must be demonstrated through **reproducible testing** and **runtime telemetry**.

RAG Security: Trust in Sources and Retrieval

A RAG system introduces a trust boundary between the model and enterprise sources. Provenance, ownership, classification, versioning, permission trimming and vector database protection determine which content can enter the context. Controls must ensure that queries, retrieved documents, permissions and generated responses remain fully traceable.

Agentic AI Security: Identity, Privileges, and Tools

An agent is a non-human identity with the ability to act. Least privilege, temporary credentials, task-specific scopes, segregation of duties and approval of high-impact operations are security requirements, not architectural options. Memory, connectors and agent-to-agent communications must remain authenticated, authorised, monitored and revocable.

AI Red Teaming: Testing the System End to End

AI red teaming must assess the system end to end, covering interactions, retrieval, data, tools, identities, policies, workflows and dependencies. For the CISO, the relevant outcome is a reproducible finding linked to impact, controls, ownership, remediation, retesting and residual risk.

Control Architecture for AI Agents



Risks, Controls and Evidence at Every Step

01 | DETECTION

RISK

High volumes of false alerts designed to overwhelm the agent.

CONTROL

Inbound deduplication and rate limiting.

COMPONENT

SIEM and orchestrator.

EVIDENCE

Log degli alert scartati.

02 | ENRICHMENT

RISK

The agent accesses more data than required.

CONTROL

Read-only connectors with task-specific scopes.

COMPONENT

Identity & access.

EVIDENCE

Access matrix and query logs.

03 | HYPOTHESIS

RISK

Fabricated or unsupported conclusion.

CONTROL

Mandatory evidence citation and minimum confidence threshold.

COMPONENT

Model gateway.

EVIDENCE

Percentage of hypotheses confirmed by analysts.

04 | CONTAINMENT

RISK

Out-of-scope actions or reused credentials.

CONTROL

Closed action catalogue, task-specific temporary credentials and analyst approval.

COMPONENT

Policy engine, secrets vault and human approval.

EVIDENCE

Rejected actions and average approval time.

05 | CLOSURE

RISK

An undetected or irreversible error.

CONTROL

Tested rollback and a kill switch that revokes the agent's credentials.

COMPONENT

Orchestrator and SIEM.

EVIDENCE

Successful rollback tests and a complete audit trail.

Prompt injection in Logs: Separating Data from Instructions

An attacker inserts an instruction in the login “device name” field, attempting to have the alert closed as a false positive. The content must enter the context as untrusted data, remain separate from the agent’s instructions, and be subject to model gateway controls.

Even if the model recommended closing the alert, the process impact and account privileges would still require **analyst approval**.

Resilience therefore depends on separating probabilistic assessment from **deterministic enforcement**.

What the case demonstrates: the control chain continues to operate even when the model makes an error, because permissions and policies remain external to it.

PART I

Risk: A New Attack Surface and Operational Control

SECURE AI LIFECYCLE

05

Secure AI Lifecycle: Gates, Evidence and Residual Risk

A Secure AI posture is built throughout the lifecycle. Each phase must clearly define risk, ownership, control requirements, **release conditions** and **evidence**. This makes go-live a **residual risk** decision, while operations remain subject to monitoring, incident management and review.

Selection and Classification

Have the use case, impact and level of autonomy been identified?

Use case profile, classification and owner.

Data and Model

Are sources, data, the model and dependencies governed?

Data lineage, provenance, inventory and assessments.

Development and Threat Modeling

Are abuse cases, requirements and controls defined before implementation?

Threat model and security requirements.

Testing and AI Red Teaming

Can the system withstand the relevant scenarios?

Testing, findings, remediation and residual risk.

Deployment and Runtime

Is the release authorised and observable?

Approval, logging, dashboard e runbook.

Incident and Change Management

Are incidents, changes and decommissioning governed?

Playbooks, audit trails, change assessments and access revocation.

Regulation and Standards: From Requirements to Controls

Scope Referenced in the Presentation

AI Act	➤	Classification, governance, obligations, transparency, human oversight and risk management.
GDPR	➤	Data limitation, DPIAs, retention, risk management and data security.
NIS2	➤	Risk management, system and network security, and supply chain security.
DORA	➤	ICT risk management, business continuity, disaster recovery, third-party risk governance and TLPT.
NIST AI RMF	➤	AI governance and risk management.
ISO/IEC 42001	➤	Artificial intelligence management system.

