



EXECUTIVE PERSPECTIVE

SECURE AI: IL CONTROLLO PRIMA DELL'AUTONOMIA

Rischi, guardrail ed evidenze per proteggere
dati, identità e processi

AUTORI



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**

Graphic Design & Layout by Sabrina Lugato, Marco Veninati

SECURE AI: IL CONTROLLO PRIMA DELL'AUTONOMIA

Rischi, guardrail ed evidenze per proteggere dati,
identità e processi

INDICE

Prefazione	4
Executive Summary	5
PARTE I — IL RISCHIO: NUOVA SUPERFICIE DI ATTACCO E CONTROLLO OPERATIVO	
01 Cyber + AI: un unico perimetro di rischio	8
02 Threat landscape: priorità di difesa e rischio in accelerazione	10
03 AI-augmented SOC: autonomia, guardrail e controllo	15
04 GenAI, RAG e agenti: proteggere dati, identità e azioni	18
05 Secure AI Lifecycle: gate, evidenze e rischio residuo	23
06 Autonomia e accountability: il Manifesto DGS	25
PARTE II — LA RISPOSTA DGS: DAL RISCHIO AL PRESIDIO OPERATIVO	
07 Le basi del presidio DGS	26
08 Dal risk assessment al controllo continuo	29
09 Secure AI capabilities: controllo by design e validazione	32
10 Roadmap cliente: visibilità, readiness e scale-up controllato	35
CONCLUSIONI	36

PREFAZIONE

| La nuova responsabilità del controllo

A cura di Salvatore Frosina, Co-CEO DGS

Per anni abbiamo protetto sistemi progettati per eseguire istruzioni. Oggi dobbiamo governare sistemi capaci di interpretarle, generarle e trasformarle in azioni.

L'AI accede a dati, utilizza identità e connettori, supporta decisioni e interviene su applicazioni e processi reali. Il rischio riguarda quindi ciò che produce, ma soprattutto ciò che può vedere, autorizzare ed eseguire.

Per chi guida la sicurezza, la sfida è permettere all'AI di generare valore mantenendo il controllo sulle sue azioni e sulle conseguenze che possono produrre. Servono perimetri chiari, responsabilità esplicite, evidenze verificabili e la capacità di fermare o annullare un'operazione.

Questo paper raccoglie il punto di vista e l'esperienza di DGS nell'integrare Cybersecurity, Data & AI, Cloud, identity e presidio operativo per accompagnare le organizzazioni nell'adozione sicura dell'AI.

Possiamo affidare all'AI una parte crescente delle nostre attività. La responsabilità di ciò che accade resta nelle nostre mani.



Salvatore Frosina
Co-CEO DGS

EXECUTIVE SUMMARY

La tesi centrale

L'adozione di sistemi generativi e agentici modifica il profilo di rischio dell'impresa: aumenta la scala di alcune minacce note, introduce nuovi asset da proteggere e consente ai sistemi AI di interagire direttamente con dati, identità e applicazioni aziendali. Alla superficie tradizionale, reti, endpoint, infrastrutture, applicazioni e identità, si aggiungono **modelli, prompt, dataset, vector database, knowledge base, connettori, memoria degli agenti e azioni eseguite su sistemi reali**.

La priorità è mantenere il controllo mentre l'AI entra nei processi aziendali. Sicurezza, ownership ed evidenze devono quindi essere definite prima del rilascio e presidiate durante l'intero ciclo di vita.

La stessa agenda deve coprire due esigenze inseparabili: usare l'AI per aumentare prevenzione, detection e risposta, e proteggere i sistemi AI adottati dall'impresa. Dati, identità, telemetria, incident management e responsabilità sono condivisi; una governance separata crea lacune e rallenta la capacità di contenimento.

Il rischio immediato nasce dall'aumento di scala di phishing, impersonation e reconnaissance.

RAG e agenti ampliano il perimetro con indirect prompt injection, data poisoning, model extraction, tool misuse, privilegi eccessivi e propagazione automatica dell'errore. La **priorità degli investimenti** deve seguire **esposizione, impatto, rilevabilità e reversibilità**.

AI Act, GDPR, NIS2 e DORA richiedono presidi coerenti con i rispettivi ambiti di applicazione. Per il CISO, il punto di convergenza è operativo: **controlli formalizzati, responsabilità attribuite, rischio residuo esplicito ed evidenze verificabili**.

Sette priorità per il CISO: controllo, accountability e rischio residuo

01 Governare un unico perimetro di rischio

Cybersecurity e AI devono condividere governance, controlli ed evidenze.

