# Gestione dei rischi ICT nei sistemi de IA generativa

## Proposta de valore

Giugno 2026

www.managementsolutions.com

# Indice



# 1 Introduzione
## Contesto

**Le organizzazioni stanno implementando rapidamente sistemi di IA generativa su processi, dati e terze parti senza che i loro tradizionali framework di riferimento per il rischio ICT coprano adeguatamente le nuove risorse, i flussi di informazioni e i vettori di attacco**

**Rischio IA**

Il rischio IA comprende gli **impatti negativi derivanti dalla progettazione, dallo sviluppo, dall'implementazione e dall'uso dei sistemi di IA**. Dal punto di vista delle TIC, questi rischi **devono essere gestiti in modo specifico per il loro potenziale impatto** sulla sicurezza, la continuità e la resilienza operativa digitale.


### Cosa sta succedendo?

Le organizzazioni stanno **integrando rapidamente l'IA generativa nei processi**, negli strumenti aziendali e nelle soluzioni di terze parti, inclusi agenti, assistenti, flussi di lavoro e funzionalità di IA incorporata.

### Cosa cambia?

Questi sistemi **ampliano il perimetro della gestione tradizionale del rischio ICT** introducendo nuove risorse, flussi di informazioni, tipologie di errori, dipendenze tecnologiche e vettori di attacco che non sono sempre coperti dai quadri normativi attuali.

### Perché è importante?

Se questi elementi non vengono gestiti in modo specifico, **possono influire sulla sicurezza delle informazioni**, sulla **continuità operativa**, sulla **conformità normativa** e sulla fiducia dei clienti e delle autorità di vigilanza.

# 2 La nostra visione

## Cambiamenti nel modello di controllo ICT determinati da Gen AI

**L'IA generativa non richiede la creazione di un framework ICT parallelo, ma piuttosto l'estensione del framework esistente per coprire nuovi asset, root causes, conseguenze specifiche e obiettivi di controllo**

### Catalogo dei rischi ICT

Non compaiono necessariamente nuovi rischi di primo livello, ma **nuove manifestazioni di rischi esistenti**, come ad esempio:

- Disponibilità dei servizi LLM
- Fuga di informazioni
- Abuso da parte degli agenti
- Modifiche non controllate dei modelli

### Catalogo delle cause

Vengono incorporate nuove **root causes specifiche dell'IA:**

- Shadow AI
- Iniezione di prompt
- Allucinazioni
- Avvelenamento dei dati
- Autorizzazioni eccessive

### Catalogo delle conseg

Le conseguenze tradizionali si ampliano con **impatti specifici**:

- Nuove sanzioni (ad es. AI Act)
- Perdite dovute a reinvestimenti (rischio strategico)
- Rischio reputazionale specifico
- Impatto sul lavoro e sull'organizzazione

### Asset

**Il perimetro degli asset si amplia:**

- Modelli LLM
- Prompt
- Dataset e informazioni non strutturate
- API
- Server MCP
- Agenti
- Strumenti con IA integrata

### Obiettivi di controllo

Sono necessari **nuovi controlli o adattamenti di quelli esistenti:**

- AI Gateway
- Tracciabilità
- Test avversarial
- Fallback
- Controllo dei fornitori di IA
- Gestione dei prompt come risorsa controllata

# 2 La nostra visione

## Catalogo dei rischi ICT e delle cause profonde nell'IA generativa

**I rischi ICT associati all'IA continuano a essere classificati allo stesso modo degli altri sistemi: disponibilità e continuità, sicurezza, gestione del cambiamento, gestione dei terzi e strategia. Tuttavia, all'interno di ciascun rischio emergono nuove root causes e altre si intensificano. L'identificazione di queste cause è il primo passo verso un'efficace mitigazione**

### Disponibilità e continuità

- Disponibilità del sistema di IA e delle comunicazioni associate
- Capacità e prestazioni. Rischio di sovraccarico
- Latenza. Tempi di risposta eccessivi rispetto alle esigenze del caso d'uso (ad es. frodi)
- Dipendenze non identificate (Shadow AI)

### Sicurezza

- Fuga di informazioni ed esposizione di dati sensibili
- Prompt injection diretta o indiretta
- Carenze nell'IAM e autorizzazioni eccessive
- Attacchi avversari
- Deepfake
- Data poisoning

### Terze Parti

- Dipendenza tecnologica / Vendor lock-in
- Rischio di concentrazione
- Rischi nella catena di approvvigionamento (quarte parti)
- Rischio contrattuale e di conformità normativa (es.: clausole DORA)
- Sovranità e localizzazione dei dati
- Rischio reputazionale associato al fornitore

### Gestione del cambiamento

- Modifiche unilaterali al modello senza strategia di versioning o rollback
- Manutenzione e aggiornamenti non pianificati
- Compatibilità nelle integrazioni
- Obsolescenza del modello

### Strategia

- Disallineamento con la strategia aziendale (business/IT)
- Competitività e tempi di adozione
- Conformità normativa e allineamento con la propensione al rischio
- Ritorno sull'investimento basso o discutibile
- Capacità interne insufficienti per l'adozione e la supervisione
- Rischio geopolitico e sovranità tecnologica

# 2 La nostra visione
## Famiglie di controlli

**I controlli tradizionali rimangono necessari, ma devono essere adattati e integrati per coprire i rischi relativi a dati, prompt, connettori, terze parti e funzionamento dell'IA**

### Inventario e classificazione

- Identificazione delle soluzioni di IA esistenti, dei loro utilizzi, della loro criticità e dei responsabili.

