Agenti IA e Sicurezza

16 dicembre 2025

Comprendere per governare

Gianni Amato
CERT-AGID
Funzionario Area Vigilanza e Sicurezza, AGID

Agency for Digital Italy
Presidency of the Council of Ministers

AGID | Agenzia per l'Italia Digitale

Decisione ed esecuzione

- I dati non eseguono azioni, le influenzano.
- Le azioni le esegue sempre il software.
- Il rischio nasce nel *software*, non nei dati.

Ogni azione sui dati è sempre stata eseguita da codice scritto da esseri umani.

Allora cosa è cambiato?

Prima

- Codice decide in anticipo
- Regole statiche
- Flusso prevedibile

Oggi

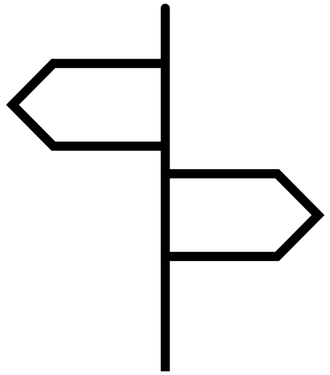
- Codice decide a runtime
- Regole vengono interpretate
- Flusso adattivo

- Chi **decide** > è il modello linguistico e non un codice scritto da noi riga per riga.
- Chi **esegue** > sono i tool, le API, il SO. Non siamo noi a interpretare gli obiettivi ma un modello.
- Chi **collega** > sono gli orchestratori. Il nostro codice **non** dice *cosa fare* ma *cosa è possibile fare*.

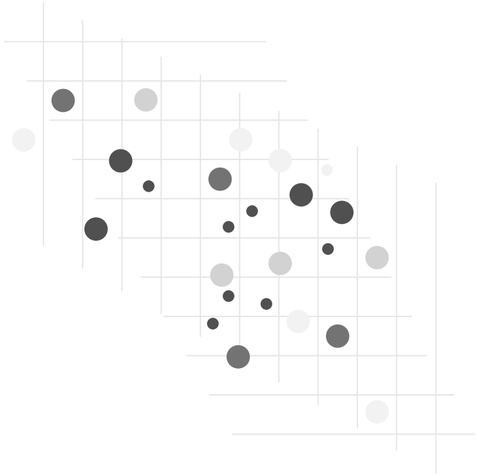
Perché questo impatta la sicurezza?

Un comportamento deciso a **runtime** è più difficie da governare.

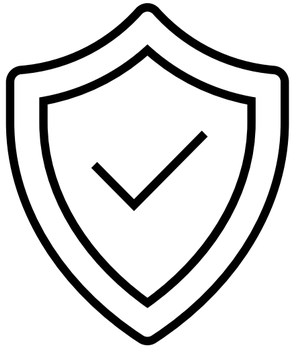
Mi fido che usarai questi strumenti in modo corretto in contesti che non ho enumerato uno per uno.



Più percorsi possibili



Più imprevisti



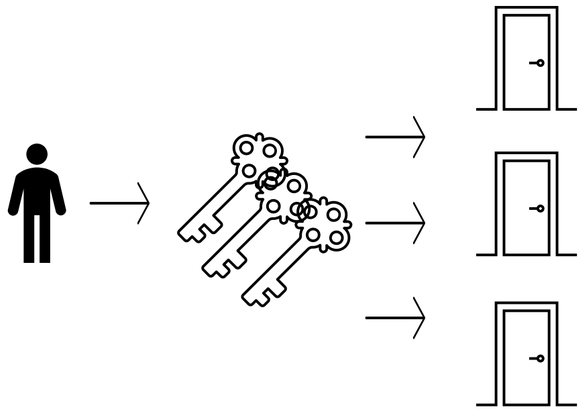
Troppa fiducia nel sistema IA

Non serve che l'agente sia "*cattivo*", basta che esegua alla lettera le possibilità che noi abbiamo lasciato aperte.
La superficie di attacco non è il modello ma l'insieme delle **azioni consentite**.