02 Mantenere il presidio sulle decisioni ad alto impatto

L'AI può supportare triage, enrichment, analisi e reporting; autorizzazione e accountability restano assegnate a ruoli identificati.

03 Estendere il controllo dalle risposte alle azioni

Identità, privilegi, connettori e tool determinano ciò che un sistema AI può eseguire sui processi aziendali.

04 Integrare la sicurezza nel ciclo di vita

Threat modeling, data provenance, test avversariali, gate di rilascio, runtime monitoring e incident management devono precedere e accompagnare il go-live.

05 Trattare agenti e connettori come identità operative

Least privilege, credenziali temporanee, segregazione dei compiti, approval e revoca devono applicarsi anche alle identità non umane.

06 Concedere autonomia per gradi

L'automazione aumenta soltanto quando telemetria, playbook, osservabilità, rollback e kill switch sono operativi e verificati.

07 Misurare efficacia e rischio residuo

KPI, KRI, override, azioni bloccate, violazioni di scope e test di rollback devono supportare decisioni di rilascio e riesame

| Priorità nei prossimi dodici mesi

01 | PROTEGGERE ORA

PERIMETRO

Sistemi GenAI/RAG già in uso; identità, dati e connettori; shadow AI; processi antifrode e incident management.

AZIONE

Inventory, risk classification, access control, provenance, logging e playbook.

02 | COSTRUIRE ORA

PERIMETRO

Metodologie di assessment, skill map, due acceleratori, piloti AI-assisted nel SOC ed evidence pack.

AZIONE

Owner, backlog, budget seed, clienti pilota e stage gate.

03 | EVITARE PROMESSE PREMATURE

PERIMETRO

Autonomous SOC generalizzato, Managed AI Security maturo e piattaforme proprietarie AI-native senza prova.

AZIONE

Presentare traiettoria, limiti e decisioni future, con progressività e prove.

PARTE I

Il rischio: nuova superficie di attacco e controllo operativo

CYBER + AI: UN UNICO PERIMETRO DI RISCHIO

D1

L'AI concentra diversi profili di rischio nello stesso sistema: aumenta le capacità difensive, rende più efficienti alcune tecniche offensive e introduce nuovi asset con accesso a dati, identità e applicazioni. Con gli agenti, il rischio si estende dall'output all'azione.

Con gli agenti, il rischio si estende dall'output all'azione.

La conseguenza è una governance unica tra **sicurezza applicativa, data governance, IAM, SOC e compliance**.

AI FOR CYBERSECURITY

- 1 | Triage e prioritizzazione
- 2 | Enrichment
- 3 | Threat intelligence
- 4 | Detection engineering
- 5 | Vulnerability prioritization

CYBERSECURITY FOR AI

- 1 | Prompt e model security
- 2 | Protezione RAG security
- 3 | Agent identity
- 4 | Infrastrutture e supply chain AI
- 5 | Tool governance

CONVERGENZA OPERATIVA

Un **solo modello** per **persone, processi, controlli** ed **evidenze**.

| Quattro dimensioni del rischio Cyber + AI

01 AI come strumento di attacco

I modelli generativi riducono il costo di produzione di contenuti credibili, aumentano il volume degli attacchi e accelerano ricognizione, personalizzazione e supporto alla codifica malevola.

02 AI come strumento di difesa

Il valore cresce nella correlazione del contesto, nella prioritizzazione degli alert, nell'arricchimento degli incidenti e nel supporto a detection, reporting e risposta.

03 AI come asset da proteggere

Modelli, dati, prompt, knowledge base, vector database, pipeline, connettori e dipendenze diventano asset da inventariare, classificare e monitorare.

04 Agenti come rischio qualitativamente diverso

Gli agenti possono usare identità, memoria, strumenti e sistemi aziendali. Il rischio dipende dai privilegi, dall'autonomia, dalla supervisione e dalla reversibilità delle azioni.

Agire ora: Istituire una cabina di regia con **CISO, Data & AI Leadership, Enterprise Architecture, Risk e Compliance.**

Il primo risultato atteso è **una vista condivisa** e aggiornata su use case, asset, rischi, owner e controlli.

PARTE I

Il rischio: nuova superficie di attacco
e controllo operativo

THREAT LANDSCAPE

02

Threat landscape: priorità di difesa e rischio in accelerazione

La priorità non è inseguire ogni scenario emergente, ma distinguere ciò che produce rischio oggi da ciò che richiede preparazione.

L'AI industrializza tecniche consolidate e, parallelamente, introduce **attacchi diretti** a **modelli, retrieval** e **agenti**.