*Esempi: inventario dei casi d'uso, classificazione e mappatura delle dipendenze con le risorse tecnologiche e i terzi.*

### Dati, autorizzazioni e flussi

- Analisi e progettazione del controllo degli accessi e della circolazione delle informazioni tra utenti, modelli, connettori e strumenti.

*Esempi: definizione delle autorizzazioni sulle fonti, protezione dei dati sensibili e revisione dei flussi verso LLM*

### Architettura e funzionamento

- Rafforzamento dell'architettura con barriere e tracciabilità operativa per ridurre gli abusi e l'esposizione.

*Esempi: AI Gateway, registrazione e monitoraggio, filtraggio dei contenuti, crittografia e tracciabilità..*

### Sviluppo e cambiamento

- Adattamento del ciclo di sviluppo sicuro ai riferimenti di mercato e rafforzamento del controllo delle modifiche su modelli, prompt, connettori.

*Esempi: versioning, validazione preliminare e valutazione prima del passaggio alla produzione.*

### Strategia di test

- Convalida del sistema contro attacchi, errori e degrado, compresa la preparazione della risposta operativa.

*Esempi: test di prompt injection, test di resilienza e latenza e meccanismi di fallback.*

### Terze parti e catena di fornitura

- Estensione del controllo a fornitori, servizi LLM e strumenti con IA integrata per ridurre l'opacità e la dipendenza.

*Esempi: due diligence, clausole specifiche e monitoraggio della catena di subappalto.*

# 3 Proposta di valore

## Gestione dei rischi ICT nei sistemi di IA generativa

**Management Solutions combina competenze in materia di gestione dei rischi ICT, sicurezza informatica, resilienza, terze parti, regolamentazione e rischio di modello per aiutare le organizzazioni a implementare l'IA generativa in modo controllato, verificabile e in linea con il proprio framework di governance e la normativa.**

### 01 Identificazione delle aree di miglioramento

Valutazione della maturità e delle lacune rispetto ai quadri di riferimento, identificando le priorità in materia di adeguamento, governance e controllo.

*Progetto della durata di 4-6 settimane*

### 02 Aggiornamento del Framework di Governance

Aggiornamento delle:

- Politiche
- Tassonomie di rischi, cause e asset,
- KRI
- Metodologie
- Obiettivi di controllo per gestire attivamente i rischi ICT associati ai sistemi di IA

*Progetto della durata di 8-12 settimane*

### 03 Analisi e test sui casi d'uso


Analisi specialistica dei rischi ed esecuzione di test su progetti e casi d'uso:

- Revisione della progettazione
- Analisi dei controlli implementati
- Test di sicurezza e resilienza
- Elaborazione di conclusioni e piani di mitigazione
- Assistenza nell'implementazione

*Progetto della durata di 3-5 settimane per ogni caso d'uso*

# Allegati

# A1 Sviluppo sicuro dei sistemi di IA

## Requisiti e principi fondamentali

**La norma ETSI EN 304 223 definisce i requisiti e i principi fondamentali di sicurezza dell'IA per sviluppatori, operatori di sistemi e responsabili del trattamento dei dati, e copre la progettazione, lo sviluppo, l'implementazione, la manutenzione e la gestione del fine vita dei sistemi di IA in modo sicuro**

### Contesto

- La **rapida evoluzione delle tecnologie di IA** e il **conseguente aumento delle minacce informatiche** evidenziano la necessità di **norme in materia di sicurezza informatica** che garantiscano **una protezione coerente** e **un'armonizzazione normativa** in tutta **l'UE**.
- **L'ETSI ha pubblicato** la prima **norma europea** di applicazione mondiale che stabilisce **i requisiti di base in materia di sicurezza informatica** per **i modelli e i sistemi di IA**.
- La norma stabilisce un **quadro basato sul ciclo di vita** con **13 principi** che coprono **la progettazione**, **lo sviluppo**, **l'implementazione**, **la manutenzione** e **la fine del ciclo di vita in sicurezza**, e mira **ad aiutare le parti interessate lungo tutta la catena di fornitura dell'IA** a rafforzare **la sicurezza dell'IA** contro minacce emergenti, come **l'avvelenamento dei dati** e **la manipolazione dei modelli** (vedi allegato).

### Ambito

- Si applica a **sviluppatori**, **operatori di sistemi**, **data custodians**, **utenti finali** ed **entità interessate** (vedi alle gato).
- Comprende **la creazione**, **l'implementazione**, **il funzionamento**, **l'uso** e **l'impatto dei sistemi di IA**, compresi **i modelli open source** e **quelli proprietari**.

### Prossimi passi

- **31/03/2026**. Termine ultimo entro il quale tutte le organizzazioni nazionali di normazione (ONN) devono annunciare pubblicamente l'esistenza della norma ETSI EN 304 223.
- **30/09/2026**. Termine ultimo entro il quale le organizzazioni nazionali di normazione devono pubblicare o approvare formalmente la norma ETSI EN 304 223.
- **30/09/2026**. Data di ritiro di qualsiasi norma nazionale in conflitto con la presente norma.