The **AI Act** timeline requires a phased governance approach:

1 August 2024: entry into force;

2 February 2025: prohibitions, definitions and AI literacy obligations;

2 August 2025: governance and obligations for general-purpose AI models;

2 August 2026: application of most provisions and transparency requirements.

PART I

Risk: A New Attack Surface and Operational Control

AUTONOMY AND ACCOUNTABILITY

D6

Autonomy and Accountability: The DGS Manifesto

Autonomy is ultimately a matter of **accountability**.

The principle stated in IBM's 1979 management training manual remains relevant today:

"A computer can never be held accountable, therefore a computer must never make a management decision." [1]

When judgement or action is delegated to AI, responsibility shifts upstream to the design of its authority, limits and operating conditions.

Every level of autonomy must be **authorised**, **observable** and **subject to review**.

Management accepts the risk; the CISO defines requirements and control thresholds; **owners** and operations maintain **evidence**, containment capabilities and accountability for outcomes.

- 01** **Granting autonomy is a management decision.** It means determining in advance how much authority to give the system and which consequences the organisation is prepared to accept.
- 02** **Uncertainty does not shift accountability.** Those authorising an autonomous system must understand the accepted risk and be able to intervene when it exceeds the defined threshold.
- 03** **Decisions must remain traceable and auditable.** The organisation must know where AI operates and be able to review material decisions.
- 04** **Autonomy should be granted incrementally.** Broader operational capabilities require equally robust control and response capabilities.

Our mission: DGS supports organisations in the **secure adoption of AI**.

Our approach combines risk assessment, control design, technology implementation, testing and operational oversight. The goal is to make every **AI system identifiable, authorised, observable** and, where necessary, **reversible**.

PART II

The DGS Response: From Risk to Operational Oversight

THE FOUNDATIONS OF THE DGS APPROACH

D7

The CISO needs an approach grounded in **actual exposure**, where every decision can be verified.

DGS brings together **Cybersecurity**, **Data & AI**, Cloud and System Integration, connecting assessment, control design, implementation, testing and operational oversight.

The scope is defined by **data**, **identities**, **privileges**, connectors, autonomy and process impact.

Guiding Principle:
Autonomy is acceptable only within a defined scope, with clear accountability, verifiable evidence and reversible actions.

| DGS Capabilities for Operational Oversight

01 | **Managed Security Services**

Transforming cybersecurity from an operational function into a strategic driver of business protection by combining threat intelligence, advanced automation and specialist expertise.

02 | **Identity Security**

Governance of identities, privileges, credentials and non-human identities used by agents and connectors.

03 | **Cloud Security and CNAPP**

Control over configurations, workloads, security posture and dependencies across cloud and hybrid architectures.

04 | **OT/IoT and Critical Infrastructure**

Continuity, resilience and risk control in mission-critical environments.

| DGS Capabilities for Operational Oversight

05 | CyLABs

Experimentation and validation environments for assessments, testing, benchmarking and control verification.

06 | Data & AI Adoption and system integration

Strategic AI adoption, definition of usage guardrails, data, GenAI, RAG and agent architectures, and integration with enterprise processes, applications and controls.

07 | Advisory and governance

Risk assessment, control model design, accountability and compliance evidence.

These capabilities deliver value when tailored to the use case and linked to an **owner**, a **risk decision** and a defined set of **evidence**. The level of oversight depends on the agreed scope and the technologies deployed.

PART II

The DGS Response: From Risk to Operational Oversight

FROM RISK ASSESSMENT TO CONTINUOUS CONTROL

D8

The DGS approach enables security decisions to be made progressively. Implementation follows asset identification, risk classification and the definition of operating conditions. Each phase must conclude with a **verifiable output**, an assigned **owner** and a **residual risk** decision.

Assessment

Which systems, data, identities, connectors and actions fall within scope?

Inventory, classification, asset map, risk register and control gaps.

Threat modeling

Which scenarios could compromise confidentiality, integrity, availability or the ability to act?

Abuse cases, attack paths, security requirements and remediation priorities.

Control design

Which controls must remain deterministic and external to the model?

Architecture, IAM, policy engine, approval workflows, logging and rollback design.

Validation

Can the system withstand the relevant scenarios, and do the controls produce the expected evidence?