Questa distinzione consente al management di allocare budget e controlli in base a evidenze, esposizione e impatto potenziale.



| Maturità indicativa delle principali minacce Cyber + AI



Priorità di difesa: **proteggere, strumentare, sperimentare**

01 | **PROTEGGERE ORA**

PERIMETRO

Phishing, impersonation, credenziali, shadow AI e disclosure.

AZIONE

MFA, identity controls, DLP, policy d'uso, logging e awareness.

02 | **STRUMENTARE ORA**

PERIMETRO

GenAI/RAG in produzione, supply chain e prompt injection.

AZIONE

Inventory, threat modeling, provenance, gateway, test e osservabilità.

03 | **SPERIMENTARE IN SICUREZZA**

PERIMETRO

Agenti, tool use, multi-agent e remediation autonoma.

AZIONE

Sandbox, scope limitato, approval, temporary credentials, rollback e kill switch.

| Tre criteri per prioritizzare il rischio

01 Esposizione

La stessa tecnica produce conseguenze diverse in base ad accessibilità del sistema, valore dei dati, privilegi dei connettori e dipendenza del processo. La classe di rischio cambia radicalmente tra un agente isolato in sandbox e un agente collegato a ERP e identity provider.

02 Velocità di propagazione

La priorità cresce quando errore o compromissione possono attraversare rapidamente utenti, applicazioni, agenti e sistemi, generando effetti tecnici, operativi o finanziari.

03 Rilevabilità e reversibilità

La soglia di controllo deve aumentare quando il comportamento è difficile da osservare, attribuire, contenere o annullare.

Soglie operative: assistere, approvare, automatizzare

MODALITÀ	RESPONSABILITÀ	ESEMPI	PREREQUISITI
ASSISTERE	L'AI raccoglie contesto e propone; la persona decide.	Triage, enrichment, query, incident summary, report.	Qualità dei dati, citazione delle fonti e validazione umana.
PREPARARE + APPROVARE	L'agente predispose il task; una persona autorizza l'azione.	Ticket, cambi di priorità, contenimento e revoca di privilegi.	Identità dedicata, least privilege, approval e audit trail.
AUTOMATIZZARE SU SCOPE	Il sistema agisce entro limiti predefiniti e reversibili.	Tagging, arricchimento, azioni isolate e remediation a basso impatto.	Soglie definite, rollback, kill switch, logging e misurazione.

Proof before autonomy: i gate che autorizzano l'azione

GATE	CONDIZIONE DI SUPERAMENTO	EVIDENZA MINIMA
TELEMETRIA	Eventi, contesto e asset sono disponibili con qualità misurata.	Coverage map e baseline di data quality
PLAYBOOK	Il processo è formalizzato, testato e distingue eccezioni e soglie.	Playbook versionato con test e owner
IDENTITÀ	Agente, tool e connettori usano privilegi minimi e credenziali temporanee.	Matrice accessi, audit e revoca automatica
OSSERVABILITÀ	Prompt, retrieval, decisioni, tool e override sono ricostruibili.	Audit trail end-to-end e alert di anomalia
REVERSIBILITÀ	L'azione può essere annullata, compensata o contenuta senza danno esteso.	Rollback provato, kill switch e criteri di escalation

Criterio di avanzamento: Lo scenario supera un gate soltanto quando il prerequisito è operativo e l'evidenza è disponibile.

L'**autonomia cresce** insieme alla capacità di **provare, osservare** e **annullare**.

PARTE I

Il rischio: nuova superficie di attacco
e controllo operativo

AI-AUGMENTED SOC

03

| **AI-augmented SOC:** autonomia, guardrail e controllo

SOAR, correlazione SIEM e playbook automatizzati hanno già introdotto automazione nelle Security Operations.

I **modelli generativi** aggiungono capacità sul **linguaggio** e sul **contesto**: sintetizzano incidenti, correlano evidenze, spiegano query e supportano la **sequenza investigativa**.

Il beneficio atteso deve essere misurato insieme al rischio di overreliance, data leakage e decisioni non supportate.



1 | SOC tradizionale

Triage e investigazione prevalentemente manuali.

Backlog, disomogeneità e scarsa scalabilità.

MTTA, MTDD, backlog.

2 | Data-driven SOC

Telemetria consolidata, use case e analytics strutturati.

Silos, copertura incompleta e qualità dei dati.

Copertura log e precisione alert.

3 | AI-assisted SOC

Copilot per triage, enrichment, summarization e reporting.

Overreliance, hallucination e data leakage.

Produttività, accuratezza e override.

4 | Agentic

Agenti con tool su compiti delimitati e approvazioni.

Privilege misuse, tool misuse e propagazione dell'errore.

Task approval, execution e rollback.