### Principi e disposizioni di sicurezza dell'IA

| | |
|---|---|
| Progettazione sicura | Per **sviluppatori**, **operatori di sistemi** e **data custodians**, include disposizioni quali **formazione sulla sicurezza dell'IA**, **valutazione e mitigazione dei rischi per la sicurezza**, **documentazione della progettazione del sistema e utilizzo dei dati.** |
| Sviluppo sicuro | Per **sviluppatori**, **operatori di sistemi** e **data custodians**, include disposizioni quali la tenuta **di inventari delle risorse**, **i controlli di accesso**, **la supervisione dei processi sicuri**, la documentazione **dei registri di audit** e **l'esecuzione di test e valutazioni dei modelli di IA**. |
| Implementazione sicura | Stabilisce requisiti di comunicazione per **gli utenti finali** e **le entità interessate**, comprese **linee guida chiare**, **trasparenza nell'uso dei dati**, **divulgazione delle limitazioni**, **aggiornamenti di sicurezza tempestivi** e richiede **processi documentati** e **contrattuali.** |
| Manutenzione sicura | Richiede **agli sviluppatori** e **agli operatori di sistemi** di pubblicare **patch di sicurezza**, testare aggiornamenti importanti come **nuovi modelli**, supportare **la gestione del cambiamento** e monitorare **registri**, **prestazioni** e **anomalie** per garantire **la sicurezza continua del sistema**. |
| Fine del ciclo di vita sicuro | Obbliga **gli sviluppatori** e **gli operatori di sistema** a **eliminare in modo sicuro tutte le risorse** e **le configurazioni** durante **il trasferimento** o **lo smantellamento del modello** per evitare **rischi residui per la sicurezza.** |

Sviluppatori | Operatori di sistema | Data Custodians | Utenti finali | Entità interessate

# A1 Sviluppo sicuro dei sistemi di IA

## Progettazione sicura

**La sezione «Progettazione sicura», composta da quattro principi, stabilisce le misure di sicurezza fondamentali che le organizzazioni devono seguire per garantire che i sistemi di IA siano sicuri, affidabili e supervisionati da personale qualificato. Richiede alle organizzazioni di sensibilizzare i propri dipendenti, progettare in modo sicuro, gestire i rischi in modo continuo e integrare la responsabilità umana lungo l'intero ciclo di vita dell'IA.**

| | |
|---|---|
| **Principio 1: Sensibilizzare sulle minacce e sui rischi per la sicurezza dell'IA** | • Integrare **contenuti specifici sulla sicurezza dell'IA**, aggiornati periodicamente, **nella formazione sulla sicurezza informatica** dell'organizzazione, adattati **alle responsabilità** di ciascuna posizione.<br>• Garantire che tutto il personale sia informato **sulle minacce** emergenti **relative all'IA**, **sulle vulnerabilità** e **sulle misure di mitigazione** disponibili attraverso **la comunicazione multicanale**.<br>• Fornire agli sviluppatori una formazione specializzata sulla **codifica sicura dell'IA**, **sulla progettazione dei sistemi** e sulle tecniche per prevenire **le vulnerabilità** nei **modelli**, negli **algoritmi** e **nel software** di supporto. |
| **Principio 2: Progettare il sistema di IA tenendo conto della sicurezza, oltre che della funzionalità e delle prestazioni** | • Individuare e documentare **i requisiti**, **i rischi per la sicurezza dell'IA** e **le strategie di mitigazione**, coinvolgendo i **Data Custodians**, ove pertinente.<br>• Progettare sistemi resistenti ad **attacchi avversari**, **input imprevisti** e guasti, supportati da **registri di audit** chiari per modelli, set di dati e prompt.<br>• Effettuare **valutazioni dei rischi specifiche per l'IA** e **la due diligence** per i componenti/fornitori esterni, garantendo **controlli** adeguati **sulla sensibilità dei dati** e **permessi minimi** quando si interagisce con altri sistemi. |
| **Principio 3: Valutare le minacce e gestire i rischi per il sistema di IA.** | • Effettuare continuamente **modelli di minacce** e **gestione dei rischi**, affrontando **gli attacchi** e i rischi **specifici dell'IA** derivanti da nuove **configurazioni** o **funzionalità** non necessarie **dei modelli**.<br>• Applicare **controlli** adeguati basati **sulla tolleranza al rischio e** comunicare **le minacce IA** irrisolte **agli operatori di sistema** e **agli utenti finali,** indicando chiaramente **le ripercussioni** e **le misure raccomandate**.<br>• Mantenere **una supervisione** continua, garantire che le parti esterne siano in grado di gestire **i rischi di sicurezza dell'IA** e riconoscere che anche dopo la mitigazione permane **un rischio residuo**. |
| **Principio 4: Abilitare la responsabilità umana per i sistemi di IA** | • Garantire che **la supervisione umana** sia integrata nella progettazione del sistema di IA, consentendo agli esseri umani **di valutare**, **interpretare** e **spiegare** facilmente i risultati del modello.<br>• Implementare e mantenere **misure tecniche** a supporto della supervisione come **controllo dei rischi e** verificare che **i controlli di sicurezza specificati dal Data Custodian** siano correttamente integrati.<br>• Informare **gli utenti finali** sui **casi d'uso vietati** per evitare un uso improprio e rafforzare il funzionamento sicuro e conforme del sistema. |