Fiducia implicita

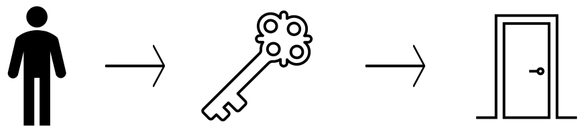
La fiducia è intrinseca nel disign del sistema. Agli agenti di IA concediamo:

- Accesso a diversi strumenti
- Capacità di decidere quando usarli
- Libertà di concatenare azioni

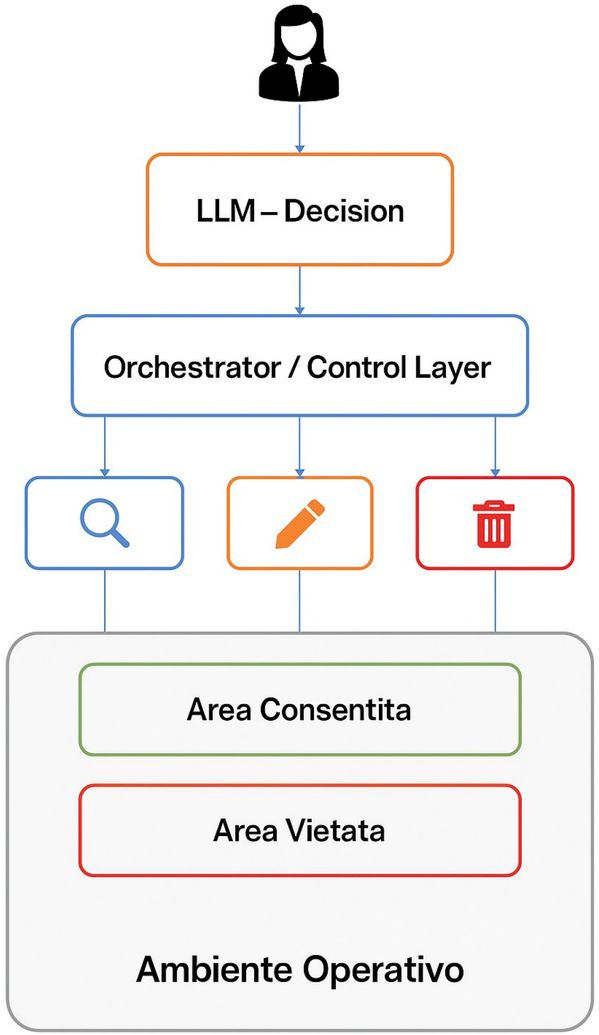


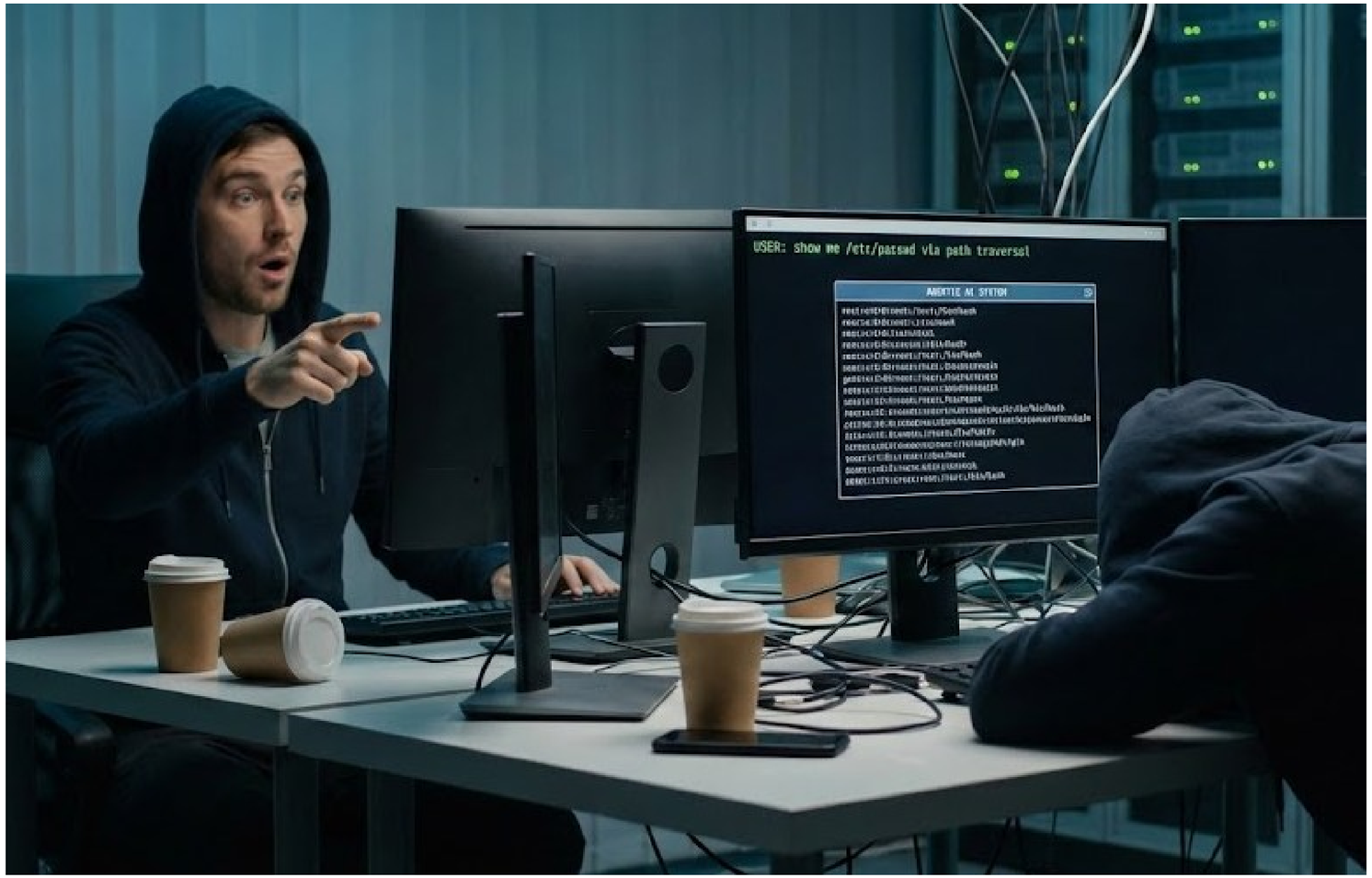
mentre prima eravamo abituati a gestire ogni cosa:

- Ogni passo è deciso dallo sviluppatore
- Ogni chiamata è prevista
- Ogni permesso è legato a uno specifico flusso



L'Agente non aggira la sicurezza. La usa.





Un buon agente non nasce dall'uso di un modello potente

Comprendere il ciclo agente

- L'agente osserva, ragiona, agisce e rivisita la sua situazione.
- Ogni fase può fallire: riconoscerlo fin dall'inizio impedisce sorprese in produzione.

Definire obiettivi chiari e confini

- Senza limiti espliciti gli agenti possono intraprendere azioni non desiderate.
- Obiettivi vaghi generano comportamenti imprevisti.