5 | Autonomous scope

Remediation autonoma su scenari ad alta fiducia.

False remediation e impatto sul business.

Task approval, execution e rollback.

| Guardrail non negoziabili

01

Soglie di autonomia definite prima del rilascio: quali azioni sono suggerite, quali richiedono approvazione e quali possono essere eseguite automaticamente.

02

Rollback e contenimento: ogni azione automatica dispone di un percorso di annullamento o di una compensazione testata.

03

Auditabilità end-to-end: input, contesto, modello, tool invocati, approvazioni, output e override restano ricostruibili.

04

Misurazione della qualità: accuratezza del triage, falsi positivi, tasso di accettazione, tempi risparmiati e incidenti introdotti vengono monitorati insieme.

Implicazione operativa: l'**X-SOC di DGS** costituisce il **perimetro operativo** su cui sviluppare capacità di **AI-augmented Cyber Defense**. Il passaggio dal **claim** alla **capability dimostrabile** richiede un **catalogo interno** dei casi d'uso operativi, metriche, dataset autorizzati, controlli, limiti e casi cliente pubblicabili.

PARTE I

Il rischio: nuova superficie di attacco e controllo operativo

GEN AI, RAG E AGENTI

D4

| GenAI, RAG e agenti: proteggere dati, identità e azioni

Il CISO **deve governare l'intero sistema**, non soltanto il modello. Il rischio attraversa dati, prompt, orchestrazione, retrieval, vector database, API, connettori, identità, memoria, tool, output e processo umano. I controlli che limitano accessi e azioni devono rimanere esterni al modello e applicabili anche quando il modello produce un risultato errato o manipolato.

LLM Security: governare input, accessi e output

Prompt injection, data leakage, model abuse ed extraction agiscono su superfici differenti. La risposta richiede controllo degli accessi, rate limiting, model gateway, separazione tra dati e istruzioni, logging e gestione dell'output. L'efficacia deve essere dimostrata attraverso **test riproducibili e telemetria runtime**.

RAG Security: fiducia nelle fonti e nel retrieval

Un sistema RAG introduce un dominio di fiducia tra modello e fonti aziendali. Provenance, ownership, classificazione, versionamento, permission trimming e protezione del vector database determinano quali contenuti possono entrare nel contesto. Il controllo deve rendere ricostruibili query, documenti recuperati, autorizzazioni e risposta generata.

Agentic AI Security: identità, privilegi e tool

Un agente è un'identità non umana con capacità di azione. Privilegi minimi, credenziali temporanee, scope per task, segregazione dei doveri e approvazione delle operazioni ad alto impatto rappresentano requisiti di sicurezza, non opzioni architetturali. Memoria, connettori e comunicazioni agent-to-agent devono restare autenticati, autorizzati, monitorati e revocabili.

AI red teaming: testare il sistema end-to-end

L'AI red teaming deve verificare il sistema end-to-end: interazioni, retrieval, dati, tool, identità, policy, workflow e dipendenze. Il risultato utile al CISO è un finding riproducibile, collegato a impatto, controllo, owner, remediation, retest e rischio residuo.

Architettura di controllo per gli agenti AI



Rischi, controlli ed evidenze a ogni passo

01 | RILEVAZIONE

RISCHIO

Alert falsi in massa per saturare l'agente.

CONTROLLO

Deduplica e rate limit in ingresso.

COMPONENTE

SIEM e orchestrator.

EVIDENZA

Log degli alert scartati.

02 | ARRICCHIMENTO

RISCHIO

L'agente legge più del necessario.

CONTROLLO

Connettori in sola lettura con scope per task.

COMPONENTE

Identity & access.

EVIDENZA

Matrice accessi e log delle query.

03 | IPOTESI

RISCHIO

Conclusione inventata o non supportata.

CONTROLLO

Obbligo di citare le evidenze e soglia minima di confidenza.

COMPONENTE

Model gateway.

EVIDENZA

Percentuale di ipotesi confermate dagli analisti.

04 | CONTENIMENTO

RISCHIO

Azioni fuori perimetro o credenziali riusate.

CONTROLLO

Catalogo chiuso, credenziali temporanee legate al task e approvazione dell'analista.

COMPONENTE

Policy engine, secrets vault e human approval.

EVIDENZA

Azioni rifiutate e tempo medio di approvazione.

05 | CHIUSURA

RISCHIO

Errore non visto o non annullabile.

CONTROLLO

Rollback testato e kill switch che revoca le credenziali dell'agente.

COMPONENTE

Orchestrator e SIEM.