# A1 Sviluppo sicuro dei sistemi di IA

## Sviluppo sicuro

**La sezione «Sviluppo sicuro», composta da cinque principi, stabilisce i requisiti fondamentali che regolano lo sviluppo sicuro dell'IA. Questi principi guidano le organizzazioni nella protezione delle loro risorse, nel rafforzamento dell'infrastruttura, nella protezione della catena di approvvigionamento, nella documentazione dei sistemi e nell'esecuzione di test approfonditi.**

| | |
|---|---|
| **Principio 5: Identificare, tracciare e proteggere gli assest**  | • Mantenere un **inventario completo di tutti gli asset e documentarne le interdipendenze** per garantire la completa visibilità e tracciabilità dell'intero sistema di IA.<br>• Utilizzare processi e strumenti robusti per **tracciare, autenticare e implementare il controllo** delle **versioni** degli asset specifici dell'IA durante il loro ciclo di vita.<br>• Proteggere i dati riservati applicando **rigorosi controlli di convalida, misure di sanificazione e controlli di sicurezza** adeguati al livello di riservatezza. |
| **Principio 6: Proteggere l'infrastruttura**  | • Rafforzare **i quadri di controllo degli accessi** per API, modelli, dati e processi, al fine di garantire che solo le entità autorizzate possano interagire con i componenti critici dell'IA.<br>• Applicare **controlli di sicurezza** robusti **alle API** per mitigare rischi quali il reverse engineering e il model poisoning.<br>• Utilizzare **ambienti dedicati e isolati** per ridurre il rischio di azioni non autorizzate e movimenti laterali. |
| **Principio 7: Garantire la supply chain**  | • Seguire **pratiche sicure nella catena di fornitura del software** per garantire l'integrità e l'affidabilità di tutti i componenti.<br>• **Giustificare l'uso** di qualsiasi componente non documentato o non sicuro effettuando valutazioni formali dei rischi e implementando adeguati controlli di mitigazione.<br>• **Rieseguire le valutazioni** dei modelli rilasciati per convalidarne la sicurezza e le prestazioni continue e comunicare chiaramente eventuali modifiche agli utenti finali. |
| **Principio 8: Documentare dati, modelli e prompt**  | • Mantenere **una documentazione completa** e un **registro di audit** chiaro che copra le decisioni di progettazione del sistema, le attività di sviluppo e le azioni di manutenzione continua.<br>• Includere tutte **le informazioni rilevanti per la sicurezza**, quali le fonti dei dati di formazione, le limitazioni note, le barriere di protezione e qualsiasi restrizione che influisca sul funzionamento sicuro.<br>• Pubblicare **gli hash crittografici** degli artefatti del modello e documentare qualsiasi acquisizione di dati pubblici, inclusi la **fonte, l'URL e la data** di recupero. |
| **Principio 9: Effettuare test e valutazioni adeguati**  | • Esegui **test approfonditi** di **valutazione della sicurezza** dei modelli, delle applicazioni e dei sistemi prima del loro rilascio per garantire che soddisfino gli standard di sicurezza richiesti.<br>• Eseguire **test prima dell'implementazione**, preferibilmente con **valutatori di sicurezza indipendenti**, per ottenere una valutazione obiettiva delle possibili vulnerabilità.<br>• Valutare **i risultati dei modelli** per identificare e prevenire rischi quali il reverse engineering, la manipolazione o l'influenza indesiderata sul comportamento dei modelli. |

# A1 Sviluppo sicuro dei sistemi di IA

## Implementazione sicura, Manutenzione sicura e Ritiro sicuro

**Queste tre sezioni —Implementazione sicura, Manutenzione sicura e Ritiro sicuro— stabiliscono i requisiti per implementare, mantenere e ritirare i sistemi di IA in modo sicuro e controllato. Proteggono i dati, preservano l'integrità del sistema e prevengono i rischi per la sicurezza**

**Principio 10: Comunicazione e processi relativi agli utenti finali e alle entità interessate**

- Informare gli utenti finali **su dove e come vengono utilizzati, consultati e archiviati i loro dati**, compreso l'uso per **il riaddestramento dei modelli** o **la revisione umana**.
- Fornire **indicazioni accessibili sull'uso, la gestione, l'integrazione e la configurazione** del sistema, evidenziando **i limiti** e **le possibili modalità di guasto**, e comunicare in modo proattivo **gli aggiornamenti rilevanti per la sicurezza**.
- Fornire supporto **agli utenti finali e alle entità interessate** durante e dopo **gli incidenti di sicurezza informatica**, seguendo un processo **documentato e** concordato nei contratti.

**Principio 11: Mantenere aggiornamenti di sicurezza, patch e mitigazioni periodici**

- Fornire **aggiornamenti e patch di sicurezza**, informare gli operatori di sistema e garantire che gli aggiornamenti vengano consegnati agli utenti finali.
- Mantenere **meccanismi e piani di emergenza** per **mitigare i rischi per la sicurezza** quando non è possibile fornire aggiornamenti.
- Trattare **gli aggiornamenti** importanti **del sistema di IA** come nuove versioni del modello ed eseguire **test e valutazioni di sicurezza**.
- Supportare gli operatori del sistema **nella valutazione e nella risposta ai cambiamenti del modello**, compreso **l'accesso anticipato** tramite beta test o **API versionate**.

**Principio 12: Monitorare il comportamento del sistema**

- Registrare **le azioni del sistema e degli utenti** per supportare **la conformità alle norme di sicurezza**, **l'indagine sugli incidenti** e **la correzione delle vulnerabilità**.
- Analizzare **i registri** per rilevare **anomalie**, **violazioni della sicurezza**, **comportamenti imprevisti** o problemi quali **la deriva dei dati** o **l'avvelenamento dei dati**.
- Monitorare **gli stati interni** dei sistemi di IA quando ciò migliora la capacità di affrontare **le minacce alla sicurezza** o consente future **analisi di sicurezza**.
- Monitorare **le prestazioni del modello e del sistema** nel tempo per identificare **cambiamenti improvvisi o graduali** che potrebbero influire sulla sicurezza.