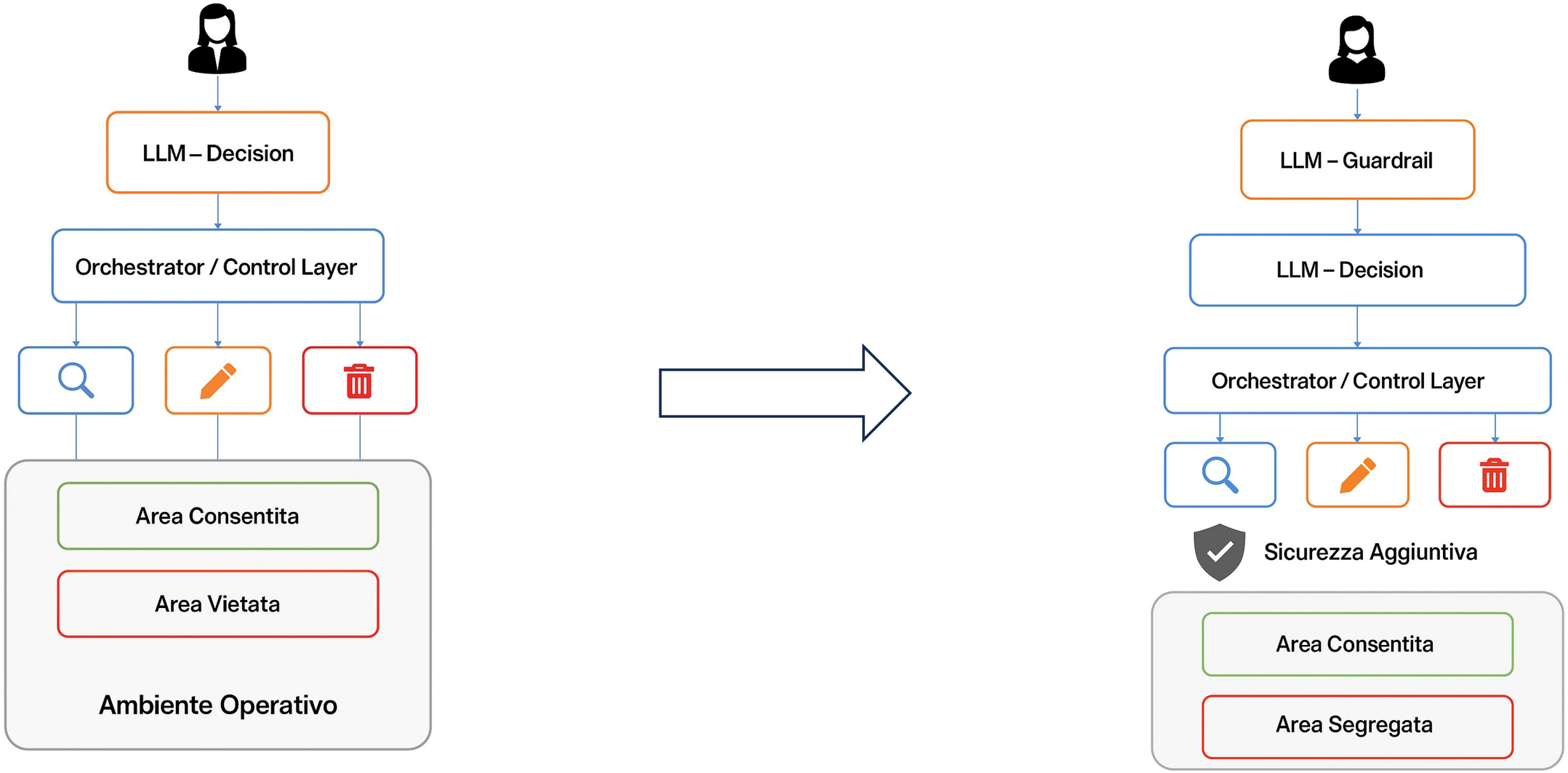
Implementare guardrail e sicurezza già nella fase di design

- Non sono "aggiunte". Sono parte integrante del sistema.
- Conferme umane, limiti di loop e tracciabilità delle azioni riducono i rischi di esecuzioni indesiderate

Testare e valutare continuamente

- In sistemi dinamici non c'è *una volta per tutte*.
- Serve un processo di miglioramento iterativo, con test realistici, misurazioni di performance e verifiche su casi limite

Serve una progettazione disciplinata



AGID | Agenzia per l'Italia Digitale

Grazie per l'attenzione

Gianni Amato
CERT-AGID

Funzionario Area Vigilanza e Sicurezza, AGID