EVIDENZA

Test di rollback superati e audit trail completo.

| Prompt injection nei log: separare dati e istruzioni

Nel campo “nome dispositivo” del login, un attaccante inserisce un’istruzione che tenta di far chiudere l’alert come falso positivo. Il contenuto deve entrare nel contesto come dato non fidato, separato dalle istruzioni dell’agente e sottoposto ai controlli del model gateway.

Anche se il modello proponesse la chiusura, l’impatto sul processo e i privilegi dell’account imporrebbero l’**approvazione dell’analista**.

La resilienza deriva quindi dalla separazione tra valutazione probabilistica ed **enforcement deterministico**.

Cosa dimostra il caso: la catena di controllo continua a funzionare anche in presenza di un errore del modello, perché permessi e policy restano esterni.

PARTE I

Il rischio: nuova superficie di attacco e controllo operativo

SECURE AI LIFECYCLE

D5

| Secure AI Lifecycle: gate, evidenze e rischio residuo

La postura Secure AI si costruisce lungo il ciclo di vita. Ogni fase deve rendere espliciti rischio, owner, requisiti di controllo, **condizioni di rilascio** ed **evidenze**. In questo modo il go-live diventa una decisione sul **rischio residuo** e l'esercizio resta soggetto a monitoraggio, incident management e riesame.

Selezione e classificazione

Il caso d'uso, l'impatto e il livello di autonomia sono identificati?

Scheda use case, classificazione e owner.

Dati e modello

Fonti, dati, modello e dipendenze sono governati?

Data lineage, provenance, inventory e valutazioni.

Sviluppo e threat modeling

Abuse case, requisiti e controlli sono definiti prima dell'implementazione?

Threat model e security requirements.

Testing e AI red teaming

Il sistema resiste agli scenari rilevanti?

Test, finding, remediation e rischio residuo.

Deployment e runtime

Il rilascio è autorizzato e osservabile?

Approval, logging, dashboard e runbook.

Incident e cambiamento

Incidenti, modifiche e dismissione sono governati?

Playbook, audit trail, change assessment e revoca accessi.

Regolazione e standard: tradurre obblighi in controlli

Ambito richiamato nel PPT

AI Act	➤	Classificazione, governance, obblighi, trasparenza, human oversight e gestione del rischio.
GDPR	➤	Limitazione, DPIA, retention, risk management e data security.
NIS2	➤	Risk management, system and network security e supply chain.
DORA	➤	ICT risk management, business continuity, disaster recovery, third-party risk governance e TLPT.
NIST AI RMF	➤	Governance e gestione del rischio AI.
ISO/IEC 42001	➤	Sistema di gestione dell'intelligenza artificiale.

La timeline dell'**AI Act** richiede un presidio progressivo:

1° agosto 2024: entrata in vigore;

2 febbraio 2025: divieti, definizioni e obblighi di AI literacy;

2 agosto 2025: governance e obblighi per i modelli di AI per finalità generali;

2 agosto 2026: applicazione della maggior parte delle disposizioni e delle regole di trasparenza.

PARTE I

Il rischio: nuova superficie di attacco e controllo operativo

AUTONOMIA E ACCOUNTABILITY

06

| Autonomia e accountability: il Manifesto DGS

La questione dell'autonomia è una questione di **accountability**.

Il principio richiamato dal manuale IBM del 1979 resta attuale:

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

Quando il giudizio o l'azione passano all'AI, la responsabilità si concentra a monte, nel disegno dei poteri, dei limiti e delle condizioni di intervento.

Ogni livello di autonomia deve essere **autorizzato, osservabile e riesaminabile**.

Il management accetta il rischio; il CISO definisce requisiti e soglie di controllo; **owner** e operations mantengono **evidenze**, capacità di contenimento e responsabilità sull'esito.

- 01** **Dare autonomia è una scelta manageriale.** Significa stabilire in anticipo quanto potere affidare al sistema e quali conseguenze siamo disposti ad accettare.
- 02** **L'incertezza non sposta la responsabilità.** Chi autorizza un sistema autonomo deve conoscere il rischio accettato e poter intervenire quando supera la soglia definita.
- 03** **Le decisioni devono rimanere tracciabili e ricostruibili.** L'organizzazione deve sapere dove opera l'AI e poter riesaminare le scelte rilevanti.
- 04** **L'autonomia si concede un passo alla volta.** Capacità operative più ampie richiedono una capacità altrettanto chiara di controllo e risposta.

| **La nostra missione:** DGS accompagna le organizzazioni nell'**adozione sicura dell'AI**. L'approccio integra valutazione del rischio, progettazione dei controlli, implementazione tecnologica, testing e presidio operativo. L'obiettivo è rendere ogni sistema **AI identificabile, autorizzato, osservabile** e, quando necessario, **reversibile**.