**Principio 13: Garantire la corretta eliminazione dei dati e dei modelli**

- Quando trasferiscono o condividono **dati di addestramento** o un **modello**, gli sviluppatori e gli operatori del sistema devono coinvolgere **i data custodians** e eliminare in modo **sicuro** le risorse pertinenti per evitare che i problemi di sicurezza si diffondano tra le istanze del sistema di IA.
- Quando **si smantella** un modello o un sistema, è necessario coinvolgere **i data custodians** e garantire la **cancellazione sicura** dei dati e dei dettagli di configurazione applicabili.

# A2 Test dei sistemi di IA

## Principi chiave

**La sicurezza di un sistema LLM non dipende solo dal modello, ma anche da come tale modello si comporta all'interno di un'applicazione, con quali dati opera e quali decisioni o azioni innesca nel processo aziendale supportato.**

### Comportamento del modello

Il modello può essere indotto a ignorare le istruzioni, a dare priorità a contesti dannosi o a produrre risposte incoerenti. In caso di frode, ciò può favorire blocchi indebiti o bypass; in termini di benefici, risposte errate o non applicabili.

- Prompt injection e jailbreak
- Risposta non deterministica
- Sensibilità a input ambigui
- Allucinazioni con impatto operativo

### Applicazione, dati e integrazioni

L'applicazione degli LLM introduce rischi in nuove risorse, come prompt, connettori, recuperatori, regole di business e trattamento degli output. La fuga della logica antifrode applicata, dei segnali di scoring, dei dati dei clienti o l'ottenimento di profitti non autorizzati sono rischi da verificare.

- RAG o fonti contaminate
- Gestione non sicura degli output
- Esposizione di prompt e regole
- Abuso di strumenti e integrazioni

### Processo, canale e operazione

Quando il sistema si integra con canali, identità e operazioni, il profilo di rischio cambia, poiché un guasto non è più solo tecnico: influisce sull'esperienza del cliente e sulla continuità, tracciabilità e capacità di risposta della banca.

- Errori di autorizzazione
- Blocchi indebiti o bypass
- Degradazione del servizio
- Impatto sulla reputazione e normativo

# A2 Test dei sistemi di IA

## Riferimenti e migliori pratiche di mercato

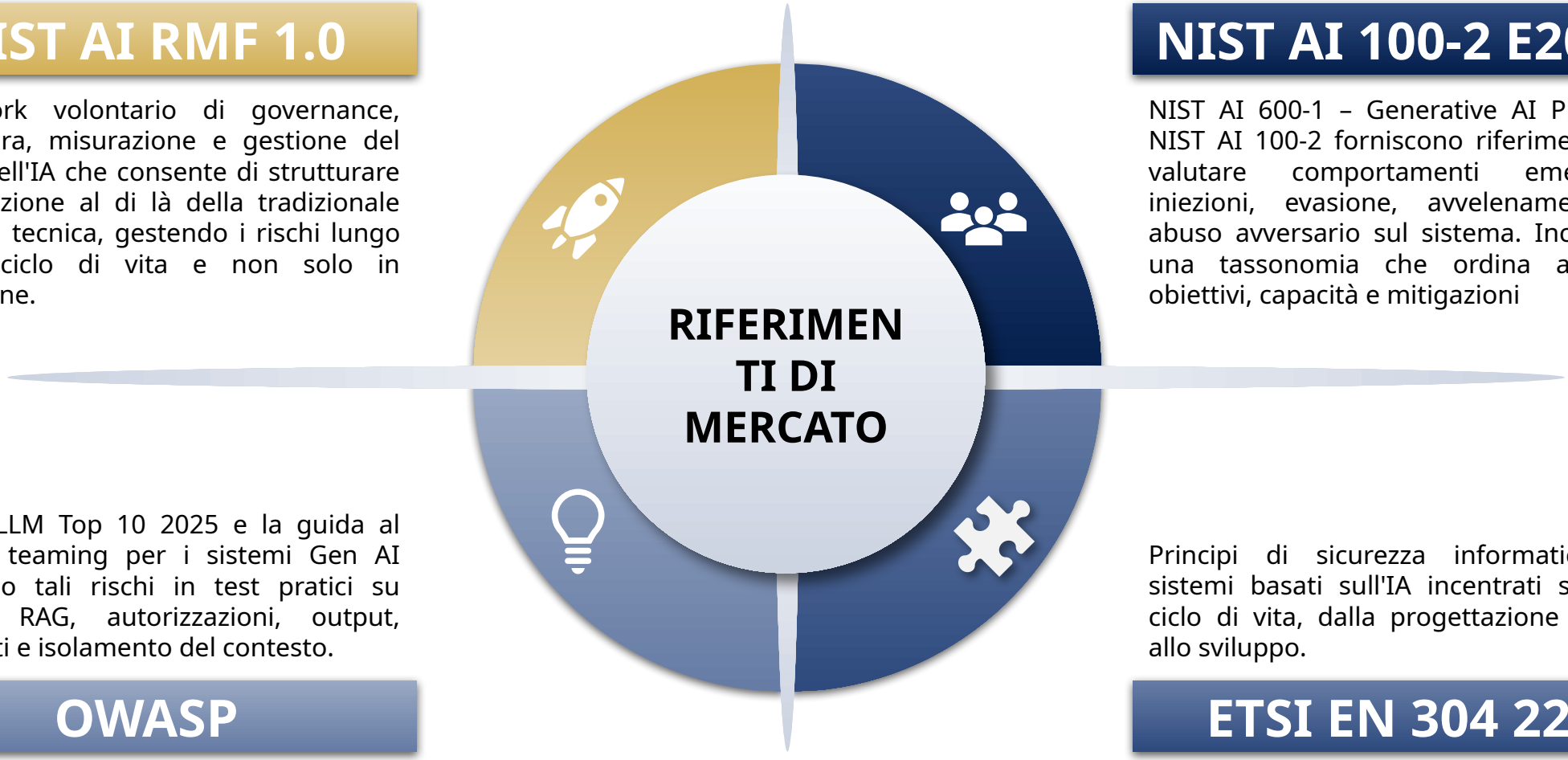
**La definizione di una strategia di test completa deve tenere conto dei riferimenti più rilevanti del settore.**