Testing, reproducible findings, remediation, retesting and residual risk.

Release

Which conditions authorise go-live?

Release gates, approvals, runbooks and acceptance criteria.

Operations

How can abuse, drift, scope violations and incidents be detected?

Telemetry, detection use cases, KPIs/KRIs, playbooks and audit trails.

Change and Decommissioning

How should changes to models, sources, prompts, connectors or privileges be reassessed?

Change assessments, access revocation, evidence updates and decommissioning.

From Model to Policy: Separating Hypotheses from Enforcement

The model can analyse context, formulate hypotheses and propose actions. Enforcement must remain deterministic. IAM, policy engines, approval workflows and segregation of duties must constrain actions even when the model fails, is manipulated or operates with incomplete information.

LEVEL	FUNCTION	EVIDENCE
AI MODEL	Analyses context, formulates hypotheses and proposes actions.	Prompt, context, sources, model version and confidence level.
POLICY AND IDENTITY	Verify scope, identities, privileges and permitted actions.	Policy decision, credentials used and blocked actions.
HUMAN OVERSIGHT	Approves high-impact or difficult-to-reverse operations.	Approver role, rationale, timestamp and override.
RUNTIME AND RESPONSE	Detect anomalies, revoke access and trigger containment or rollback.	Alerts, timeline, revocation, rollback outcome and audit trail.

| **Accountability:** Who Decides, Who Controls, Who Is Accountable

01 Business owner

Risk acceptance and process impact.

02 AI / Product owner

Quality, behaviour and management of the AI system.

03 CISO

Cyber risk requirements and oversight.

04 Data owner

Data quality, access and provenance.

05 Legal / Compliance

Interpretation of the applicable regulatory framework.

06 Operations / SOC

Monitoring, incident management and response.

PART II

The DGS Response: From Risk to Operational Oversight

SECURE AI CAPABILITIES

09

Secure AI capabilities: control by design and validation

Secure AI requires coordinated expertise and clear accountability. DGS brings together advisory, design, implementation, testing and operational support to move each use case from unmeasured exposure to a **controlled, observable** and **documented environment**.

AREA	ACTIVITIES	EXPECTED EVIDENCE
AI RISK & SECURITY ASSESSMENT	Inventory, classification and assessment of data, identities, privileges and dependencies.	Risk register, heat map, control gaps and remediation roadmap.
LLM AND RAG SECURITY	Threat modelling, provenance reviews, permission trimming, gateway controls and testing.	Attack paths, findings, control design and retesting.
AGENT SECURITY	Analysis of non-human identities, tools, memory, credentials and actions.	Tool map, privilege matrix, approval design, kill switch and audit trail.
SECURE ARCHITECTURE	Integration of IAM, policy engines, logging, segregation of duties and runtime controls.	Architectural blueprint, requirements and validation plan.
AI RED TEAMING	Adversarial testing across retrieval, tools and workflows.	Reproducible findings, replay pack, remediation and regression testing.
SECURITY MONITORING	Design of telemetry and detection use cases for AI systems and agents.	Dashboards, rules, KPIs/KRIs and incident playbooks.
GOVERNANCE AND COMPLIANCE	Linking assets, owners, controls, changes and evidence.	Control map, evidence backlog and review process.

Cybersecurity for AI Case: Containing a Compromised RAG System

A seemingly legitimate document enters a RAG assistant's knowledge base containing instructions designed to manipulate the model.

The risk spans the entire chain: ingestion, indexing, retrieval, generation, action and response.

For the CISO, the key question is whether a **compromised source** can be detected, isolated and removed before causing information disclosure or out-of-scope actions.

01 | INGESTION

RISK

Unauthorised source or altered document.

CONTROL

Provenance, ownership, classification and approval.

EVIDENCE

Source, version, hash and approver.

02 | INDEXING

RISK

Malicious content enters the corpus.

CONTROL

Scanning, quarantine and version control.

EVIDENCE

Control results and rationale.

03 | RETRIEVAL

RISK

The system retrieves out-of-scope content.

CONTROL

Permission trimming and context policies.

EVIDENCE

Queries, retrieved documents and permissions.

04 | GENERATION

RISK

The model follows instructions embedded in the data.

CONTROL

Data-instruction separation, model gateway and output validation.

EVIDENCE

Prompt, context, applied rules and response.