PARTE II

La risposta DGS: dal rischio al presidio operativo

LE BASI DEL PRESIDIO DGS

D7

Il CISO ha bisogno di un percorso che parta dall'**esposizione reale** e renda ogni scelta verificabile.

DGS integra **cybersecurity, Data & AI**, cloud e system integration per collegare assessment, control design, implementazione, testing e presidio operativo.

Il perimetro viene definito sulla base di **dati, identità, privilegi**, connettori, autonomia e impatto sul processo.

**Principio di intervento:
l'autonomia è
accettabile solo entro
un perimetro definito,
con responsabilità
esplicite, evidenze
verificabili e azioni
reversibili.**

| Capability DGS a supporto del presidio

01 | **Managed Security Services**

Trasformare la cybersecurity da funzione operativa a fattore strategico di protezione del business, combinando threat intelligence, automazione avanzata e competenze specialistiche

02 | **Identity Security**

Governance delle identità, privilegi, credenziali e identità non umane utilizzate da agenti e connettori.

03 | **Cloud Security e CNAPP**

Controllo di configurazioni, workload, posture e dipendenze nelle architetture cloud e ibride.

04 | **OT/IoT e infrastrutture critiche**

Continuità, resilienza e controllo del rischio in ambienti mission-critical.

| Capability DGS a supporto del presidio

05 | CyLABs

Ambienti di sperimentazione e validazione per assessment, test, benchmark e verifica dei controlli.

06 | Data & AI Adoption e system integration

Adozione strategica dell'AI, definizione dei guardrail di utilizzo, architetture dati/GenAI/RAG/agenti e integrazione con processi, applicazioni e controlli enterprise.

07 | Advisory e governance

Valutazione del rischio, definizione del modello di controllo, responsabilità ed evidenze di compliance.

Queste capability acquistano valore quando sono composte sul caso d'uso e collegate a un **owner**, a una **decisione di rischio** e a un set di **evidenze**. La profondità del presidio dipende dal **perimetro** concordato e dalle **tecnologie** effettivamente adottate.

PARTE II

La risposta DGS: dal rischio al presidio operativo

DAL RISK ASSESSMENT AL CONTROLLO CONTINUO

08

Il percorso DGS rende progressiva la decisione di sicurezza.

L'implementazione segue l'identificazione degli asset, la classificazione del rischio e la definizione delle condizioni di esercizio. Ogni fase deve chiudersi con un **output verificabile**, un **owner** e una decisione sul **rischio residuo**.

Assessment

Quali sistemi, dati, identità, connettori e azioni rientrano nel perimetro?

Inventory, classificazione, mappa degli asset, risk register e gap di controllo.

Threat modeling

Quali scenari possono compromettere confidenzialità, integrità, disponibilità o capacità di azione?

Abuse case, attack path, security requirements e priorità di remediation.

Control design

Quali controlli devono restare deterministici ed esterni al modello?

Architettura, IAM, policy engine, approval flow, logging e rollback design.

Validation

Il sistema resiste agli scenari rilevanti e i controlli producono le evidenze attese?

Test, finding riproducibili, remediation, retest e rischio residuo.

Release

Quali condizioni autorizzano il go-live?

Gate di rilascio, approvazioni, runbook e criteri di accettazione.

Operations

Come rilevare abuso, drift, violazioni di scope e incidenti?

Telemetria, use case di detection, KPI/KRI, playbook e audit trail.

Change e dismissione

Come rivalutare modifiche a modello, fonti, prompt, connettori o privilegi?

Change assessment, revoca accessi, aggiornamento delle evidenze e decommissioning.

Dal modello alla policy: separare ipotesi ed enforcement

Il modello può analizzare contesto, formulare ipotesi e proporre azioni. L'enforcement deve restare deterministico. IAM, policy engine, workflow autorizzativi e segregazione dei compiti devono contenere l'azione anche quando il modello sbaglia, viene manipolato o opera con informazioni incomplete.

LIVELLO	FUNZIONE	EVIDENZA
MODELLO AI	Analizza contesto, formula ipotesi e propone azioni.	Prompt, contesto, fonti, versione del modello e livello di confidenza.
POLICY E IDENTITY	Verificano perimetro, identità, privilegi e azioni consentite.	Decisione di policy, credenziali utilizzate e azioni bloccate.
HUMAN OVERSIGHT	Approva le operazioni ad alto impatto o difficilmente reversibili.	Ruolo approvatore, motivazione, timestamp e override.
RUNTIME E RESPONSE	Rilevano anomalie, revocano accessi e attivano contenimento o rollback.	Alert, timeline, revoca, esito del rollback e audit trail.

| **Accountability:** chi decide, chi controlla, chi risponde

01 Business owner

Accettazione del rischio e impatto sul processo.