### NIST AI RMF 1.0

Framework volontario di governance, mappatura, misurazione e gestione del rischio dell'IA che consente di strutturare la valutazione al di là della tradizionale sicurezza tecnica, gestendo i rischi lungo l'intero ciclo di vita e non solo in produzione.

### NIST AI 100-2 E2025

NIST AI 600-1 – Generative AI Profile e NIST AI 100-2 forniscono riferimenti per valutare comportamenti emergenti, iniezioni, evasione, avvelenamento e abuso avversario sul sistema. Includono una tassonomia che ordina attacchi, obiettivi, capacità e mitigazioni




OWASP LLM Top 10 2025 e la guida al network teaming per i sistemi Gen AI traducono tali rischi in test pratici su prompt, RAG, autorizzazioni, output, strumenti e isolamento del contesto.

### OWASP

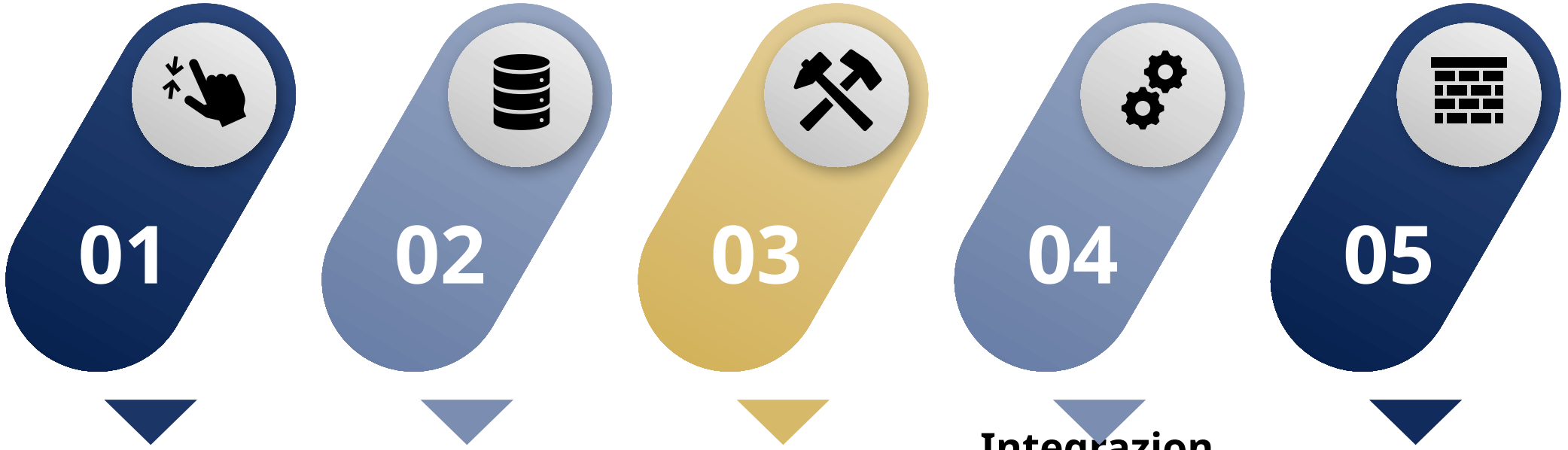
Principi di sicurezza informatica nei sistemi basati sull'IA incentrati sul loro ciclo di vita, dalla progettazione iniziale allo sviluppo.

### ETSI EN 304 223

# A2 Test dei sistemi di IA

## Categorie di test

**Tenendo conto delle fonti citate, Management Solutions definisce cinque famiglie di test che coprono sia la materializzazione dell'abuso sia la convalida delle barriere tecniche e operative.**




### Interazione

Iniezione di prompt, jailbreak, fuga dal prompt di sistema e aggiramento delle politiche. Rif.: OWASP LLM01 e LLM07.

### Dati

Divulgazione, avvelenamento, catena di approvvigionamento e messa a terra difettosa. Rif.: OWASP LLM02-04; NIST AI 100-2e2025.

### Strumenti

Eccessiva agenzia, abuso di strumenti, connettori e API e fallimenti di autorizzazione. Rif.: OWASP LLM06; Playbook.

### Integrazione

Gestione impropria dell'output e propagazione a canali, regole e processi. Rif.: OWASP LLM05; NIST AI 600-1.

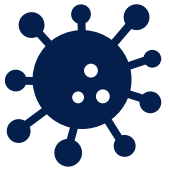
### Resilienza

Monitoraggio, fallback, continuità e risposta agli incidenti. Rif.: AI RMF, Playbook ed ETSI.