05 | ACTION

RISK

The output triggers a connector or tool.

CONTROL

Closed action catalogue, least privilege and approval.

EVIDENCE

Identity, tool call, policy, approval and outcome.

06 | RESPONSE

RISK

The compromise remains undetected.

CONTROL

Alerts, source isolation, re-indexing and an incident playbook.

EVIDENCE

Alerts, source isolation, re-indexing and an incident playbook.

Resilience criterion: A control is effective when a compromised source can be **identified**, **isolated** and **removed** without losing **incident traceability**.

CISO Metrics: Effectiveness, Efficiency and Control

1 | Effectiveness

Triage accuracy, confirmed findings, contained scenarios and prevented incidents.

2 | Efficiency

MTTA, MTTD, average investigation time, backlog reduction and time saved.

3 | Control

Human overrides, policy-blocked actions, scope violations, successful rollbacks and audit trail completeness.

A metric is useful when it supports a decision. Each indicator must therefore have an **owner**, data source, measurement frequency, **threshold** and **associated action**. Targets and tolerances depend on **impact**, industry and **accepted risk**.

PART II

The DGS Response: From Risk to Operational Oversight

CLIENT ROADMAP

10

Client Roadmap: Visibility, Readiness and Controlled Scale-Up

The roadmap starts with visibility across the scope and progresses only when controls produce **evidence**. Inventory, **ownership** and risk classification precede the pilot; testing, telemetry and runbooks precede deployment; expansion to new use cases follows a **residual risk** review.

0–30 days

Inventory of use cases and assets, ownership, risk classification and immediate priorities.

A shared view of systems, data, identities, connectors and actions.

31–60 days

Threat modelling, control gaps, target architecture and a pilot within a defined scope.

Requirements, controls, test plan and go-live criteria.

61–90 days

Validation, telemetry, playbooks, KPIs/KRIs and operational readiness.

Evidence pack, residual risk, runbook and release decision.

3–12 months

Phased expansion to new use cases, continuous monitoring and control reviews.

A repeatable, measurable capability embedded in security processes.

CONCLUSIONS

| Maintaining **Control** When the Model Fails

Generative and agentic AI expands the attack surface by accessing data, using identities, interacting with applications and modifying real-world processes. Risk extends beyond what the system generates to what it can authorise, execute and propagate.

Security posture depends on the ability to maintain control when the model fails.

Least privilege, external policies, **human oversight**, observability, **audit trails** and rollback turn autonomy from a technological promise into a manageable risk decision.

DGS supports organisations in the secure adoption of AI and its controlled use in cyber defence, connecting assessment, control design, implementation, testing and operational oversight. The intended outcome is an identifiable, authorised, observable and reversible system.

Closing Principle:
We can entrust AI with an increasing share of our actions.
Responsibility for what happens remains in our hands.

| About DGS

DGS S.p.A. is an Italian technology partner that has supported the growth of businesses and public-sector organisations for 29 years. Drawing on extensive technology and market expertise, DGS enables the secure modernisation of processes, systems and infrastructure.

Through an integrated portfolio spanning Cybersecurity, AI & Data, Digital Engineering, Management Consulting and Smart Enterprise solutions, DGS helps organisations build solid foundations and digital trust. Its AI modernisation programmes and Secure by Design approach improve operational performance, resilience and process sustainability.

DGS operates through a network of Centres of Excellence and partnerships with leading global technology vendors. The Group brings together more than 2,700 professionals and holds over 2,500 technical certifications.

Intellectual Property and Document Confidentiality

Intellectual Property

Unless expressly stated otherwise, all content, information, methodologies, models, software, materials, graphic representations and documentation contained in this document are the exclusive property of DGS S.p.A.

The content may not be reproduced, distributed, published, modified or used, in whole or in part, without the prior written consent of DGS S.p.A.

Confidentiality

This document contains confidential and proprietary information belonging to DGS S.p.A. and is intended solely for purposes related to the evaluation of the application.

The content may not be disclosed, distributed or communicated to third parties without the prior written consent of DGS S.p.A., except where strictly necessary for evaluation activities conducted by the initiative organisers.

The information contained in this document is provided for illustrative purposes only and does not grant, license or transfer any intellectual property, industrial property or know-how rights to third parties.

CONTACT US

 @dgs_spa

 DGS SpA

 www.dgsspa.com



visit dgsspa.com