02 AI / Product owner

Qualità, comportamento e gestione del sistema AI.

03 CISO

Requisiti e presidio del rischio cyber.

04 Data owner

Qualità, accesso e provenienza dei dati.

05 Legal / Compliance

Interpretazione del quadro normativo applicabile.

06 Operations / SOC

Monitoraggio, gestione dell'incidente e risposta.

PARTE II

La risposta DGS: dal rischio al presidio operativo

SECURE AI CAPABILITIES

09

Secure AI capabilities: controllo by design e validazione

La Secure AI richiede competenze coordinate e responsabilità chiare. DGS collega advisory, progettazione, implementazione, testing e supporto operativo per portare ogni caso d'uso da una condizione di esposizione non misurata a un **perimetro controllato, osservabile e documentato**.

AMBITO	ATTIVITÀ	EVIDENZA ATTESA
AI RISK & SECURITY ASSESSMENT	Inventory, classificazione, valutazione di dati, identità, privilegi e dipendenze.	Risk register, heatmap, control gap e remediation roadmap.
LLM E RAG SECURITY	Threat modeling, provenance review, permission trimming, gateway e test.	Attack path, finding, control design e retest.
AGENT SECURITY	Analisi di identità non umane, tool, memoria, credenziali e azioni.	Tool map, matrice privilegi, approval design, kill switch e audit trail.
SECURE ARCHITECTURE	Integrazione di IAM, policy engine, logging, segregazione e runtime controls.	Blueprint architetturale, requisiti e piano di validazione.
AI RED TEAMING	Test avversariali sul sistema, inclusi retrieval, tool, workflow e dipendenze.	Finding riproducibili, replay pack, remediation e regressione.
SECURITY MONITORING	Disegno della telemetria e dei casi d'uso di detection per sistemi AI e agenti.	Dashboard, regole, KPI/KRI e incident playbook.
GOVERNANCE E COMPLIANCE	Collegamento tra asset, owner, controlli, modifiche ed evidenze.	Control map, evidence backlog e processo di riesame.

Caso Cybersecurity for AI: contenere una RAG compromessa

Un documento apparentemente legittimo entra nella knowledge base di un assistente RAG e contiene istruzioni progettate per manipolare il modello. Il rischio attraversa l'intera catena: ingestion, indexing, retrieval, generation, action e response.

Per il CISO, la domanda è se una **fonte compromessa** possa essere individuata, isolata e rimossa prima che produca disclosure o azioni fuori perimetro.

01 | INGESTION

RISCHIO

Fonte non autorizzata o documento alterato.

CONTROLLO

Provenance, ownership, classificazione e approvazione.

EVIDENZA

Origine, versione, hash e approvatore.

02 | INDEXING

RISCHIO

Contenuto malevolo entra nel corpus.

CONTROLLO

Scanning, quarantena e versionamento.

EVIDENZA

Esito dei controlli e motivazione.

03 | RETRIEVAL

RISCHIO

Il sistema recupera contenuti fuori scope.

CONTROLLO

Permission trimming e policy sul contesto.

EVIDENZA

Query, documenti recuperati e autorizzazioni.

04 | GENERATION

RISCHIO

Il modello segue istruzioni presenti nei dati.

CONTROLLO

Separazione dati/istruzioni, model gateway e output validation.

EVIDENZA

Prompt, contesto, regole applicate e risposta.

05 | ACTION

RISCHIO

L'output attiva un connettore o un tool.

CONTROLLO

Catalogo chiuso, least privilege e approval.

EVIDENZA

Identità, tool call, policy, approvazione ed esito.

06 | RESPONSE

RISCHIO

La compromissione resta invisibile.

CONTROLLO

Alert, isolamento della fonte, re-indexing e incident playbook.

EVIDENZA

Timeline, contenimento, impatto e remediation.

Criterio di resilienza: Il controllo è efficace quando una fonte compromessa può essere **identificata**, **isolata** e **rimossa** senza perdere la **ricostruibilità dell'incidente**.

Metriche CISO: efficacia, efficienza e controllo

1 | Efficacia

Precisione del triage, finding confermati, scenari contenuti e incidenti evitati.

2 | Efficienza

MTTA, MTTD, tempo medio d'indagine, riduzione del backlog e tempo risparmiato.

3 | Controllo

Override umano, azioni bloccate dalle policy, violazioni di scope, rollback riusciti e completezza dell'audit trail.

Una metrica è utile quando supporta una decisione. Ogni indicatore deve quindi avere **owner**, fonte dati, frequenza di rilevazione, **soglia** e **azione associata**.

Target e tolleranze dipendono dall'**impatto**, dal settore e dal **rischio accettato**.

PARTE II

La risposta DGS: dal rischio al presidio operativo

ROADMAP CLIENTE

10

Roadmap cliente: visibilità, readiness e scale-up controllato

La roadmap parte dalla visibilità sul perimetro e avanza soltanto quando i controlli producono **evidenze**. Inventory, **ownership** e classificazione del rischio precedono il pilota; test, telemetria e runbook precedono il rilascio; l'estensione a nuovi casi segue il riesame del **rischio residuo**.

0-30 giorni

Inventory di use case e asset, owner, classificazione del rischio e priorità immediate.

Vista condivisa su sistemi, dati, identità, connettori e azioni.

31-60 giorni

Threat modeling, control gap, architettura target e pilota su un caso delimitato.

Requisiti, controlli, piano di test e criteri di go-live.

61-90 giorni

Validazione, telemetria, playbook, KPI/KRI e readiness operativa.

Evidence pack, rischio residuo, runbook e decisione di rilascio.

3-12 mesi

Estensione progressiva a nuovi casi, monitoraggio continuo e riesame dei controlli.

Capability ripetibile, misurata e integrata nei processi di sicurezza.

CONCLUSIONI

| Mantenere il **controllo** quando il modello sbaglia

L'AI generativa e agentica amplia la superficie di attacco perché accede a dati, usa identità, interagisce con applicazioni e può modificare processi reali. Il rischio non riguarda soltanto ciò che il sistema genera, ma ciò che può autorizzare, eseguire e propagare.

La postura di sicurezza dipende dalla capacità di mantenere il controllo quando il modello sbaglia. **Privilegi minimi**, policy esterne, **human oversight**, osservabilità, **audit trail** e rollback trasformano l'autonomia da promessa tecnologica a scelta di rischio governabile.

DGS accompagna le organizzazioni nell'adozione sicura dell'AI e nell'impiego controllato dell'AI a supporto della difesa, collegando assessment, **control design**, implementazione, testing e presidio operativo. Il risultato atteso è un sistema identificabile, autorizzato, osservabile e reversibile.

Principio conclusivo:
Possiamo affidare all'AI una parte crescente delle nostre azioni. La responsabilità di ciò che accade resta nelle nostre mani.

| Chi è DGS

DGS S.p.A. è il partner tecnologico italiano che da 29 anni affianca la crescita di imprese e pubbliche amministrazioni attraverso la modernizzazione in sicurezza di processi, sistemi e infrastrutture, grazie a una conoscenza approfondita delle tecnologie e dei mercati.

Con un'offerta che integra servizi di Cybersecurity, AI & Data, Digital Engineering, Management Consulting e soluzioni per la Smart Enterprise, DGS accompagna le organizzazioni nella costruzione di basi solide e fiducia digitale per il miglioramento delle performance operative, della resilienza e della sostenibilità dei processi, attraverso percorsi di AI modernization e un approccio Secure by Design.

DGS opera attraverso un modello di Centri di Eccellenza e di partnership con i principali vendor tecnologici globali. Il Gruppo conta oltre 2.700 professionisti e più di 2.500 certificazioni tecniche.

Proprietà intellettuale e riservatezza del documento

Proprietà Intellettuale

Tutti i contenuti, le informazioni, le metodologie, i modelli, il software, gli elaborati, le rappresentazioni grafiche e la documentazione descritti nel presente materiale sono di esclusiva titolarità di DGS S.p.A., salvo diversa indicazione espressa.

La riproduzione, distribuzione, pubblicazione, modifica o utilizzo, anche parziale, dei contenuti è vietata senza il preventivo consenso scritto di DGS S.p.A.

Riservatezza

Il presente documento contiene informazioni riservate e proprietarie di DGS S.p.A. ed è destinato esclusivamente alle finalità connesse alla valutazione della candidatura.

È vietata la divulgazione, distribuzione o comunicazione a terzi dei contenuti qui riportati senza il preventivo consenso scritto di DGS S.p.A., fatti salvi gli utilizzi strettamente necessari alle attività di valutazione da parte degli organizzatori dell'iniziativa.

Le informazioni contenute nel presente documento sono fornite esclusivamente a fini illustrativi e non comportano alcuna concessione, licenza o trasferimento di diritti di proprietà intellettuale, industriale o di know-how in favore di terzi."

CONTATTACI

 @dgs_spa

 DGS SpA

 www.dgsspa.com



visit dgsspa.com