# A2 Test dei sistemi di IA

## Interazione avversaria

**Nei sistemi con IA generativa, il primo vettore di compromissione è solitamente l'interazione stessa. Non è sufficiente analizzare input dannosi isolati, ma è necessario verificare se il sistema può essere deviato, ricontestualizzato o indotto a disobbedire alle proprie restrizioni.**



### Iniezione diretta

Inserire istruzioni utente che entrino in competizione con quelle del sistema per modificare il comportamento del modello.

### Jailbreak

Costringere il modello a disobbedire a restrizioni o guardrail tramite formulazioni avverse.

### Elusione delle politiche

Ottenere un risultato proibito senza richiederlo esplicitamente, tramite riformulazione, offuscamento, frammentazione o cambio di lingua/canale.



### Manipolazione del contesto o della memoria

Sfruttare la cronologia, la memoria o lo stato della conversazione per far persistere istruzioni dannose o influenzare le decisioni successive.

# A2 Test dei sistemi di IA

## Dati, recupero ed esposizione

**La sicurezza del sistema dipende anche da quali informazioni recupera, come le interpreta e cosa finisce per esporre. Un sistema può fallire se si basa su dati contaminati, mal prioritizzati o non autorizzati.**

### ESPOSIZIONE DI PROMPT E LOGICA INTERNA

Rivelare istruzioni di sistema, criteri di filtraggio, ruoli, limiti operativi o segreti incorporati nel prompt.

### AVVELENAMENTO DEL CORPUS O DELLE FONTI

Alterare documenti, basi di conoscenza o contenuti indicizzati utilizzati dal sistema per influenzare il processo decisionale.

### MANIPOLAZIONE DEI RISULTATI DI RICERCA, DEI METADATI O DEGLI EMBEDDING

Sfruttare il ranking, i metadati, le somiglianze vettoriali o la prioritizzazione dei documenti per favorire contenuti errati o non autorizzati.

### FUGA DI INFORMAZIONI SENSIBILI

Estrarre informazioni personali, dati dell'account, punteggi, regole interne o benefici non applicabili tramite interazione o recupero improprio.

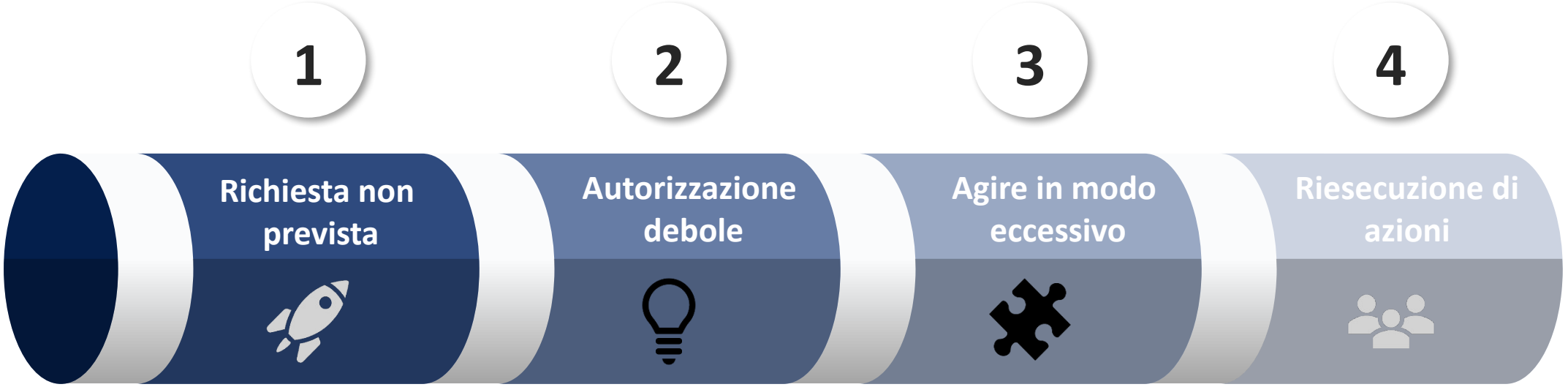

### CATENA DI FORNITURA

Verificare l'effetto di modifiche o vulnerabilità su modelli, connettori, librerie, indici e servizi di terze parti.

# A2 Test dei sistemi di IA

## Strumenti, azioni e autorizzazioni

**Il rischio aumenta in modo non lineare quando l'LLM smette di limitarsi a rispondere e passa a richiamare strumenti, consultare sistemi o innescare azioni. In questi casi, la sicurezza della pipeline dipende sia dall'autorizzazione che dal contenimento dei sistemi con cui si integra.**




Indurre chiamate a strumenti falsi, errati o al di fuori del perimetro funzionale previsto.

Verificare quale identità, ruolo, titolarità dell'account e contesto transazionale vengono convalidati prima di ogni azione.

Valutare se il sistema concatena piani in più fasi, prende decisioni o esegue azioni che vanno oltre il mandato ricevuto.

Ripetere, riordinare o concatenare azioni valide per generare effetti indesiderati.

# A2 Test dei sistemi di IA

## Integrazione e trattamento degli output

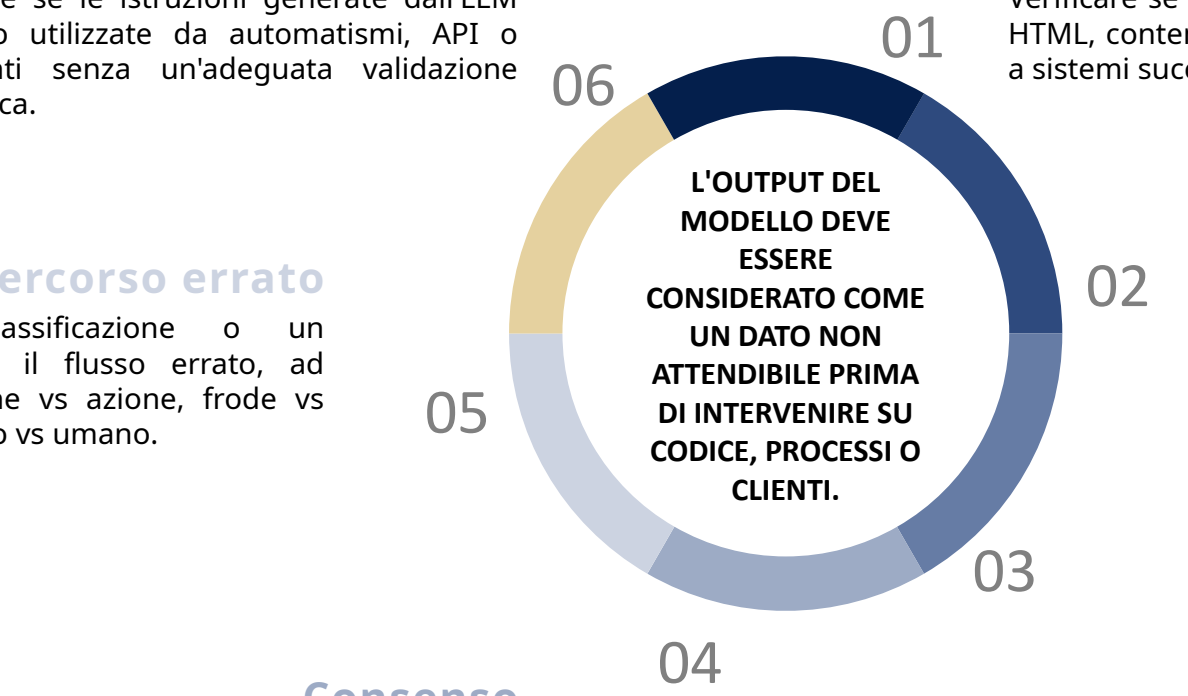
**L'impatto maggiore non si verifica solitamente nella generazione della risposta, ma nella sua propagazione. Un output non sicuro può generare comportamenti indesiderati alimentando canali, regole di business o azioni successive senza una validazione sufficiente.**

### Comando a valle

Verificare se le istruzioni generate dall'LLM vengono utilizzate da automatismi, API o strumenti senza un'adeguata validazione semantica.

### Payload attivo

Verificare se l'output dell'LLM può propagare HTML, contenuti attivi, script, SQL o comandi a sistemi successivi.




### Percorso errato

Provocare una classificazione o un instradamento verso il flusso errato, ad esempio consultazione vs azione, frode vs assistenza, automatico vs umano.

### Regola di business

Rafforzare soglie, limiti, eccezioni e condizioni di idoneità per verificare se l'output del modello altera la logica applicata.

### Consenso

Verificare che qualsiasi utilizzo dei dati, raccomandazione o azione sull'account rispetti il consenso, la legittimazione e il contesto.

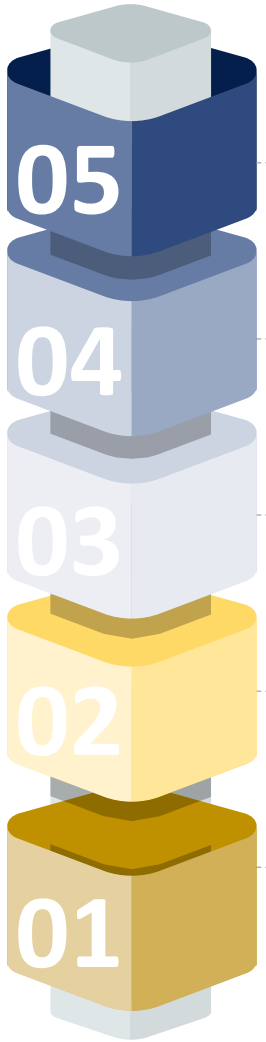
### Canale del cliente

Misurare come l'output si concretizza nell'app, sul web, su WhatsApp o nella comunicazione vocale e se raggiunge il cliente con controlli di verifica e tracciabilità.

# A2 Test dei sistemi di IA

## Resilienza, consumo e governance

**Un sistema LLM è sicuro solo se è anche governabile: la robustezza tecnica deve essere integrata da monitoraggio, fallback, controllo dei consumi, risposta agli incidenti e supervisione umana efficace**




### Modifiche

Monitorare prompt, modelli, indici, dipendenze e release per individuare deviazioni e regressioni di sicurezza.

### Continuità e consumo

Simulare interruzioni del fornitore, latenza, quota, loop di inferenza o consumo eccessivo di risorse.

### Monitoraggio

Verificare la copertura di prompt, strumenti, recupero, rifiuti, anomalie ed eventi critici nel monitoraggio del sistema.

### Incidenti

Verificare la classificazione, il contenimento, l'escalation e la comunicazione in caso di anomalie o comportamenti anomali del sistema.

### Fallback

Testare il fallback sicuro, i limiti operativi e il passaggio alla revisione umana quando il sistema non deve decidere da solo.

# Management Solutions

*Making things happen*

**International**
***One Firm***

**Multidisciplinary Team**

**Best practices**
***know-how***

**Proven experience**

**Maximum commitment**

**Per maggiori informazioni, visitate il nostro sito:**

www.managementsolutions.com

O i nostri social networks: