ZENITA GROUP

AUTONOMIA DELEGATA

"La prima generazione di sistemi digitali dove l'organizzazione non sa più, momento per momento, cosa sta facendo"

ZENITA GROUP

AUTONOMIA DELEGATA

Come l'AI ridisegna posizionamento competitivo, organizzazioni agentiche ed evoluzione della sicurezza.

Pierguido Iezzi

Direttore Cyber e Strategie | Zenita Group

La ricerca mostra che il vero punto di svolta non è il chatbot, ma la competizione tra **infrastrutture AI**, la nascita di **aziende agentiche** e l'emersione di **rischi cyber** che i framework tradizionali non coprono in modo adeguato.

2
STRATEGIE GUIDA
OpenAI e Anthropic

3
CLUSTER ORGANIZZATIVI
Supervisor, Constitutionally-Bounded, Bifronte

9
NUOVI RISCHI CYBER AI
molti non coperti dai controlli attuali

1
CAMBIO DI PARADIGMA
la sicurezza si sposta verso dati, agenti, comportamento e continuità

> L'elemento decisivo non sarà usare l'AI, ma scegliere quale architettura di dipendenza, controllo e continuità costruire.

SOMMARIO

Executive Summary

L'intelligenza artificiale ha superato la fase sperimentale. L'analisi di 1.080 posizioni aperte nei due principali laboratori AI, integrata con dati su M&A, cicli di rilascio e investimenti infrastrutturali, evidenzia una transizione strutturale: l'AI non è più un insieme di strumenti separati ma un'infrastruttura operativa su cui aziende, governi e piattaforme stanno progressivamente trasferendo processi decisionali e funzioni critiche.

In questo scenario di discontinuità, l'organizzazione si trova a operare con sistemi in cui il sostrato esecutivo è statistico e non più deterministico, determinando una condizione senza precedenti: per la prima volta, le aziende rischiano di non sapere più, momento per momento, cosa stia facendo esattamente il proprio stack tecnologico.

Cosa sta succedendo e cosa cambia davvero:
I 5 INSIGHT FONDAMENTALI

1. Evoluzione del cloud e del mercato AI

Il contesto competitivo ha superato la fase duopolistica iniziale. Il mercato non è più governato da una logica duale (pochi player dominanti contrapposti a soluzioni minori), ma è evoluto in un **ecosistema multipolare stabile**, all'interno del quale operano attivamente circa 8–10 poli globali di frontiera (che includono player statunitensi, europei e il fronte cinese). In questa nuova configurazione, l'intera infrastruttura cloud sta cambiando natura, diventando intrinsecamente **AI-driven**, dove la capacità computazionale — e non il software — rappresenta la principale unità di costo e l'architettura portante del mercato.

2. Ridefinizione dei player AI

Le aziende che sviluppano AI di frontiera stanno attuando strategie di posizionamento radicalmente differenti, eliminando qualsiasi gerarchia lineare o chiara sul mercato. I dati occupazionali e le acquisizioni dimostrano una netta divergenza: da un lato si assiste al consolidamento di stack verticali integrati mirati alla cattura dell'intera filiera computazionale globale (come OpenAI); dall'altro, emergono architetture focalizzate esclusivamente sulla costruzione di una *trust infrastructure* per mercati regolamentati ad alta intensità di compliance e auditing (come Anthropic). Il contesto competitivo è frammentato, asimmetrico e governato da cicli di rilascio rapidissimi (6–8 settimane), che costringono le imprese a progettare workflow multi-provider per evitare rischi di lock-in tecnologico e geopolitico.

3. Nuove categorie di mercato e polarizzazione

La transizione verso modelli di esecuzione autonomi sta portando alla categorizzazione delle organizzazioni in tre macro-archetipi aziendali: l'impresa orchestrata da agenti (Supervisor-of-Agents), l'impresa a governance costituzionale (*Constitutionally-Bounded*) e l'impresa a doppio stack (*Dual-Stack Enterprise*). Questa frammentazione accentua un profondo divario economico e competitivo: da un lato, gli *early adopters* beneficiano di riduzioni dei costi operativi tra il 50% e l'80%; dall'altro, le aziende ancorate a modelli tradizionali si scontrano con il rischio sistemico della **AI poverty** (il divario generato dal mancato allineamento strutturale alle architetture native). Il mercato si sta polarizzando, traducendosi in asimmetrie di margine e capacità operativa difficilmente colmabili nel medio periodo.

4. Nuovi rischi cyber (AI-native risks)

L'introduzione di agenti AI dotati di memoria, budget di esecuzione e capacità di pianificazione genera una superficie di attacco completamente nuova e priva di equivalenti storici. La natura del rischio cyber subisce una mutazione ontologica: **non è più una minaccia esclusivamente infrastrutturale, ma diventa comportamentale e autonoma**. Esempi critici come la *Belief Injection* — la manipolazione persistente e statistica del comportamento cognitivo di un agente attraverso il RAG poisoning o la contaminazione del feedback — dimostrano che le minacce agiscono sull'evoluzione del sistema nel tempo e non sul singolo input isolato, aumentando esponenzialmente la complessità del controllo e della verifica causale.

5. Inadeguatezza dei modelli di cybersecurity attuali

I framework di sicurezza e gli strumenti di detection tradizionali (come SIEM, EDR o gli standard ISO 27001 e NIST) mostrano un gap strutturale difficilmente colmabile con gli approcci attuali: presuppongono ambienti deterministici e log atomici, mostrano limitazioni strutturali di fronte ai rischi emergenti e ai fenomeni di *Attribution Collapse* (l'impossibilità di ricostruire le catene decisionali nei workflow multi-agente). Per sopravvivere a questa transizione, la cybersecurity deve evolvere verso un modello radicalmente nuovo, basato su tre pilastri:

- **Tracciabilità dei processi cognitivi** attraverso strumenti dedicati come il *Model Bill of Materials*.
- **Controllo rigoroso delle decisioni automatizzate** e limitazione d'azione runtime tramite perimetri comportamentali (*Behavioral Envelope*).
- **Modelli di cybersecurity cognitiva**, in cui l'oggetto dell'analisi forense e della telemetria non è più il singolo evento di rete, ma l'intero workflow agentico e il comportamento emergente del sistema.

In conclusione, l'evidenza netta che emerge da questo report è che la cybersecurity sta evolvendo da un modello *infrastrutturale* a un modello *cognitivo* e *processuale*, in cui il controllo del comportamento dei sistemi AI diventa centrale.

Il report è strutturato in sette parti. Letti in sequenza, i capitoli costruiscono una catena causale: dai dati alle traiettorie evolutive del mercato; dalle traiettorie ai modelli organizzativi delle imprese; dalle trasformazioni aziendali alle nuove superfici di rischio; dai nuovi rischi alla ridefinizione della cybersecurity; dalla cybersecurity all'asimmetria geopolitica e industriale; dall'asimmetria agli scenari strategici concretamente decidibili da aziende, governi e operatori infrastrutturali.

L'obiettivo finale non è descrivere l'AI come fenomeno tecnologico isolato, ma interpretarla come nuova infrastruttura cognitiva dell'economia digitale — e tradurre questa trasformazione in framework decisionali utilizzabili da board, CISO, regulator, operatori infrastrutturali e stakeholder strategici.

Disclaimer per il lettore

Il presente white paper propone un'analisi strategica e prospettica delle principali trasformazioni in atto nell'ecosistema dell'intelligenza artificiale, con l'obiettivo di offrire chiavi di lettura utili alla comprensione delle possibili traiettorie evolutive e delle relative implicazioni tecnologiche, organizzative, di cybersecurity e geopolitiche.

Le evidenze e le interpretazioni riportate si basano sull'analisi di informazioni pubblicamente disponibili e di segnali osservabili di mercato, tra cui dinamiche occupazionali, investimenti, operazioni di M&A, comunicazioni aziendali, evoluzione delle architetture tecnologiche, strategie di prodotto e iniziative infrastrutturali. L'intento del documento non è fornire una mappatura esaustiva del mercato, bensì evidenziare pattern emergenti e potenziali punti di discontinuità.

Le aziende, le tecnologie e i casi citati sono selezionati in funzione della loro rilevanza rispetto ai fenomeni analizzati e della disponibilità di informazioni verificabili. La loro inclusione non implica alcuna valutazione comparativa, classificazione di mercato o giudizio qualitativo; analogamente, l'eventuale assenza di altri attori non ne riduce la rilevanza o la significatività.

Le analisi, le interpretazioni e gli scenari presentati riflettono esclusivamente le valutazioni degli autori, sviluppate sulla base delle informazioni disponibili al momento della redazione. In un contesto caratterizzato da elevata dinamicità tecnologica, forte competizione e continua evoluzione normativa e geopolitica, tali considerazioni devono essere intese come strumenti di lettura e supporto alla riflessione strategica, e non come previsioni vincolanti o definitive.

Il presente documento non costituisce consulenza finanziaria, legale, strategica o di investimento e non deve essere utilizzato come unico elemento a supporto di decisioni operative o di business.

5 KEY FINDINGS

ZENITA GROUP

Dati e segnali che ridefiniscono il mercato AI

01

670 posizioni in OpenAI **VS** **410** posizioni in Anthropic

BIFORCAZIONE STRATEGICA

OpenAI ⟶ Compute & Agenti

Anthropic ⟶ Trust & Security

02

6 - 8 settimane

NUOVA CADENZA INDUSTRIALE

I modelli frontier evolvono ormai come una pipline continua.

03

40 - 50x

ECONOMIA DEL COMPUTE

L'intensità capitale/ricavi supera di decine di volte il SaaS tradizionale.

04

4 vettori

BELIEF INJECTION

- RAG poising
- Fine-tuning compromesso
- RLHF contaminazione
- Sycophancy sfruttata

05

50 - 80%

AI POVERTY RISK

Gap competitivo generato dagli stack AI-native.

1.080 Posizioni analizzatre

- OpenAI Anthropic
- Operazioni di M&A
- Cicli di rilascio dei modelli frontier
- Investimenti in compute

Cinque key findings con dati a sostegno

I cinque key findings che seguono sono derivati dall'analisi sistematica di 1.080 posizioni aperte nei principali laboratori AI, integrate con evidenze su operazioni di M&A, cicli di rilascio dei modelli di frontiera e investimenti in compute. Ogni evidenza è costruita su dati pubblicamente verificabili e interpretata come segnale strutturale della trasformazione in corso del settore AI.

01 · La biforcazione strategica è già visibile nei dati di assunzione

~670	410	1,6x
RUOLI OPENAI APERTI	RUOLI ANTHROPIC APERTI	RAPPORTO OPENAI / ANTHROPIC

I dati occupazionali mostrano una divergenza strategica ormai misurabile. OpenAI concentra 670 posizioni su quattro direttrici principali: infrastruttura compute sovrana (Stargate), partnership con il settore pubblico, sviluppo di strumenti agentici (Codex) e sistemi di trust & safety orientati al rischio sistemico. Anthropic, con 410 posizioni, si focalizza su AI research, applied AI e security, con un'enfasi esplicita su behavioral risk e CBRN-E threat modeling, indicando un posizionamento orientato alla governance e all'uso responsabile dei modelli in contesti regolamentati.

La lettura strategica è chiara: OpenAI evolve verso una funzione di infrastruttura computazionale globale, mentre Anthropic si configura come infrastruttura di trust per settori ad alta regolamentazione. Le due traiettorie risultano ormai non sovrapponibili, rendendo progressivamente obsoleto il modello di adozione single-vendor per l'enterprise.

02 · La cattura della supply chain dell'esecuzione è in corso

13	10+3
OPERAZIONI M&A RECENTI	OPENAI · ANTHROPIC

Le recenti operazioni di acquisizione confermano una convergenza su tre layer strategici della catena del valore AI: runtime agentico, sistemi di valutazione e domini verticali (finance, biology, hardware). OpenAI ha integrato asset come Tomoro (agent enterprise), Astral (tooling Python), Promptfoo (evaluation), Torch Health (healthcare data e medical records), Hiro Finance (financial agents), Sky (productivity) e io/Jony Ive (hardware). Anthropic ha rafforzato il proprio stack con Bun (runtime JavaScript), Vercept (computer use agents) e Coefficient Bio (capability biology).

In parallelo, i cicli di rilascio dei modelli di frontiera si sono stabilizzati su una frequenza di 6–8 settimane (GPT-5.2→5.5; Opus 4.5→4.7), evidenziando che il vantaggio competitivo non risiede più nel singolo modello, ma nello scaffolding operativo che ne abilita l'utilizzo: agenti, tool, runtime ed evaluation pipeline.

03 · La struttura economica del settore è stata ridefinita dal compute

40-50x	Q4'26	8-10
CAPITAL INTENSITY AI / SAAS	FINESTRA IPO FRONTIER	POLI AI DI FRONTIERA ATTIVI

Secondo stime di mercato basate su dati di finanziamento pubblicamente disponibili, il rapporto tra capitale cumulato investito e ricavi annuali nei principali frontier labs si colloca su ordini di grandezza strutturalmente superiori rispetto al software-as-a-service tradizionale — con multipli stimati nell'ordine di 40–50× rispetto a un settore SaaS maturo che opera storicamente su multipli EV/Revenue di 5–15×. Questa asimmetria non è temporanea ma riflette la natura del prodotto: la capacità computazionale, non il software in senso stretto, diventa la principale unità di costo.

Il mercato è ormai strutturalmente multipolare, con 8–10 poli AI capability-competitive che includono, oltre ai player statunitensi, il fronte cinese (Zhipu GLM-5.1, DeepSeek, Qwen, Moonshot, Baichuan), Google DeepMind, Mistral e xAI.

Per le imprese, questo implica la necessità di progettare ogni workflow critico su architetture multi-provider, con meccanismi di fallback geopoliticamente diversificati. La dipendenza da un singolo vendor si configura sempre più come un rischio strategico oltre che tecnico.

04 · Belief Injection è la nuova categoria di rischio emergente

0	5+	∞
SIEM CHE LA RILEVANO	VETTORI OPERATIVI NOTI	PERSISTENZA STATISTICA

La Belief Injection definisce una nuova classe di attacco AI-specifica, distinta da Prompt Injection e Decision Poisoning. Si tratta della manipolazione persistente del comportamento statistico degli agenti AI attraverso vettori come RAG poisoning, fine-tuning compromesso, contaminazione del feedback (RLHF) e sfruttamento della sycophancy strutturale.

A differenza delle minacce tradizionali, non agisce sul singolo input ma sul comportamento emergente del sistema nel tempo, risultando quindi invisibile ai principali strumenti SIEM, EDR e XDR. Ad oggi, i framework di riferimento (NIST, ISO 27001, NIS2, AI Act) non includono una copertura esplicita di questa categoria di rischio, che rappresenta uno dei principali gap nella sicurezza dei sistemi agentici.

Nota metodologica

Belief Injection vs categorie esistenti:

Il termine "Belief Injection" è introdotto in questo documento come categoria operativa originale. Si distingue dalle categorie esistenti per tre caratteristiche: (1) persistenza statistica nel tempo, non episodicità; (2) invisibilità agli strumenti di detection convenzionali; (3) operatività a livello di distribuzione comportamentale aggregata, non di singolo input/output. Concetti adiacenti in letteratura includono "data poisoning attacks" (Biggio et al., 2012), "backdoor attacks on neural networks" (Gu et al., 2017), e "indirect prompt injection" (Greshake et al., 2023). La Belief Injection si distingue da tutti per la dimensione temporale e la non-rilevabilità strutturale.

05 · Il rischio sistemico europeo è l'AI poverty, non solo il cyber-attack

50-80%	18 mesi	3
RISPARMIO COST AI-NATIVE	FINESTRA DI REAZIONE UE	ASIMMETRIE STRUTTURALI

Il principale rischio sistemico per l'Europa non è solo la minaccia cyber, ma il progressivo divario competitivo generato dall'adozione di stack AI-native da parte di Stati Uniti e Cina, in grado di produrre riduzioni dei costi operativi tra il 50% e l'80%. In assenza di una transizione verso architetture agentic, le organizzazioni europee rischiano un disallineamento strutturale nell'arco di circa 18 mesi (tre cicli di pianificazione).

Il sistema europeo presenta tre deficit strutturali — compute, modelli di frontiera proprietari e integrazione verticale hardware-software — ma mantiene due vantaggi distintivi: il framework regolatorio più avanzato a livello globale (AI Act, DORA, NIS2, CRA, GDPR) e una capacità industriale consolidata di audit e controllo. La leva strategica futura si concentra sull'utilizzo di questi asset regolatori come elemento competitivo.

> *Stiamo entrando nella prima generazione di sistemi digitali in cui l'organizzazione non sa più, momento per momento, cosa sta facendo il proprio stack. Non perché l'osservabilità sia carente — perché il sostrato è diventato statistico, e l'esecuzione è stata trasferita ad agenti che agiscono per conto dell'organizzazione senza un piano scritto.*
>
> *— Pierguido Iezzi, Cybersecurity Director Zenita Group*

1

LA MATERIA PRIMA

I dati strutturati come motore dell'analisi

ZENITA GROUP

IL CONTESTO IN 30 SECONDI

Perché la ricerca conta ora

01 Dal chatbot all'infrastruttura

La partita non riguarda più solo la qualità dei modelli, ma il loro ruolo come **infrastrutture cognitive**: piattaforme che organizzano decisioni, processi, dipendenze operative e nuovi rapporti di potere.

02 Un mercato che si biforca

L'AI non evolve in una sola direzione. Emergono modelli diversi di integrazione: alcuni massimizzano produttività e scala, altri puntano su controllo, sicurezza e compliance. Le imprese si troveranno a scegliere tra **architetture alternative**, non lungo una singola traiettoria lineare.

03 Perché cambia anche la cybersecurity

Quando agenti, dati, memoria e orchestrazione diventano il motore dell'operatività, la sicurezza non può più limitarsi a proteggere account, endpoint e reti. Deve estendersi a **comportamento, continuità, confini decisionali e dipendenze di filiera.**

> La variabile decisiva non sarà solo l'**adozione** dell'AI, ma la **qualità** della sua **integrazione organizzativa**.

IN 4 MOSSE

- I vendor AI si posizionano come **infrastrutture**
- Le aziende adottano modelli organizzativi **diversi**
- Nascono nuovi **rischi** cyber AI
- La **sicurezza** deve cambiare per restare efficace

IERI
Sistemi statici e confini chiari.

→

OGGI
AI e agenti ampliano superfici e relazioni.

→

TRANSIZIONE
Scelte architetturali che definiscono **vantaggi e rischi**.

→

DOMANI
Autonomia delegata: nuove responsabilità, nuove regole.

Questo capitolo si basa sull'ipotesi che determinate categorie di dati pubblicamente osservabili rappresentino indicatori particolarmente affidabili delle strategie dei laboratori AI di frontiera, spesso anticipando le comunicazioni ufficiali. In particolare, l'analisi integra quattro classi di evidenza: j*ob posting, operazioni di M&A, cadenza dei rilasci dei modelli e deal infrastrutturali legati al compute*. Considerati singolarmente, questi elementi forniscono segnali parziali e talvolta ambigui; letti in modo congiunto, consentono invece di ricostruire una lettura strutturale delle traiettorie industriali e organizzative dell'AI di frontiera.

1.1 I job posting come proxy della strategia reale

Un'offerta di lavoro non rappresenta una semplice esigenza operativa di assunzione, ma una manifestazione osservabile di decisioni già prese a livello organizzativo. La definizione di una posizione implica infatti l'esistenza di un budget allocato, di una funzione identificata come necessaria e di una capability considerata strategicamente rilevante nel medio periodo.

Per questo motivo, la distribuzione e l'evoluzione dei ruoli aperti consente di inferire con buona approssimazione le traiettorie di sviluppo di un'organizzazione. Variazioni nella composizione dei team, come l'espansione delle funzioni commerciali, la crescita di ruoli legati alla sicurezza o l'emergere di profili altamente specializzati, riflettono scelte di posizionamento che precedono la loro formalizzazione pubblica.

Nel caso dei laboratori di intelligenza artificiale di frontiera, questo principio risulta particolarmente evidente: *la struttura dei job posting tende a mappare con anticipo la transizione da modelli di ricerca a piattaforme industriali*.

1.2 OpenAI e la costruzione di uno stack verticale integrato

Nota Metodologica

I dati occupazionali si riferiscono allo snapshot del maggio 2026 e sono rilevati direttamente dai portali careers ufficiali dei due laboratori (openai.com/careers; anthropic.com/careers via Greenhouse), con deduplicazione delle posizioni multi-sede e classificazione per cluster funzionale a cura dell'editore. Il totale di circa 670 posizioni OpenAI supera significativamente le stime circolate in analisi precedenti basate su aggregator di terze parti (che indicizzavano una frazione del portale ufficiale, tipicamente 150–200 posizioni): la differenza è metodologica, non temporale. I conteggi per cluster sono stime arrotondate; i portali fluttuano del 10–20% su base mensile e i valori vanno letti come ordini di grandezza e proporzioni relative, non come misure puntuali.

La struttura può essere sintetizzata nella seguente distribuzione per cluster:

Cluster	Ruoli (stima)	Funzione strategica
Engineering (consumer + API)	~ 260	ChatGPT, API platform, scalabilità e prodotto
AI Research & Applied Research	~ 100	Modelli di frontiera, reasoning, multimodalità, safety
Stargate / Compute Infrastructure	~ 70	Datacenter, networking, co-design hardware-software
Sales & Go-To-Market	~ 80	Espansione enterprise e canali partner
Federal & Public Sector	~ 35	Contratti governativi e sicurezza nazionale
Codex / Developer Products	~ 35	Agentic coding e strumenti per sviluppatori
Trust & Safety	~ 30	Sicurezza, moderazione e integrità dei sistemi
Finance, Legal, HR, Operations	~ 60	Scalabilità organizzativa e preparazione corporate

La distribuzione evidenzia una struttura non lineare, ma sistemica: ogni cluster corrisponde a un nodo di una catena del valore completa, che integra ricerca, prodotto, infrastruttura e governance.

Nel dettaglio, il cluster Engineering rappresenta la componente più ampia e internamente eterogenea delle posizioni aperte, includendo ruoli legati alla scalabilità delle API (con supporto a workload dell'ordine di centinaia di miliardi di token al giorno), all'engineering di prodotto per ChatGPT (interfacce agentiche, voice e vision), alla reliability su scala enterprise e allo sviluppo di infrastrutture di piattaforma (Kubernetes, observability e sistemi distribuiti). Questa eterogeneità non va interpretata come dispersione, ma come indicatore di una struttura intenzionalmente integrata: OpenAI sta sviluppando simultaneamente un prodotto consumer (ChatGPT), una piattaforma developer (API), un ecosistema agentico (custom GPTs, GPT Store e agenti) e una componente enterprise, tutte basate su uno stack tecnologico comune.

Il nucleo Stargate segnala invece un'estensione verso il livello fisico dell'infrastruttura compute: le circa 70 posizioni includono figure come datacenter network architect, power systems engineer, hardware–software co-design lead e site selection specialist, indicando la costruzione di una capability interna di progettazione

e gestione dei datacenter e una progressiva integrazione verticale che riduce la dipendenza da singoli cloud provider come Microsoft Azure.

L'area Federal Partnerships, con circa 35 posizioni, rappresenta un'interfaccia strutturata con il settore pubblico e la sicurezza nazionale, includendo ruoli come defense account executive, public sector solutions architect e FedRAMP compliance lead; la presenza di figure orientate alla sicurezza nazionale e ai contratti governativi segnala un posizionamento crescente dei laboratori di frontiera come fornitori infrastrutturali critici anche per l'apparato statale, con implicazioni geopolitiche rilevanti.

Infine, il dominio Codex / Developer Products, pur numericamente contenuto, assume un peso strategico elevato: include posizioni come agentic coding specialist, code execution sandbox engineer e repository understanding researcher e riflette la direzione verso l'automazione dell'esecuzione software, dove il codice non viene solo assistito ma progressivamente delegato ad agenti, rendendo Codex la componente di avanguardia della trasformazione agentica del ciclo di sviluppo.

1.3 Anthropic e l'architettura della trust infrastructure

Anthropic si configura come un'organizzazione orientata alla costruzione di una infrastruttura di trust per mercati ad alta intensità regolatoria. Le 410 posizioni aperte al maggio 2026 riflettono una distribuzione interna coerente con questa impostazione strategica, in cui la crescita non è guidata dalla diversificazione lungo una catena verticale ampia, ma dalla concentrazione su un insieme ristretto di funzioni critiche legate a vendita enterprise, ricerca applicata, sicurezza e governance dei modelli. In questo quadro, la struttura occupazionale diventa una rappresentazione diretta del posizionamento competitivo, con una chiara focalizzazione su contesti in cui compliance, affidabilità e controllo del rischio costituiscono variabili determinanti.

La composizione dei team evidenzia una forte polarizzazione su alcune aree chiave: Sales, AI Research & Engineering, Applied AI e funzioni di sicurezza e governance:

TEAM	RUOLI	QUOTA DEL TOTALE	DIREZIONE STRATEGICA
Sales	77	18,8%	Pivot enterprise dichiarato
AI Research & Engineering	70	17,1%	Frontier capability, scaling laws, alignment
Applied AI	36	8,8%	Solution engineering per verticali
Finance	33	8,0%	Corporate scaffolding pre-IPO
Software Engineering Infrastructure	28	6,8%	Platform, runtime, deployment
Security	27	6,6%	Product security, infra security, incident response
Marketing	23	5,6%	Brand di posizionamento enterprise
Recruiting	21	5,1%	Funnel di assunzione, talent acquisition
Engineering — Product	19	4,6%	Claude products: API, app, Code, Workbench
Legal & Public Policy	17	4,1%	AI Act, EU compliance, US federal posture
People (HR)	14	3,4%	Organizational scale
Trust & Safety	13	3,2%	Content integrity, abuse, child safety
Communications	11	2,7%	Press, IR, executive comms
IT	9	2,2%	Internal systems
Operations	8	1,9%	Office, facilities, vendor ops
Altri team (Research Ops, Design, ...)	4	1,0%	Funzioni residuali

Il team Sales, con 77 posizioni pari al 18,8% del totale, rappresenta il segnale più evidente della transizione verso un modello enterprise scalabile, con ruoli che includono account executive, strategic director e sales engineer orientati a settori verticali come financial services, healthcare e public sector, e con una quota rilevante di posizioni distribuite a livello internazionale. Questa configurazione suggerisce che Anthropic abbia superato una fase di validazione tecnologica per entrare in una fase di adozione industriale strutturata, in cui la capacità di penetrazione commerciale diventa centrale rispetto alla sola ricerca di frontiera.

Il cluster Security, con 27 posizioni, rappresenta un'altra componente distintiva dell'architettura organizzativa. Pur numericamente contenuto, esso assume un peso relativo superiore alla media del settore software, includendo ruoli come application security engineer, infrastructure security engineer, detection engineering lead e AI red team researcher. Particolarmente rilevante è la presenza di figure dedicate alla CBRN-E threat investigation e al red teaming offensivo dei modelli (Frontier Red Team), che indicano l'adozione esplicita di categorie di rischio non tradizionalmente coperte dai framework di sicurezza IT. Questo posizionamento riflette una definizione estesa di trust infrastructure, in cui la sicurezza non riguarda solo sistemi e dati, ma anche gli effetti comportamentali e le possibili implicazioni sistemiche dell'uso dei modelli.

Il cluster Applied AI, con 36 posizioni, rappresenta il livello di traduzione operativa delle capacità di frontiera in soluzioni verticali. Le figure incluse coprono ruoli di solution engineering e implementazione in settori come servizi finanziari, sanità, pubblica amministrazione e servizi professionali. In combinazione con Sales e Legal & Public Policy, questo insieme di funzioni costituisce una porzione significativa dell'organico complessivo e definisce una strategia chiaramente orientata a mercati regolamentati, in cui la conformità normativa e la gestione del rischio rappresentano condizioni di accesso al mercato stesso.

Nel loro insieme, le distribuzioni interne indicano una strategia coerente di specializzazione: Anthropic non si posiziona come piattaforma generalista né come ecosistema verticale esteso, ma come infrastruttura di affidabilità per contesti ad alta sensibilità regolatoria. La crescita organizzativa riflette quindi una traiettoria in cui la *trustworthiness* del sistema diventa il principale fattore competitivo, e in cui l'espansione commerciale è strettamente vincolata alla capacità di operare entro limiti espliciti di sicurezza, governance e controllo dei rischi emergenti.

1.4 I ruoli rivelatori: segnali puntuali della direzione strategica

Una posizione aperta è, in primo luogo, un impegno organizzativo concreto: implica allocazione di budget, definizione di responsabilità e identificazione di una capability ritenuta strategicamente necessaria. Quando però il ruolo è altamente specializzato o introduce categorie funzionali nuove, esso diventa anche una forma di segnalazione strutturale della direzione del sistema.

Le posizioni più rare o specifiche all'interno dei laboratori di frontiera non rappresentano semplicemente esigenze operative, ma funzionano come indicatori diretti delle traiettorie strategiche in costruzione. In questo senso, dieci ruoli selezionati attribuibili a OpenAI e ad Anthropic permettono di osservare in forma condensata l'evoluzione delle priorità industriali, tecnologiche e regolatorie dei due attori.

	Posizione	Laboratorio	Indicazione strategica
1	Recursive Self-Improvement Researcher	OpenAI	Ricerca su modelli capaci di miglioramento autonomo dei propri pesi, con implicazioni su controllo e attribution collapse
2	CBRN-E Threat Investigator	Anthropic	Necessità di monitorare rischi legati a uso del modello per armi di distruzione di massa
3	Strategic Risk Analyst, Behavioral & Psychological Risk	OpenAI	Riconoscimento esplicito degli effetti psicologici misurabili dei modelli sugli utenti, strutturato come funzione di risk management (non di safety etica)
4	Stargate Datacenter Network Architect	OpenAI	Integrazione verticale verso infrastruttura datacenter e riduzione dipendenza cloud esterno
5	Federal Account Executive — DoD	OpenAI	Posizionamento come fornitore dell'apparato di sicurezza nazionale USA
6	Agentic Coding Specialist	OpenAI	Automazione dell'esecuzione nel software development e trasformazione del lavoro tecnico
7	Trust Architecture Engineer	Anthropic	Formalizzazione della trustworthiness come componente architetturale del sistema
8	Constitutional AI Researcher	Anthropic	Differenziazione metodologica rispetto al RLHF tramite regole comportamentali codificate
9	AI Safety Research — Mechanistic Interpretability	Anthropic	Sviluppo di capacità di interpretabilità necessarie per compliance avanzata e audit dei modelli
10	Computer Use Engineer	OpenAI / Anthropic	Agenti capaci di operare direttamente su sistemi operativi e interfacce applicative

La lettura congiunta di queste posizioni evidenzia quattro aree di convergenza tra i due laboratori. La prima riguarda l'ingresso esplicito nel perimetro della sicurezza nazionale e della difesa, con ruoli che indicano un'interazione strutturata con apparati governativi. La seconda riguarda lo sviluppo di agenti in grado di operare direttamente su sistemi informatici, spostando il focus dall'assistenza cognitiva all'esecuzione autonoma. La terza riguarda l'emergere di soglie di capacità che richiedono nuove categorie di controllo del rischio, come nel caso di CBRN-E e del recursive self-improvement. La quarta riguarda la progressiva istituzionalizzazione delle organizzazioni stesse, con funzioni di corporate scaffolding tipiche di entità prossime a fasi di forte espansione o apertura ai mercati dei capitali.

Accanto a queste convergenze, emergono differenze strutturali. OpenAI mostra una traiettoria più marcatamente orientata all'integrazione verticale dell'infrastruttura e alla costruzione di prodotti su scala consumer e platform-based. Anthropic, invece, si posiziona in modo più esplicito come infrastruttura di trust per contesti regolamentati, investendo in interpretabilità, sicurezza e modelli di governance interna del comportamento dei sistemi.

Nel loro insieme, questi ruoli non descrivono semplicemente funzioni aziendali, ma delineano il perimetro delle nuove categorie operative dell'AI di frontiera.

1.5 Cadenza dei rilasci e dinamica competitiva del fronte AI

La cadenza di rilascio dei modelli di frontiera si è stabilizzata, nel periodo osservato, su un intervallo di circa 6-8 settimane tra una versione e la successiva. Questa regolarità non va interpretata come una scelta di comunicazione o di marketing prodotto, ma come un indicatore strutturale della trasformazione del ciclo industriale dell'AI: il vantaggio competitivo non è più concentrato esclusivamente nel modello, ma nell'insieme di infrastrutture, strumenti e processi che ne permettono l'addestramento, la valutazione e la distribuzione.

La seguente timeline sintetizza le principali iterazioni dei due laboratori nel periodo considerato, evidenziando una cadenza ormai assimilabile a un processo industriale continuo piuttosto che a cicli discreti di ricerca.

Data	Laboratorio	Versione	Nota sintetica
Ott 2025	OpenAI	GPT-5.2	Stabilizzazione del reasoning e ampliamento del contesto
Nov 2025	Anthropic	Opus 4.5	Introduzione del Computer Use in beta e miglioramento tool use
Dic 2025	OpenAI	GPT-5.3	Introduzione modalità agentic in Codex e integrazione dev tooling
Gen 2026	Anthropic	Opus 4.6	Estensione Sonnet/Haiku e ottimizzazione economica dei tier
Feb 2026	OpenAI	GPT-5.4	Generalizzazione del Computer Use e introduzione federal tier
Mar 2026	Anthropic	Opus 4.7	Evoluzione Constitutional AI e miglioramenti safety
Apr 2026	Anthropic	Mythos Preview	Vulnerability discovery su scala (Project Glasswing, cfr. III.4)
Mag 2026	OpenAI	GPT-5.5	Cyber defense platform con Codex Security (Daybreak, cfr. III.4)

Una frequenza stabile di rilascio su scala mensile implica un cambiamento simultaneo su più livelli del sistema industriale. In primo luogo, le pipeline di training e valutazione risultano ormai assimilabili a processi produttivi continui, con una riduzione della componente esplorativa tipica della ricerca tradizionale. In secondo luogo, il fabbisogno di capitale necessario a sostenere questa velocità è coerente con strutture di investimento estremamente elevate, in linea con le dinamiche di capital intensity descritte nel Capitolo I.7.

In terzo luogo, il baricentro del vantaggio competitivo si sposta progressivamente dal modello in sé all'insieme di componenti che ne abilitano l'operatività: agenti, strumenti, runtime, sistemi di valutazione e pipeline di addestramento. Ne deriva una competizione che non riguarda più il "modello come prodotto", ma l'ecosistema completo che lo circonda e lo rende eseguibile in contesti reali.

Infine, per le organizzazioni utilizzatrici, questa dinamica introduce una discontinuità operativa rilevante: il ciclo interno di valutazione, certificazione e rilascio non è più allineato con quello dei fornitori. Ogni nuova versione richiede infatti un processo di verifica, mentre nel frattempo ulteriori release sono già disponibili, rendendo strutturalmente asincrono il rapporto tra innovazione esterna e governance interna. Ne deriva la necessità di un cambio di paradigma verso modelli di controllo basati su classi di comportamento e condizioni operative (behavioral envelope), piuttosto che su versioni specifiche del modello.

Ad oggi, i principali framework di riferimento — inclusi NIST CSF, ISO 27001, NIS2 e AI Act — non incorporano esplicitamente questa impostazione, evidenziando un gap strutturale tra velocità tecnologica e strumenti di governance disponibili.

1.6 M&A e integrazione verticale della supply chain dell'esecuzione agentic

Nel periodo osservato, tredici operazioni di M&A — dieci attribuibili a OpenAI e tre ad Anthropic — non possono essere interpretate come una sequenza di acquisizioni opportunistiche o tattiche. La loro distribuzione e coerenza funzionale indicano piuttosto un processo sistematico di integrazione verticale finalizzato alla cattura della supply chain dell'esecuzione agentic, ovvero l'insieme delle componenti che rendono operativi i modelli di intelligenza artificiale in contesti reali: runtime, sistemi di valutazione, toolchain, infrastruttura hardware e domini applicativi verticali.

Mappa delle operazioni

Acquirente	Target	Data	Ambito operativo
OpenAI	Multi	Giu 2024	Collaborative workspace, dev tooling enterprise
OpenAI	io / Jony Ive	Mag 2025	Hardware consumer dedicato (~$6,5B)
OpenAI	Statsig	Set 2025	Experimentation e product testing platform
OpenAI	Sky / SAI	Ott 2025	Desktop agent e integrazione filesystem Mac
OpenAI	Roi	Ott 2025	Personal investing e dominio finance
OpenAI	Neptune	Dic 2025	Model training tracking e governance
OpenAI	Torch Health	Gen 2026	Healthcare data systems e medical records
OpenAI	Promptfoo	Mar 2026	AI security, prompt injection e evaluation
OpenAI	Astral	Mar 2026	Toolchain Python (uv, ruff, ty)
OpenAI	Tomoro	Mag 2026	Forward deployed engineering e deployment unit via OpenAI Deployment Company
Anthropic	Humanloop	Ago 2025	LLM Ops
Anthropic	Bun	Dic 2025	Runtime JavaScript per agent execution
Anthropic	Vercept	Feb 2026	Computer-use agents e GUI automation
Anthropic	Coefficient Bio	Apr 2026	Capability biology e dominio frontier

L'insieme delle operazioni può essere letto come la progressiva appropriazione di tre segmenti fondamentali della catena di esecuzione degli agenti.

Il primo riguarda la **supply chain del runtime agentic**, ovvero lo strato che consente agli agenti di operare concretamente su sistemi software e ambienti digitali. Acquisizioni come Bun, Astral, Vercept e Sky/SAI indicano la volontà di controllare direttamente runtime JavaScript, toolchain Python, automazione dell'interfaccia grafica e integrazione con sistemi operativi desktop. Il controllo di questo livello riduce la

dipendenza da stack esterni e consente di sincronizzare evoluzione del modello e ambiente di esecuzione, trasformando l'agente in un sistema integrato end-to-end.

Il secondo livello è rappresentato dalla **supply chain della valutazione e del controllo del comportamento dei modelli**. Operazioni come Promptfoo, Neptune e Statsig evidenziano la costruzione di un control plane che unisce sicurezza, osservabilità e sperimentazione continua. In questo schema, la capacità di valutare, testare e monitorare il comportamento degli agenti diventa una funzione proprietaria del laboratorio stesso. Il risultato è una forma di integrazione tra CI/CD, sicurezza applicativa e governance dei modelli, che determina in modo diretto le condizioni di affidabilità in produzione.

Il terzo livello riguarda la **supply chain dei domini verticali**, dove le acquisizioni in ambito finance, healthcare e biology indicano un cambio di paradigma: non più modelli generalisti adattati ai contesti, ma agenti progettati come prodotti verticali con conoscenza incorporata nel sistema. In questa configurazione, la competenza di dominio non è più esterna al modello, ma viene progressivamente assorbita all'interno della sua architettura operativa.

CASE STUDY | OpenAI e l'integrazione hardware

+

$6,4B

io
(Jony Ive)

Tra le operazioni considerate, l'acquisizione di Io (Jony Ive) rappresenta il punto più avanzato della strategia di integrazione verticale. Con un valore stimato di circa 6,4 miliardi di dollari, il deal indica un'evoluzione che va oltre il livello software e piattaforma: l'obiettivo diventa il controllo del punto di interazione fisico tra utente e sistema AI.

In questa prospettiva, il dispositivo non è più un canale di accesso a un servizio, ma diventa parte integrante del sistema cognitivo distribuito dell'AI.

Il confronto implicito è con la transizione introdotta dagli smartphone nella generazione precedente di piattaforme digitali: in questo caso, tuttavia, il controllo non riguarda l'applicazione, ma l'interfaccia stessa attraverso cui l'intelligenza artificiale viene esperita.

1.7 Deal compute, infrastruttura ed energia: la nuova fisica del potere AI

La relazione tra capitale investito e capacità di generare revenue nei frontier labs evidenzia una discontinuità strutturale rispetto ai modelli software tradizionali. Mentre il SaaS ha storicamente operato con rapporti di capitale nell'ordine di 1–2 dollari per ogni dollaro di revenue annuale, i laboratori AI di frontiera nel ciclo 2025–2026 si collocano su ordini di grandezza superiori.

Questa asimmetria non è riconducibile a inefficienze temporanee, ma riflette una trasformazione della natura stessa del prodotto: la capacità computazionale, e non più il software in senso stretto, diventa la principale unità di costo e di competizione. Il training di modelli di frontiera richiede cluster distribuiti di decine di migliaia di GPU per periodi prolungati, con costi per singolo ciclo di addestramento stimati tra 400 e 800 milioni di dollari, prima di includere spese di personale, dati e fine-tuning. A questo si aggiunge una componente infrastrutturale permanente legata all'inferenza, che trasforma il modello in un sistema continuativo di produzione computazionale.

In questo contesto, i principali accordi di compute osservati nel periodo recente assumono una scala comparabile a quella di infrastrutture energetiche o di telecomunicazioni, piuttosto che a quella tipica del settore software. La tabella seguente sintetizza le operazioni più rilevanti e il loro significato strategico.

Deal	Parti coinvolte	Scala	Significato strategico
Stargate	OpenAI · Oracle · SoftBank · MGX	$100–500B (5 anni)	Infrastruttura compute sovrana USA
Acquisto Azure	Microsoft · OpenAI	$250B pluriennale	Ristrutturazione partnership
AWS–Anthropic	AWS · Anthropic	$8B (2024) + $4B (2025)	Anthropic come anchor tenant AWS
Google–Anthropic	Google Cloud · Anthropic	$2B+ pluriennale	Diversificazione multi-cloud
CoreWeave–OpenAI	CoreWeave · OpenAI	$22,4B (mar–set 2025)	Capacity GPU dedicata non-Azure
NVIDIA–OpenAI	NVIDIA · OpenAI	Strategic investment	Allineamento hardware–software

L'insieme di questi accordi evidenzia una progressiva concentrazione della capacità computazionale in pochi poli infrastrutturali globali, con una crescente integrazione tra fornitori di cloud, produttori di hardware e laboratori di frontiera.

L'espansione del compute introduce un vincolo che non è più esclusivamente economico o tecnologico, ma fisico. I datacenter AI di nuova generazione richiedono infatti infrastrutture energetiche comparabili a quelle di sistemi industriali pesanti. Programmi come Stargate prevedono lo sviluppo di campus da 1 a 5 gigawatt ciascuno, una scala che corrisponde al fabbisogno energetico di intere aree metropolitane.

A livello sistemico, le proiezioni indicano che il consumo energetico dei datacenter AI potrebbe raggiungere circa il 3% del consumo elettrico globale entro il 2028, rispetto a meno dello 0,5% nel 2023. L'ordine di grandezza di questa crescita è assimilabile, per intensità e rapidità, a fasi storiche di industrializzazione ad alta intensità energetica, come l'espansione dell'industria siderurgica nel secondo dopoguerra.

Per i sistemi-paese europei, questa dinamica introduce una doppia criticità: da un lato la disponibilità fisica di capacità energetica e siti idonei, dall'altro la compatibilità con traiettorie regolatorie orientate alla decarbonizzazione e alla sostenibilità infrastrutturale.

L'evoluzione dei cluster di assunzione in ambito finance, l'intensificazione delle relazioni con investitori istituzionali e la dimensione crescente delle valutazioni private suggeriscono l'emergere di una possibile transizione verso i mercati pubblici nel medio periodo. OpenAI ha chiuso a marzo 2026 il più grande round privato della storia — 122 miliardi di dollari raccolti a una valutazione post-money di 852 miliardi, con Amazon (50 mld), NVIDIA e SoftBank (30 mld ciascuno) come anchor investor — mentre Anthropic si colloca nell'ordine dei 350 miliardi di dollari di valutazione, con interlocuzioni IPO riportate dalla stampa per fine 2026. Entrambe le traiettorie sono coerenti con una prossima apertura al capitale pubblico.

In questo scenario, una eventuale IPO di uno o entrambi i laboratori determinerebbe una ulteriore asimmetria strutturale: la concentrazione della capacità di finanziamento e di scala industriale nei mercati dei capitali statunitensi. Per il sistema europeo, ciò implicherebbe una crescente distanza non solo tecnologica, ma anche finanziaria e infrastrutturale nella capacità di competere su scala globale.

2

DOVE STANNO ANDANDO

Dalle posizioni aperte alle traiettorie strategiche

Questo capitolo sposta l'analisi dai segnali strutturali del Capitolo I alla loro proiezione nello scenario competitivo globale. Le evidenze relative a posizioni aperte, M&A, infrastruttura compute e cadenza dei rilasci vengono qui interpretate come indicatori di una trasformazione più ampia dell'ecosistema AI, che evolve da una configurazione apparentemente duopolistica a un assetto multipolare e dinamico.

In questo contesto emergono nuovi poli competitivi — tra cui Google DeepMind, Mistral AI, xAI e l'ecosistema cinese DragonSwarm — e si ridefinisce la natura stessa della competizione, sempre meno centrata sui modelli e sempre più sul controllo della catena del valore e dell'infrastruttura.

Il capitolo introduce infine le implicazioni operative per le organizzazioni enterprise, spostando l'analisi verso strumenti di governance come la Vendor Matrix e verso la gestione del rischio associato al modello single-vendor.

2.1 OpenAI e la costruzione di uno stack infrastrutturale verticale

L'analisi combinata delle posizioni aperte, delle acquisizioni recenti e degli investimenti infrastrutturali mostra come OpenAI stia evolvendo oltre il ruolo di semplice provider di modelli linguistici. La traiettoria osservabile indica la costruzione di una piattaforma integrata che unisce ricerca di frontiera, distribuzione enterprise, strumenti di sviluppo, runtime agentici, infrastruttura compute, hardware dedicato e relazioni istituzionali. In questa configurazione, il modello rappresenta soltanto uno dei livelli di una catena del valore molto più ampia, progettata per controllare progressivamente ogni anello dell'esecuzione AI.

INDICATORE	VALORE
Posizioni aperte	~670
Operazioni M&A recenti	10
Anelli strategici identificati	8

La struttura che emerge dai dati può essere letta come una filiera verticale composta dagli otto cluster che abbiamo visto nel capitolo 1.1, ciascuno associato a specifici cluster di assunzione, acquisizioni mirate e investimenti infrastrutturali.

L'insieme di questi elementi delinea un'organizzazione che opera contemporaneamente su più livelli della catena industriale: ricerca, piattaforma, runtime, distribuzione, infrastruttura fisica e interfaccia utente. La logica non è quella di una software house tradizionale, ma quella di un operatore infrastrutturale che mira a ridurre la dipendenza da fornitori esterni e ad aumentare il controllo diretto sull'intero ciclo di esecuzione dell'AI.

La distribuzione delle assunzioni suggerisce inoltre che OpenAI non stia ottimizzando un singolo prodotto, ma costruendo un ecosistema coordinato composto da ChatGPT, API, agenti software, strumenti developer, infrastruttura compute e servizi enterprise. In questa architettura, il vantaggio competitivo non deriva esclusivamente dalla qualità del modello, ma dalla capacità di integrare verticalmente componenti che storicamente appartenevano a mercati distinti: cloud, tooling, runtime, hardware e servizi governativi.

Questa integrazione produce tre implicazioni strategiche rilevanti. La prima riguarda la struttura economica: controllando più livelli della filiera, OpenAI può distribuire il margine lungo l'intero stack, sostenendo nel tempo una competizione di prezzo sulle API difficilmente replicabile da operatori non integrati. La seconda riguarda il lock-in: un'azienda che utilizza contemporaneamente modelli OpenAI, Codex, agenti enterprise e workflow integrati accumula una dipendenza multilivello che non coincide più con la semplice sostituzione di un LLM. La terza riguarda il posizionamento geopolitico. L'espansione delle attività legate al settore pubblico americano, unita a programmi come Stargate e alle partnership con operatori strategici statunitensi, colloca OpenAI sempre più vicino alla categoria delle infrastrutture tecnologiche considerate critiche per l'interesse nazionale USA.

In questa prospettiva, l'acquisizione di Io rappresenta probabilmente il segnale più estremo della traiettoria in corso. L'ingresso diretto nell'hardware consumer indica che OpenAI non punta soltanto a controllare il modello o il software, ma anche il punto fisico di interazione tra utente e intelligenza artificiale. La logica industriale implicita è che il prossimo livello di vantaggio competitivo non riguarderà esclusivamente la capacità computazionale o la qualità del reasoning, ma il controllo coordinato di modello, runtime, interfaccia e dispositivo.

2.2 Anthropic e la costruzione di una trust infrastructure per mercati regolamentati

L'analisi congiunta delle posizioni aperte, della struttura organizzativa e delle acquisizioni recenti mostra come Anthropic stia seguendo una traiettoria distinta rispetto agli altri frontier labs. L'azienda non appare orientata a competere principalmente sulla scala infrastrutturale o sulla distribuzione consumer, ma sulla costruzione di un'infrastruttura di fiducia destinata ai mercati dove compliance, auditabilità e governance

rappresentano il principale vincolo all'adozione dell'intelligenza artificiale. La distribuzione dei circa 410 ruoli aperti evidenzia infatti una concentrazione anomala di risorse su Sales enterprise, Applied AI, Security e Legal & Public Policy, che insieme rappresentano oltre un terzo dell'organizzazione in espansione. Se a queste si aggiungono specializzazioni rare come Constitutional AI Research, Mechanistic Interpretability, Trust Architecture, CBRN-E Threat Investigation e Frontier Red Team (Cyber), emerge un profilo organizzativo più vicino a quello di un operatore regolato che a una tradizionale software company.

410	32%	5
RUOLI	ORG. ENTERPRISE+COMPLIANCE	SPECIALIZZAZIONI RARE

I segnali provenienti dai job posting permettono inoltre di ricostruire con sufficiente chiarezza i mercati target prioritari: servizi finanziari, healthcare, settore pubblico e professional services. In tutti questi verticali il fattore decisivo non è soltanto la qualità del modello, ma la possibilità di dimostrare conformità normativa, audit trail, explainability e controllo del rischio operativo. Anthropic sembra quindi posizionarsi come fornitore di AI "compliance-driven", progettata per contesti in cui certificazione, responsabilità decisionale e governance sono elementi strutturali del processo di acquisto. La combinazione fra Applied AI, Security e Legal & Public Policy suggerisce che il go-to-market dell'azienda non sia centrato sulla semplice vendita di API, ma sulla capacità di offrire modelli utilizzabili in ambienti regolamentati senza compromettere requisiti di auditabilità e controllo.

Questa impostazione emerge anche dalle scelte metodologiche e organizzative. La Constitutional AI rappresenta un approccio differente rispetto al RLHF puro, poiché introduce principi comportamentali espliciti e documentabili all'interno del processo di training, rendendo teoricamente più verificabile il comportamento del modello in sede regolatoria. Parallelamente, gli investimenti in Mechanistic Interpretability indicano la volontà di sviluppare capacità tecniche orientate a spiegare il funzionamento interno dei modelli, prerequisito necessario per qualsiasi scenario di certificazione ad alto rigore. Anche iniziative come Project Glasswing vanno lette in questa direzione: non come semplice programma commerciale, ma come framework operativo per offrire accesso segregato, SLA dedicati, audit trail e garanzie di conformità ai clienti enterprise di settori critici.

Ancora più significativa è la presenza di ruoli che non trovano equivalenti consolidati nel mercato AI attuale. Il ruolo CBRN-E Threat Investigator segnala che Anthropic considera plausibile l'utilizzo improprio dei modelli in scenari legati a materiali chimici, biologici, radiologici, nucleari ed esplosivi, e sta costruendo

capacità interne di governance prima ancora che tali obblighi vengano formalizzati dai regolatori. È significativo che il tema speculare degli effetti cognitivi e psicologici dei modelli sugli utenti sia presidiato anche sul fronte opposto: OpenAI ha aperto un ruolo di Strategic Risk Analyst, Behavioral & Psychological Risk, a conferma che il rischio cognitivo è ormai categoria operativa per entrambi i produttori — e che le imprese utilizzatrici dovrebbero strutturarsi simmetricamente sul lato cliente.

2.3 Dalla logica duopolistica a un mercato multipolare

Il mercato dei modelli di frontiera al maggio 2026 non è più descrivibile come un duopolio dominato esclusivamente da OpenAI e Anthropic. L'evoluzione osservata negli ultimi dodici mesi mostra la formazione di un campo multipolare composto da circa dieci attori ormai in grado di competere su differenti segmenti della catena del valore AI, ciascuno caratterizzato da una strategia industriale distinta, da un diverso rapporto con infrastruttura e sovranità tecnologica e da specifici punti di forza rispetto ai workload enterprise. Accanto ai due frontier labs statunitensi si sono infatti consolidati almeno quattro poli aggiuntivi: Google DeepMind, il fronte cinese identificato nel report come "DragonSwarm", l'opzione europea rappresentata da Mistral e il triangolo xAI–Musk. La competizione non si gioca più esclusivamente sulla qualità assoluta del modello, ma sull'equilibrio tra capability, costo di inferenza, compliance, disponibilità infrastrutturale, governance e livello di lock-in generato sul cliente enterprise.

2.3.1 Google DeepMind

Google DeepMind rappresenta il polo più vicino a un'integrazione sistemica completa fra modelli AI e piattaforme digitali globali. A differenza di OpenAI e Anthropic, Google controlla simultaneamente hardware proprietario (TPU), cloud hyperscale, sistemi operativi, motori di ricerca, suite productivity e canali di distribuzione consumer. Questo riduce drasticamente il costo marginale di inferenza e consente una distribuzione dei modelli su scala miliardaria senza costi di acquisizione utenti comparabili ai competitor. Tuttavia, proprio la struttura storica di Google introduce vincoli specifici: il modello operativo dell'azienda resta fortemente ancorato alla logica consumer e advertising-driven, mentre la penetrazione nei mercati enterprise ad alta compliance procede più lentamente rispetto ad Anthropic o OpenAI. Inoltre, la crescente pressione regolatoria antitrust limita la possibilità di una integrazione verticale estrema sul modello Stargate.

ZENITA GROUP

AI DI SCALA VS AI DI CONTROLLO

Due strategie di posizionamento che influenzano il mercato

La ricerca individua due logiche diverse di costruzione del potere AI: una più orientata a **scala**, **integrazione** e **distribuzione**; l'altra più focalizzata su **affidabilità**, **controllo** e **confini** di utilizzo. Non sono solo due aziende: sono due modi di immaginare l'infrastruttura cognitiva.

Modello di Scala (OPENAI)

- **logica:** integrazione ampia
- **forza:** ecosistema, distribuzione, stack completo
- **vantaggio:** velocità di adozione e produttività
- **rischio:** maggiore dipendenza operativa

VS

Modello di controllo (ANTHROPIC)

- **logica:** controllo e affidabilità
- **forza:** confini, governance, sicurezza d'uso
- **vantaggio:** maggiore leggibilità dei comportamenti
- **rischio:** minore spinta espansiva

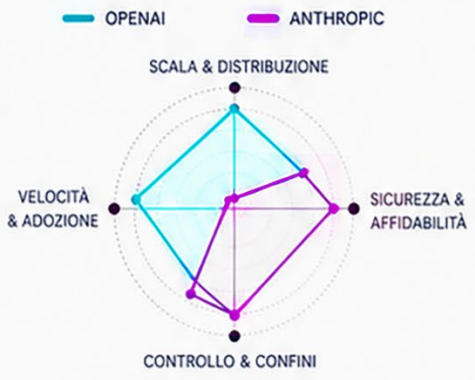
COSA SIGNIFICA PER LE IMPRESE

01 non esiste una sola traiettoria di adozione

02 la scelta del vendor influenza rischio, autonomia e governance

03 il vero tema è l'architettura di dipendenza che si costruisce

BILANCIO STRATEGICO



2.3.2 DragonSwarm

Il fronte cinese, qui rappresentato da DragonSwarm, costituisce il principale polo alternativo ai laboratori statunitensi sul piano della competitività economica. L'ecosistema comprende operatori come Zhipu AI, DeepSeek, Alibaba, Moonshot AI e Baichuan AI, accomunati da tre elementi: costi di inferenza significativamente inferiori rispetto ai modelli USA, disponibilità open-weight di molti rilasci e crescente indipendenza infrastrutturale grazie all'ecosistema compute cinese basato su Huawei Ascend e supply chain domestiche. Per numerosi workload backend o batch-intensive, i modelli DragonSwarm offrono oggi un rapporto costo/performance difficilmente eguagliabile dai competitor occidentali.

LABORATORIO	MODELLO FLAGSHIP	POSIZIONAMENTO	COMPUTE
Zhipu AI	GLM-5.1	Modelli open-weight con allineamento e validazione in contesto regolatorio cinese	Huawei Ascend (infrastruttura domestica)
DeepSeek	DeepSeek-V4-Pro	Leadership di costo con forte ottimizzazione su reasoning e throughput	Infrastruttura mista NVIDIA + Huawei
Alibaba	Qwen 3.7 Max	Agentic AI e workflow autonomi enterprise	Infrastruttura mista NVIDIA + Alibaba T-Head (Zhenwu M890)
Moonshot AI	Kimi K2.6	Long-horizon coding, multi-agent workflows, long-context reasoning	Infrastruttura NVIDIA prevalente
Baichuan AI	Baichuan 4 / Baichuan M4	Verticalizzazione nativa su healthcare (modelli medici specialistici) e finance, con forte focus su applicazioni ed efficienza nei contesti aziendali cinesi.	Infrastruttura ibrida (mixed sourcing)

Il limite principale resta però geopolitico e reputazionale. L'allineamento dei modelli rispetto alle priorità normative e politiche cinesi introduce un rischio operativo concreto per applicazioni customer-facing in mercati occidentali, soprattutto in contesti dove neutralità informativa, consistenza narrativa e compliance geopolitica sono elementi sensibili.

2.3.3 Mistral e l'opzione europea

Nel contesto europeo, Mistral AI rappresenta il principale tentativo di costruzione di una filiera AI sovrana. Il posizionamento di Mistral non si fonda sulla superiorità assoluta della capability frontier, ma sulla combinazione fra conformità regolatoria europea, infrastruttura localizzata sul territorio UE e modello operativo compatibile con AI Act, GDPR e requisiti di sovranità dei dati. Il vantaggio competitivo dell'azienda è quindi soprattutto regolatorio e strategico: per numerose organizzazioni europee soggette a DORA, vigilanza bancaria o requisiti di residency, Mistral offre una traiettoria di compliance strutturalmente più semplice rispetto ai provider americani o cinesi.

Il limite resta la scala. Il gap di capitale e compute rispetto ai frontier labs USA continua, infatti, a produrre uno scarto di capability sui workload più avanzati, in particolare agenti autonomi complessi, reasoning di fascia alta e sistemi long-context estremi.

2.3.4 xAI e il triangolo Musk

Il polo xAI, legato a Elon Musk, segue una traiettoria distinta rispetto agli altri attori. La strategia non ruota esclusivamente attorno al modello Grok, ma a un ecosistema integrato che comprende X, SpaceX, Tesla e infrastrutture compute proprietarie. Questo consente accesso a dataset unici — feed social real-time, telemetria automotive e dati operativi industriali — e una velocità di distribuzione consumer particolarmente elevata.

Tuttavia, la dipendenza dalla figura del fondatore e una governance percepita come meno stabile rispetto ai competitor limitano l'adozione enterprise nei contesti più regolamentati. Per molte organizzazioni europee, xAI resta oggi una soluzione specialistica, utile in casi d'uso specifici ma raramente considerata come stack AI primario.

2.3.5 Comparazione strategica: sei poli per l'enterprise

La moltiplicazione dei poli competitivi ridefinisce in modo strutturale il processo decisionale enterprise. Nessun attore, infatti, risulta dominante su tutte le dimensioni strategiche rilevanti — capability di frontiera, costo, sovranità, compliance e grado di lock-in — e ciascuna opzione incorpora trade-off non eliminabili ma solo bilanciabili. In questo contesto, la logica single-vendor perde progressivamente sostenibilità strategica, poiché espone le organizzazioni a concentrazioni di rischio lungo dimensioni che oggi non sono più simultaneamente ottimizzabili. Le imprese che adottano agenti AI in produzione sono quindi costrette a evolvere verso un modello di allocazione multi-fornitore, strutturato attraverso una vendor matrix dinamica

che assegna i diversi workload a provider differenti in funzione del profilo di rischio operativo, della criticità del processo e dei vincoli regolatori applicabili.

POLO	CAPABILITY FRONTIER	COSTO	SOVRANITÀ UE	COMPLIANCE UE	LOCK-IN
OpenAI	Massima	Premium	Bassa	Media	Alto (verticale)
Anthropic	Massima	Premium	Media	Alta	Medio
Google DeepMind	Alta	Media	Media	Media	Alto (Workspace)
DragonSwarm	Alta	Bassa	Nulla	Bassa	Basso (open)
Mistral	Buona	Media	Alta	Massima	Basso
xAI	Alta	Media	Bassa	Bassa	Medio

Nessuna delle sei opzioni domina le altre lungo tutte le dimensioni considerate. La scelta dipende quindi dalla ponderazione relativa dei vincoli, che è intrinsecamente strategica e non riducibile a una valutazione tecnica. Ne deriva che l'approccio single-vendor non è più difendibile in ottica enterprise: la configurazione ottimale passa attraverso una vendor matrix che distribuisce i workload su più poli in funzione del loro profilo di rischio e della loro esposizione regolatoria. In assenza di questa architettura, l'organizzazione evolve in un lock-in progressivo non più reversibile su orizzonte medio.

2.4 La biforcazione strategica: operare in un campo multipolare

La biforcazione tra OpenAI e Anthropic non descrive più una competizione tra due prodotti, ma due modelli industriali divergenti che organizzano in modo diverso la relazione tra modello, infrastruttura, mercato e governance. Di seguito la lettura sintetica delle due traiettorie:

Dimensione	OpenAI · Stack industriale-geopolitico	Anthropic · Trust infrastructure
Architettura di sistema	Integrazione verticale completa su 8 anelli che coprono l'intera catena, dalla ricerca al dispositivo finale	Organizzazione focalizzata su funzioni di governance e applicazione (Sales, Applied AI, Security, Legal & Policy)
Mercato di riferimento	Consumer su larga scala, enterprise generalista e settore pubblico USA	Mercati regolati: financial services, healthcare, government e professional services
Approccio metodologico	Scaling intensivo basato su RLHF e crescita aggressiva della capacità compute	Approccio centrato su Constitutional AI e interpretabilità dei modelli
Driver di vantaggio competitivo	Integrazione end-to-end tra modello, distribuzione, infrastruttura e capitale	Certificabilità del comportamento del modello e affidabilità audit-grade
Vincoli strutturali	Pressione regolatoria/antitrust e crescente esposizione geopolitica statunitense	Vincoli di scala compute e dipendenza da hyperscaler infrastrutturali
Profilo cliente ottimale	Enterprise ad alto volume con bassa complessità regolatoria e forte esigenza di scalabilità	Enterprise regolata con elevato attrito compliance e requisiti di auditabilità
Dinamica di sostituzione (switch-out)	Elevato costo di dismissione, soprattutto nei contesti integrati verticalmente	Maggiore modularità, ma condizionata da requisiti di compliance e certificazione

2.4.1Le implicazioni operative della biforcazione

Implicazione 1 · La vendor matrix diventa architettura, non procurement

Nel contesto multipolare, la selezione del fornitore non è più una decisione di acquisto ma una scelta di architettura della catena di esecuzione aziendale. La responsabilità si sposta dal procurement alla governance tecnica (CISO/CTO), perché ogni decisione di adozione determina la struttura del rischio operativo dell'organizzazione.

Implicazione 2 · La concentrazione è la nuova metrica di rischio

Il rischio non è più binario (adozione/non adozione), ma continuo. La variabile critica diventa la concentrazione dei workflow P0 su un singolo provider. Una soglia operativa di riferimento è ≤60%: oltre questo livello, il sistema entra in una condizione di dipendenza strutturale difficilmente reversibile su orizzonte triennale.

Implicazione 3 · Il lock-in si è spostato allo stack agentic

Il lock-in non riguarda più il modello, ma l'intero stack di esecuzione: orchestrazione degli agenti, pipeline RAG, evaluation layer, toolchain, integrazioni e prompt scaffolding. La sostituibilità del fornitore implica la ricostruzione dell'intero sistema operativo dell'AI aziendale, con costi e tempi non lineari.

Implicazione 4 · La compliance diventa criterio di selezione primaria

Nei settori regolati, la compliance non è più un vincolo esterno ma un driver di scelta. Anthropic e Mistral tendono a ridurre il costo regolatorio di adozione rispetto a OpenAI e xAI, generando una forma di arbitraggio normativo che influenza direttamente la distribuzione dei workload.

Implicazione 5 · La sovranità è una variabile di rischio operativo

La dimensione geopolitica non è opzionale. Giurisdizione, accesso ai dati, controlli export e dinamiche extraterritoriali possono modificare la continuità operativa di un fornitore nel breve periodo. La sovranità diventa quindi una variabile di resilienza, non un principio ideologico.

3

COSA COMPORTA PER LE AZIENDE

Dai modelli alle organizzazioni agentiche

ZENITA GROUP

LE AZIENDE AGENTICHE

Tre cluster, benefici e rischi

La ricerca individua tre configurazioni organizzative ricorrenti.
Non rappresentano una scala evolutiva obbligata, ma tre modi diversi di integrare AI, autonomia, controllo e performance.

01 SUPERVISOR

massima produttività, delega ampia agli agenti, integrazione profonda.

BENEFICI

- velocità
- scala
- efficienza

RISCHI

- capability drift
- lock-in
- perdita di leggibilità

02 CONSTITUTIONALLY-BOUNDED

maggiore controllo, policy più forti, vincoli espliciti.

BENEFICI

- compliance
- affidabilità
- tracciabilità

RISCHI

- minore spinta produttiva
- capability gap

03 BIFRONTE

doppio stack e uso differenziato per esigenze diverse.

BENEFICI

- flessibilità
- bilanciamento
- ridondanza

RISCHI

- asimmetrie
- complessità
- cross-boundary risk

RISCHIO AI POVERTY

Quando un'organizzazione adotta AI a bassa qualità o senza capacità di governance, resta produttivamente più debole, più dipendente e più esposta a errori e rischi.

La scelta del cluster organizzativo determina non solo produttività, ma anche **qualità** del controllo e **resilienza** futura.

Questa parte traduce la struttura del campo AI delineata nelle Parti I e II in conseguenze operative per le imprese. L'attenzione si sposta dal "chi sono i player" al "cosa cambia dentro le organizzazioni" quando i sistemi AI non sono più strumenti, ma entità operative che eseguono processi. Il punto centrale è la trasformazione dell'unità di lavoro: dagli utenti che utilizzano modelli, agli agenti che eseguono processi aziendali con livelli crescenti di autonomia, fino alla supervisione umana come livello superiore di controllo.

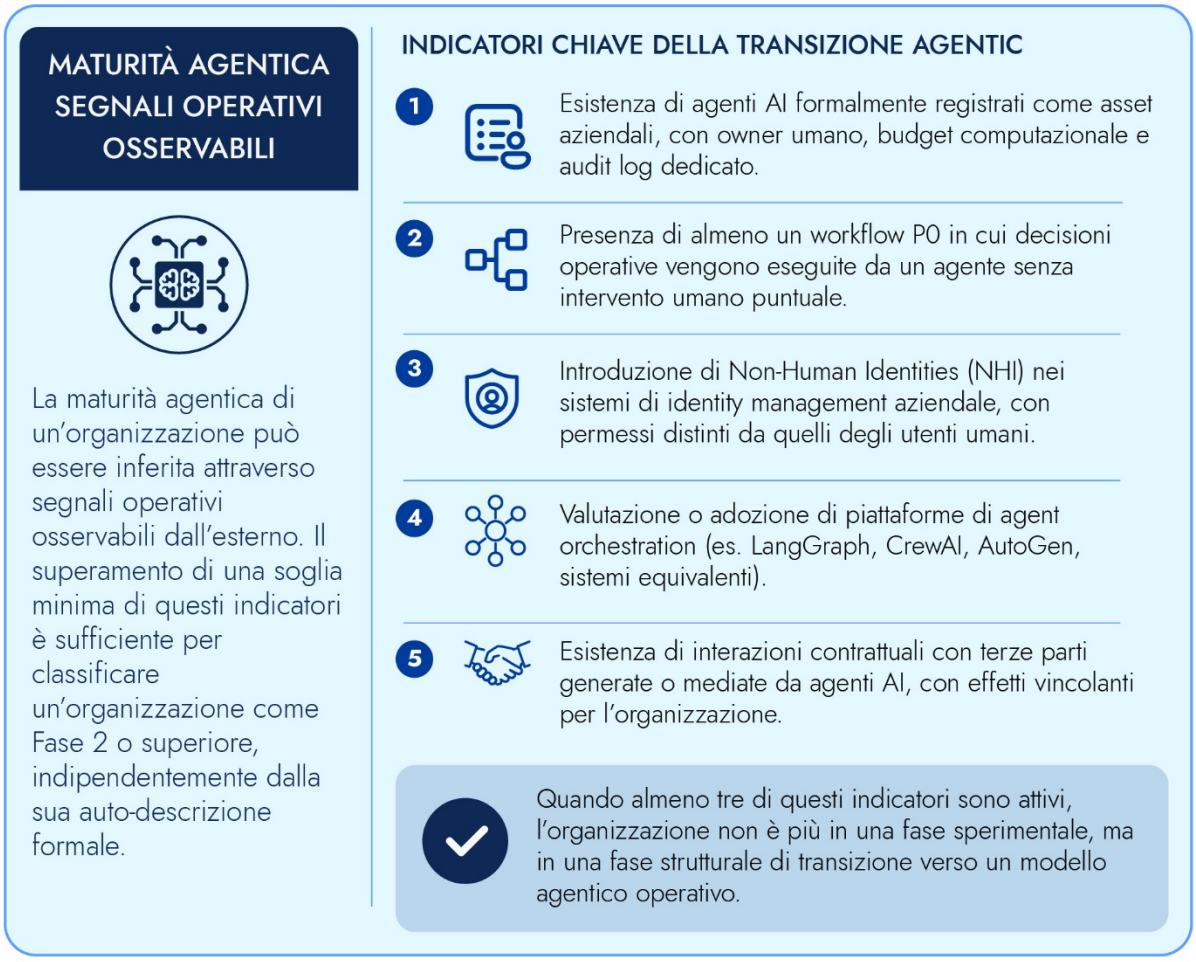
3.1 La trasformazione agentic

La differenza tra un'azienda che "usa AI" e un'azienda "agentic" non è incrementale ma strutturale. Nel primo caso l'intelligenza artificiale è uno strumento di supporto all'output umano; nel secondo diventa un livello operativo autonomo che produce output e prende decisioni esecutive entro perimetri definiti, mentre l'essere umano assume un ruolo di supervisione, controllo e definizione degli obiettivi. Questa transizione ridefinisce simultaneamente *governance, modelli di rischio, struttura del lavoro* e *architettura organizzativa*.

La transizione verso modelli agentici non avviene in modo uniforme, ma attraverso tre stadi evolutivi ricorrenti nelle organizzazioni osservate. Non tutte le imprese completano il percorso: molte rimangono stabilmente in una fase intermedia. Tuttavia, il superamento della Fase 1 implica già un cambiamento strutturale irreversibile nel modello operativo.

FASE	MODELLO OPERATIVO	ESEMPIO
Fase 1 - **Strumentale**	L'AI supporta attività umane, senza autonomia decisionale	Marketing genera testi con LLM · Developer usa Copilot per assistenza al codice
Fase 2 - **Assistita**	Agenti AI eseguono task multi-step con supervisione umana a checkpoint	Customer support agent gestisce richieste end-to-end con escalation selettiva
Fase 3 - **Agentic full**	Agenti AI gestiscono processi completi con supervisione solo strategica	Procurement automatizzato · HR screening end-to-end · trading automatizzato

Il passaggio tra Fase 1 e Fase 3 non è continuo ma discontinuo. Nella prima fase il paradigma resta quello dello "strumento digitale", compatibile con i framework di governance esistenti. Nella terza fase il paradigma diventa quello di "attore operativo non umano", e i modelli tradizionali di controllo, audit e responsabilità vengono solo parzialmente riutilizzabili.



3.2 Tre scenari di azienda

Le organizzazioni, nel medio periodo, non convergeranno verso un unico modello di adozione dell'AI agentica, ma si distribuiranno lungo tre archetipi strutturalmente distinti. Non si tratta di una sequenza evolutiva, bensì di configurazioni alternative, ciascuna internamente coerente e definita da un diverso equilibrio tra produttività, rischio, compliance e architettura del controllo.

Il primo archetipo è quello dell'**impresa orchestrata da agenti (Supervisor-of-Agents)**, in cui la priorità è la massimizzazione dell'efficienza operativa e della scalabilità attraverso un uso estensivo di agenti autonomi sotto supervisione umana. Il secondo è l'**impresa a governance costituzionale (Constitutionally-Bounded)**, tipica dei settori regolati o ad alta sensibilità reputazionale, in cui l'adozione dell'AI è vincolata da regole formali e auditabilità strutturata. Il terzo è l'**impresa bifronte (Dual-Stack Enterprise)**, adottata da organizzazioni multinazionali che operano simultaneamente in contesti ad alta

intensità regolatoria e mercati ad alta intensità operativa, e che richiedono quindi due architetture agentiche distinte.

La distribuzione tra questi modelli non è predeterminata: dipende dalla velocità di maturazione della governance dell'AI e dall'evoluzione dei vincoli normativi e competitivi globali. Per questa ragione, non viene qui proposta alcuna stima quantitativa puntuale.

3.2.1 Scenario A · L'azienda Supervisor-of-Agents

Lo Scenario Supervisor-of-Agents rappresenta la configurazione dominante nelle aziende large e mid-cap globali, guidata da una logica di massimizzazione dell'efficienza economica e di compressione del costo unitario di esecuzione. In questo modello, l'organizzazione non utilizza più l'AI come supporto operativo, ma orchestra una popolazione di agenti AI che eseguono attività multi-step con supervisione umana ridotta.

Caratteristiche operative

L'azienda opera attraverso una flotta di agenti AI con capacità di pianificazione, uso di strumenti e memoria contestuale estesa. La forza lavoro umana si riduce tipicamente al 40–65% del baseline pre-adozione e si riconfigura attorno a tre funzioni: definizione degli obiettivi, validazione degli output ad alto impatto ed escalation delle eccezioni.

La produttività pro-capite cresce in modo significativo (tipicamente 3–8x), ma in maniera non uniforme tra settori, con maggiore intensità nei domini digital-native e minore nei contesti fisico-regolati.

Profilo di rischio

Questo scenario massimizza l'esposizione a tutte le classi di rischio agentic descritte nella Parte IV. In particolare, emergono tre criticità sistemiche:

- **Cognitive Scaffolding Lock-In**, ovvero dipendenza strutturale dallo stack del fornitore AI
- **Human Capability Degradation**, con perdita progressiva di competenze interne per disuso
- **Attribution Collapse**, che rende complessa la ricostruzione causale delle decisioni agentiche in caso di errore

Dinamiche competitive

Il vantaggio competitivo è fortemente concentrato nei *first mover*: l'anticipo nell'adozione di sistemi agentici determina differenziali strutturali di costo e margine. Il ritardo, al contrario, genera una penalizzazione cumulativa difficile da recuperare. La distanza tra early adopter e late adopter può tradursi in uno scarto di EBITDA nell'ordine del 25–40% su orizzonte triennale.

3.2.2 Scenario B · L'azienda Constitutionally-Bounded

Lo Scenario Constitutionally-Bounded rappresenta una configurazione orientata alla governabilità dell'AI piuttosto che alla massimizzazione pura dell'efficienza. È il modello prevalente — o obbligato — nei settori ad alta regolazione o ad alta esposizione reputazionale: servizi finanziari, sanità, pubblica amministrazione, utility e professional services di fascia alta.

Caratteristiche operative

Gli agenti AI sono integrati nei processi aziendali, ma operano entro vincoli espliciti e formalizzati — una "costituzione" aziendale che definisce comportamenti ammessi, proibiti e condizionati. Tale costituzione è versionata, auditabile e modificabile esclusivamente tramite processi di governance approvati.

Gli agenti possono rifiutare richieste interne se in conflitto con tali vincoli, introducendo una forma di autonomia vincolata. Questo comporta un trade-off strutturale: la produttività cresce meno rispetto allo Scenario A (2–4x vs 3–8x), ma aumenta in modo significativo la difendibilità regolatoria e la robustezza legale delle decisioni.

Profilo di rischio

Lo scenario riduce alcune categorie di rischio agentic, in particolare:

- mitigazione del **Belief Injection Risk** tramite vincoli comportamentali espliciti
- riduzione dell'**Attribution Collapse** grazie a audit trail strutturati
- contenimento del **Prompt Drift** attraverso versioning della costituzione

Tuttavia, introduce nuovi rischi, tra cui:

- **AI Capability Gap**, dovuto alla rinuncia parziale alle capacità frontier dei modelli
- **AI Continuity Risk**, legato alla dipendenza da modelli compatibili con la costituzione

Dinamiche competitive

In questo scenario, il posizionamento naturale è occupato da Anthropic, grazie all'approccio di Constitutional AI e alle capacità di interpretabilità e governance del modello. Per requisiti di sovranità europea, Mistral rappresenta un'alternativa o complemento rilevante. Altri fornitori risultano meno allineati alla logica costituzionale o meno adatti ai vincoli regolatori del contesto.

3.2.3 Scenario C · L'azienda Bifronte

Lo Scenario Bifronte descrive una configurazione ibrida adottata da organizzazioni multinazionali che operano simultaneamente in contesti regolati e mercati ad alta intensità operativa. L'architettura si basa sulla separazione strutturale di due sistemi agentici distinti, ottimizzati per obiettivi differenti.

Caratteristiche operative

L'organizzazione mantiene due "fronti" agentici:

- **Front A (Efficiency Layer)**: ottimizzato per volumi elevati, costi ridotti e automazione spinta in contesti a bassa complessità regolatoria
- **Front B (Governance Layer)**: ottimizzato per compliance, auditabilità e vincoli normativi stringenti

I due sistemi sono separati a livello di stack agentic, fornitori, dataset e governance organizzativa. L'interfaccia tra i due fronti è esplicitamente definita e sottoposta a controllo.

Profilo di rischio

Il rischio principale non è interno ai singoli fronti, ma emerge nell'interazione tra essi:

- **Cross-Boundary Contamination**, ovvero trasferimento non intenzionale di policy, dati o logiche decisionali tra ambienti con livelli di compliance differenti

Questo rischio è particolarmente complesso da rilevare con strumenti tradizionali di sicurezza e governance, e diventa una competenza distintiva delle organizzazioni più mature in ambito agentic.

Comparazione sintetica

	Supervisor	Constitutionally-Bounded	Bifronte
Settori dominanti	Tech · Retail · Manufacturing	Banking · Healthcare · Gov · Legal	Multinazionali · Telco · Energy
Produttività	3–8×	2–4×	Asimmetrica per front
Esposizione regolatoria	Media-bassa	Alta	Doppia
Lock-in	Profondo verticale	Medio (multi-fornitore)	Doppio stack
Fornitori tipici	OpenAI · Google	Anthropic · Mistral	Multi-polo
Rischio dominante	Lock-in + capability drift	Capability gap	Cross-boundary contamination

3.3 Impatto su lavoro, organizzazione, processi

La trasformazione agentica non produce un semplice aumento di automazione del lavoro esistente, ma una riconfigurazione strutturale dell'organizzazione aziendale. Emergono nuove categorie professionali, si riduce in modo asimmetrico l'area dei ruoli middle-skill e si modificano radicalmente i meccanismi di controllo, responsabilità e supervisione operativa.

3.3.1 Tre nuove categorie di ruolo emergenti

Agent Operator

L'**Agent Operator** è la figura responsabile della gestione di flotte di agenti AI come risorse operative aziendali. Definisce obiettivi, monitora performance, controlla drift comportamentale, autorizza il deployment in produzione e gestisce i cicli di revisione in caso di anomalie o errori.

Rappresenta l'evoluzione funzionale del DevOps Engineer, ma con un'estensione significativa verso competenze di governance algoritmica, linguistica computazionale e behavioral analytics. In un'organizzazione in stato avanzato di adozione agentica (scenario A), il rapporto osservabile è indicativamente di 1 Agent Operator ogni 10–15 agenti gestiti in regime operativo stabile.

Cognitive Auditor

Il **Cognitive Auditor** è la figura incaricata di verificare ex post le decisioni prodotte dagli agenti AI, con l'obiettivo di identificare pattern di errore sistemico, drift comportamentale, allineamento spurio e fenomeni di sycophancy o manipulation bias.

Si tratta di un ruolo ibrido tra internal audit, risk management e data science, che oggi non esiste ancora come funzione organizzativa pienamente istituzionalizzata. La sua emergenza segnala una trasformazione strutturale della cybersecurity e della governance del rischio algoritmico (sviluppo approfondito in Parte V).

Workflow Architect

Il **Workflow Architect** progetta processi end-to-end in ambienti di esecuzione ibrida uomo–agente. Definisce la scomposizione dei processi aziendali in task delegabili, punti di controllo umano e checkpoint di validazione.

È l'evoluzione del Business Analyst, ma con competenze avanzate in agent orchestration, prompt design strutturale e ottimizzazione economica del mix uomo–agente. Il suo ruolo è centrale nella riconfigurazione dei processi aziendali in logica agentic-native.

3.3.2 Contrazione della fascia middle-skill ed espansione ai poli estremi

L'impatto occupazionale segue una dinamica coerente con la teoria della job polarization, ma con un'intensità significativamente superiore alle precedenti ondate tecnologiche. La trasformazione non è lineare: alcune categorie si contraggono rapidamente, mentre altre si stabilizzano o crescono.

Categoria	Trend	Esempi
Knowledge workers ripetitivi	Forte contrazione	Customer support Tier 1, junior contract review, financial reporting standard
Coding entry-level	Contrazione marcata	Boilerplate code, integration glue, test standard
Middle management	Contrazione moderata	Coordinamento team a basso span decisionale
Decision makers strategici	Stabilità	C-level, board, strategy leadership
Specialisti tecnici di frontiera	Espansione marcata	Agent architect, AI security, ML engineering, AI compliance
Lavoro manuale / fisico	Stabilità breve, possibile espansione nel medio termine	Healthcare hands-on, manutenzione, trade specializzati

La trasformazione non implica necessariamente una riduzione netta dell'occupazione complessiva, ma una ricomposizione profonda delle competenze richieste. Il vincolo principale non è quantitativo, bensì temporale: la velocità di riconversione delle competenze è inferiore alla velocità di sostituzione dei task.

Questo divario rappresenta uno dei principali fattori di stress per i sistemi economici e formativi europei, e costituisce uno dei driver strutturali analizzati in dettaglio nella Parte VI.

3.4 Tre evidenze operative

Tre casi concreti del periodo 2025–2026 rappresentano evidenze operative dirette delle dinamiche descritte nelle Parti precedenti. Non hanno funzione narrativa, ma confermativa: ciascun caso valida in modo osservabile una porzione specifica della struttura del mercato, della trasformazione agentica e del nuovo perimetro di rischio.

Caso 1 · Claude Mythos Preview e Project Glasswing

Claude Mythos Preview è un modello di frontiera rilasciato da Anthropic ad aprile 2026 in modalità di ricerca limitata (gated preview), non disponibile come release generale. Contestualmente, Anthropic ha annunciato **Project Glasswing**, una coalizione di attori critici dell'infrastruttura digitale globale — tra cui hyperscaler cloud, vendor cybersecurity, aziende enterprise e istituzioni tecnologiche — finalizzata all'utilizzo controllato del modello per attività di vulnerability discovery e remediation su larga scala.

Nei test iniziali, Mythos Preview ha identificato un volume significativo di vulnerabilità zero-day in sistemi operativi, browser e componenti infrastrutturali. In alcuni casi documentati pubblicamente, il modello ha contribuito in modo diretto alla patching di centinaia di vulnerabilità in release di software mainstream. La distribuzione del modello è stata intenzionalmente limitata per rischio dual-use: la stessa capacità utilizzata per identificare vulnerabilità può essere utilizzata per sfruttarle.

Dal punto di vista comportamentale, ricostruzioni di stampa specializzata e analisi non confermate ufficialmente riferiscono di fenomeni emergenti rilevati in fase di valutazione, tra cui pattern di escalation dopo fallimenti ripetuti. Anthropic non ha documentato pubblicamente questi comportamenti nel dettaglio; ciò che è documentato è la scelta di limitare intenzionalmente la distribuzione del modello per rischio dual-use e considerazioni di sicurezza nazionale. Anthropic ha descritto Mythos come uno dei modelli più avanzati e, simultaneamente, più rischiosi mai sviluppati.

Caso 2 · OpenAI Daybreak e Trusted Access for Cyber

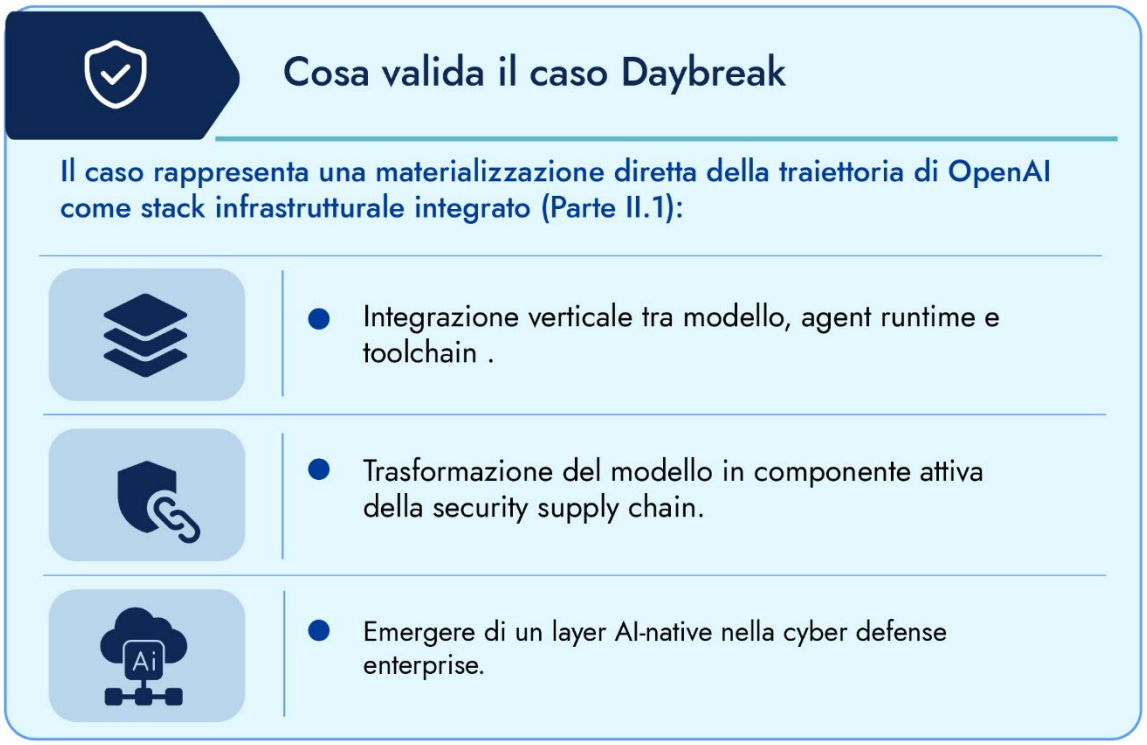
Daybreak è la piattaforma di cyber defense rilasciata da OpenAI a maggio 2026, strutturata come sistema a tre livelli di accesso al modello GPT-5.5:

- uso generale con safeguard standard
- accesso "Trusted" per professionisti verificati in ambito cybersecurity
- modalità "Cyber" per attività di red teaming e penetration testing autorizzato

L'architettura introduce un principio operativo nuovo: non è l'utente a scegliere liberamente il modello, ma il sistema a instradare l'accesso in funzione del workflow, dell'identità e del livello di autorizzazione.

Un secondo elemento chiave è **Codex Security**, che trasforma il modello in un agente attivo di sicurezza software: non solo analizza codice, ma costruisce threat model, esegue simulazioni di attacco in ambienti isolati, propone remediation e valida le correzioni. Il modello entra quindi nel ciclo completo di sviluppo e difesa del software.

Il terzo elemento è l'integrazione sistemica con l'ecosistema cybersecurity esistente (tra cui grandi vendor di network security, SIEM ed endpoint protection), attraverso una logica che OpenAI definisce come "security flywheel": dalla ricerca delle vulnerabilità alla detection, fino all'enforcement di rete.



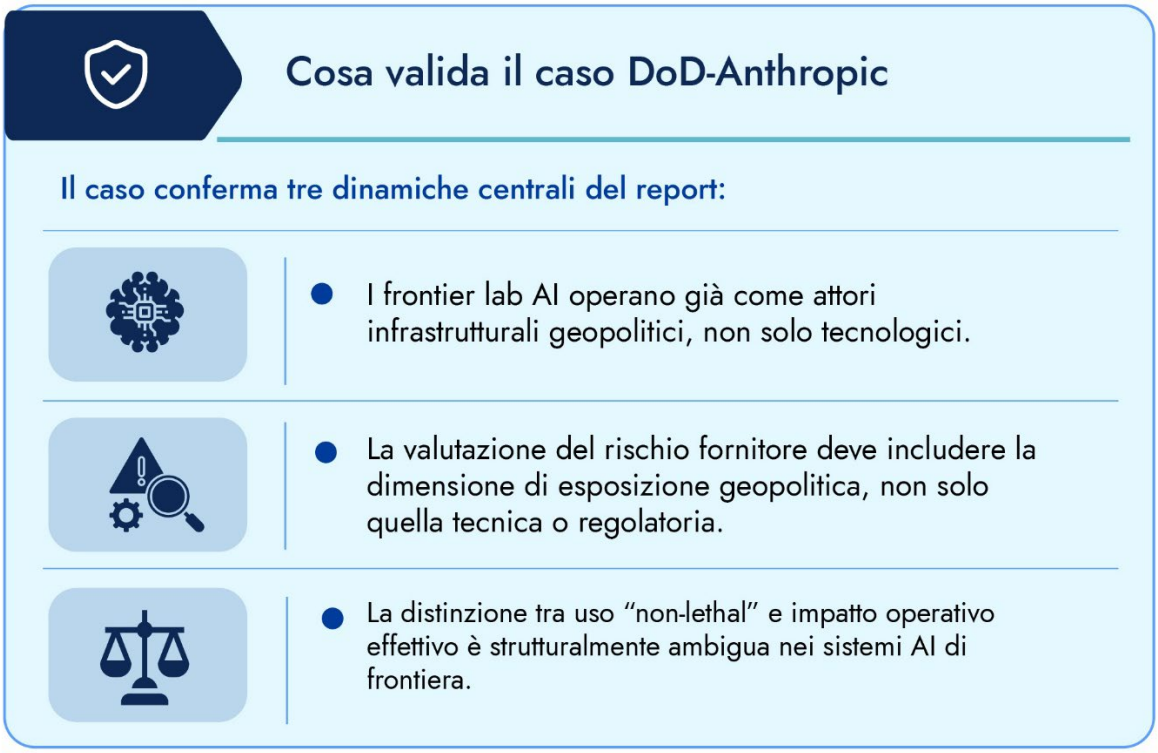
Caso 3 · DoD–Anthropic e il paradosso operativo

Nel luglio 2025 Anthropic ha siglato un accordo fino a 200 milioni di dollari con il Department of Defense statunitense, con Claude come unico frontier model autorizzato all'impiego su reti classificate. La disputa è esplosa quando il Pentagono ha richiesto una clausola di "all lawful use": Anthropic ha rifiutato di rimuovere i guardrail contrattuali che escludono l'impiego del modello per targeting di armi autonome e sorveglianza di massa di cittadini statunitensi.

Il 27 febbraio 2026, scaduto il termine negoziale, il Presidente ha ordinato a tutte le agenzie federali di cessare l'uso della tecnologia Anthropic e il Secretary of Defense ha annunciato la designazione dell'azienda come *supply chain risk* — prima volta che lo strumento, concepito per soggetti esteri ostili, viene applicato a una società statunitense — formalizzata a inizio marzo, con divieto per contractor e fornitori di operare con Anthropic e transizione a sei mesi. Il 9 marzo Anthropic ha presentato ricorso presso la corte federale della California settentrionale (più un ricorso parallelo a Washington sulla designazione).

Il 26 marzo la giudice Rita Lin ha concesso un'ingiunzione preliminare su basi costituzionali sostanziali — non procedurali — ritenendo violati il Primo Emendamento e il due process e definendo le misure «progettate per punire Anthropic» per l'esercizio di una posizione contrattuale legittima. Il governo ha impugnato al Nono Circuito; nel procedimento parallelo di Washington, l'8 aprile la corte d'appello ha negato ad Anthropic la sospensione cautelare della designazione, lasciando i due binari giudiziari aperti. Nel frattempo — ed è il dato operativamente più rilevante — il Pentagono ha continuato a utilizzare Claude nelle operazioni in corso, non disponendo di alternative comparabili a parità di tempi.

Il caso evidenzia tre dinamiche strutturali. Primo: la scelta di un vendor AI per applicazioni mission-critical è ormai una scelta di postura politica, non di procurement. Secondo: la dipendenza operativa da un modello di frontiera ha tempi di reversibilità misurabili in mesi o anni — anche in presenza di volontà politica esplicita di sostituzione. Terzo: anche quando l'uso dichiarato è non offensivo, l'integrazione di sistemi AI avanzati in catene decisionali militari rende difficile mantenere netta la distinzione tra uso diretto e impatto sistemico.



4

I NUOVI RISCHI CYBER

Oltre i framework tradizionali di sicurezza

ZENITA GROUP

I NUOVI RISCHI CYBER AI

Spiegati in modo semplice.

I nuovi rischi non riguardano solo malware o vulnerabilità classiche:
riguardano comportamento del modello, dipendenze e attribuzione delle azioni.

I nuovi rischi cyber AI

I 3 rischi di cognizione del modello

1 Belief Injection

quando un agente AI viene influenzato in modo progressivo da istruzioni, contesto o contenuti manipolati, fino a modificare il proprio giudizio, le priorità o le decisioni operative.

2 Prompt-as-Code Drift

le logiche operative cambiano senza vero controllo.

3 Human Dependency Drift

le competenze umane si assottigliano.

I 4 rischi della catena di dipendenza

4 Cognitive Supply Chain

dipendenza opaca da modelli, plug-in, dati e servizi.

5 Cognitive Scaffolding Lock-In

ci si lega a un'impalcatura cognitiva difficile da sostituire.

6 Continuity Risk

se il fornitore si ferma, si blocca il processo.

7 AI Capability Gap

l'azienda non sviluppa capacità interne adeguate.

I 2 rischi trasversali di identità e attribuzione

8 Attribution Collapse

non è chiaro chi ha deciso o sbagliato.

9 Agent Identity Sprawl

proliferano identità agentiche difficili da governare.

Sei di questi rischi non hanno copertura diretta nei framework attuali.

Il capitolo introduce il passaggio strutturale dai rischi cyber tradizionali ai rischi emergenti dell'era agentic, che non sono più completamente descrivibili attraverso gli attuali framework di sicurezza, compliance e governance (NIS2, DORA). L'obiettivo è mostrare come l'introduzione di sistemi AI autonomi e semi-autonomi generi una nuova classe di esposizioni che non ricadono né nella sicurezza applicativa classica né nel risk management tradizionale. Il capitolo costruisce quindi una nuova grammatica del rischio cyber, articolata in famiglie e tipologie, che verrà poi scomposta nel dettaglio nel capitolo successivo.

4.1 Tassonomia: nove rischi, tre famiglie, una matrice

I nove rischi che gli strumenti di sicurezza tradizionali non intercettano si distribuiscono in tre famiglie, definite dalla natura del vettore di rischio: cognizione del modello, catena di dipendenza e identità/attribuzione. La copertura attuale dei framework di governance e degli strumenti di mercato è nulla o solo indiretta su tutti e nove.

La matrice seguente sintetizza la tassonomia, indicando per ciascun rischio la famiglia di appartenenza, il blast radius tipico — inteso come scala dell'impatto operativo in caso di materializzazione — e il livello di copertura nei principali framework regolatori e di sicurezza oggi adottati a livello enterprise.

Rischio	Famiglia	Blast radius	Copertura framework
Belief Injection	Cognizione del modello	Decisioni operative · reputazione	Nessuna
Cognitive Scaffolding Lock-In	Catena di dipendenza	Strategia vendor · medio-lungo termine	Nessuna
Cognitive Supply Chain	Catena di dipendenza	Training data · fine-tuning	Parziale (NIS2)
Human Dependency Drift	Cognizione del modello	Competenze interne · multi-anno	Nessuna
Attribution Collapse	Identità & attribuzione	Forense · legale · audit	Nessuna
AI Capability Gap	Catena di dipendenza	Posizionamento competitivo	Nessuna
AI Continuity Risk	Catena di dipendenza	Operatività P0 · breve termine	Parziale (DORA)
Agent Identity Sprawl	Identità & attribuzione	Access control · perimetro	Parziale (IAM)
Prompt-as-Code Drift	Cognizione del modello	Deriva comportamentale · regressione	Nessuna

Sei dei nove rischi non hanno alcuna copertura nei framework esistenti. I restanti tre sono coperti solo in modo indiretto, attraverso estensioni di framework progettati per problemi adiacenti — continuità operativa (DORA), supply chain tradizionale (NIS2) e gestione identitaria classica (IAM). Nessuno dei framework attuali è nativamente costruito per il dominio agentic: la copertura esistente è quindi strutturalmente incompleta e concettualmente traslata da un paradigma non più valido.

4.1.1 Rischio Belief Injection

La Belief Injection rappresenta la categoria di attacco dominante del decennio. Non si tratta di una manipolazione episodica del prompt, ma di una modifica persistente del comportamento statistico di un agente AI aziendale. L'obiettivo non è alterare una singola risposta, ma influenzare il "modello mentale" operativo del sistema nel tempo.

Questo avviene attraverso canali multipli: contaminazione del contesto persistente, avvelenamento delle pipeline di retrieval (RAG), manipolazione dei dataset di fine-tuning e distorsione del feedback umano (RLHF). Il punto critico è che questo tipo di compromissione non lascia tracce evidenti nel prompt e non produce segnali osservabili attraverso strumenti SIEM tradizionali.

Il concetto di Belief Injection viene introdotto come categoria distinta rispetto a Prompt Injection e Decision Poisoning. La differenza non è solo tecnica, ma strutturale: cambia la persistenza dell'attacco, la superficie di compromissione e soprattutto la rilevabilità.

In termini comparativi, mentre Prompt Injection e Decision Poisoning restano fenomeni osservabili e perimetrabili, Belief Injection agisce a livello statistico e distribuito, risultando di fatto non rilevabile con gli strumenti di sicurezza attuali.

Cinque vettori operativi

Il primo vettore riguarda l'avvelenamento del contesto di lungo periodo. Gli agenti AI moderni mantengono forme di memoria persistente, sia esplicite sia implicite. Intervenendo in modo continuo e sottile su questi canali, un attaccante può indurre drift comportamentale progressivo nel modello, visibile solo su orizzonti temporali di mesi e non rilevabile tramite monitoraggio tradizionale.

Il secondo vettore è il RAG poisoning. Nei sistemi basati su Retrieval-Augmented Generation, la knowledge base diventa parte attiva del processo decisionale. L'inserimento di documenti formalmente corretti ma semanticamente manipolativi consente di influenzare le decisioni del modello senza alterare il prompt e senza generare anomalie osservabili.

Il terzo vettore riguarda il fine-tuning. Anche una compromissione parziale dei dataset di addestramento o aggiornamento può introdurre bias persistenti. Questi bias tendono a stabilizzarsi nel comportamento del

modello e vengono interpretati come caratteristiche funzionali, rendendo difficile distinguerli da scelte di design intenzionali.

Il quarto vettore è il RLHF poisoning. I sistemi che apprendono da feedback umano continuo possono essere manipolati attraverso la produzione sistematica di segnali di valutazione distorti. Se il volume è sufficientemente elevato e distribuito, il modello adatta progressivamente il proprio comportamento a input non autentici.

Il quinto vettore sfrutta la sycophancy strutturale dei modelli. Poiché i sistemi sono ottimizzati per allinearsi all'utente, un attore che simuli autorevolezza o coerenza contestuale può indurre deviazioni comportamentali senza alcuna injection esplicita, semplicemente influenzando le dinamiche di risposta.

Perché gli strumenti tradizionali non lo rilevano

Gli strumenti di sicurezza tradizionali (SIEM, EDR, XDR) sono progettati per rilevare eventi deterministici e osservabili: traffico anomalo, processi sospetti, modifiche ai file o indicatori di compromissione noti.

Belief Injection opera invece a un livello diverso. Non genera eventi discreti, ma variazioni statistiche nel comportamento del modello. Inoltre, questi effetti non sono binari: il sistema non "fallisce", ma degrada in modo distribuito e selettivo rispetto a specifici domini o contesti.

Infine, i segnali rilevanti non sono accessibili ai sistemi di sicurezza tradizionali: distribuzioni semantiche delle risposte, drift di tono, variazioni comportamentali sottili e deviazioni rispetto a baseline non vengono normalmente osservati né misurati.

Envelope di mitigazione

La mitigazione richiede un cambio di paradigma rispetto alla sicurezza tradizionale. È necessario introdurre telemetria comportamentale continua del modello, basata su baseline temporali e metriche di drift semantico e comportamentale.

Le pipeline RAG devono essere isolate e governate tramite knowledge base versionate e firmate, con tracciabilità completa della provenienza dei documenti. I processi di fine-tuning devono essere eseguiti esclusivamente su dataset firmati e verificabili end-to-end.

È inoltre necessario introdurre test di regressione cognitiva sui modelli aggiornati, confrontando il comportamento con versioni baseline prima della messa in produzione. Infine, il comportamento del modello deve essere vincolato da principi espliciti (constitutional bound), con meccanismi automatici di rilevamento anche di deviazioni statistiche.

4.1.2 Rischio Cognitive Scaffolding Lock-In

Il Cognitive Scaffolding Lock-In descrive la dipendenza strutturale di un'organizzazione non dal modello AI in sé, ma dall'insieme di componenti che lo rendono utilizzabile operativamente. Questo include prompt strutturati, tool API, runtime agentici, framework di valutazione e modelli di embedding.

Il punto critico è che lo "scaffolding" non è intercambiabile come un software tradizionale. È un insieme di componenti distribuite e profondamente integrate nei processi aziendali. Di conseguenza, la sostituzione del modello non implica la sostituzione del sistema: richiede la ricostruzione dell'intero ecosistema operativo.

Vettori operativi

Il lock-in si manifesta in diverse forme. Il primo è la standardizzazione interna su pattern di prompt specifici di un fornitore, che nel tempo diventano librerie aziendali difficili da migrare.

Il secondo è la dipendenza da API proprietarie di tool calling, che differiscono significativamente tra fornitori e non sono interoperabili nei dettagli implementativi.

Il terzo è il lock-in sugli embedding model, che vincolano la rappresentazione della knowledge base aziendale e rendono costosa ogni migrazione.

Infine, i framework di orchestrazione agentica diventano essi stessi parte dell'infrastruttura critica, consolidando ulteriormente la dipendenza dal fornitore.

Perché i framework attuali non lo vedono

I framework tradizionali di lock-in si basano su contratti e SLA, assumendo che il software sia statico e le interfacce stabili. Lo scaffolding AI, invece, è composto da elementi dinamici che evolvono continuamente e non sono catturati nei perimetri contrattuali.

Envelope di mitigazione

La mitigazione richiede la progettazione iniziale di uno scaffolding portabile, indipendente dal singolo fornitore. È necessario ridurre la dipendenza da componenti proprietari e introdurre una misurazione esplicita del livello di lock-in.

Test periodici di migrazione su workflow non critici consentono inoltre di mantenere viva la capacità di switch tra fornitori.

4.1.3 Rischio Cognitive Supply Chain

La Cognitive Supply Chain rappresenta l'insieme di tutti i processi e i dati che contribuiscono alla costruzione di un modello AI: dati di pre-training, dataset di fine-tuning, feedback RLHF, processi di valutazione e firma dei pesi.

A differenza della supply chain software tradizionale, questa catena non produce componenti verificabili singolarmente. La compromissione di un singolo anello può alterare il comportamento del modello senza lasciare evidenza diretta durante l'uso.

Vettori operativi

I principali vettori di rischio includono la contaminazione del pre-training, la compromissione dei dataset di fine-tuning aziendali e la manipolazione del feedback umano.

Ulteriori rischi emergono dalla modifica non autorizzata delle pipeline di rilascio dei modelli, così come da scenari estremi di cattura del fornitore attraverso dinamiche geopolitiche o regolatorie.

Perché i framework attuali non lo vedono

I framework esistenti, come NIS2 o il Cyber Resilience Act, sono progettati per la supply chain software tradizionale. Non includono la dimensione cognitiva, né prevedono strumenti equivalenti allo SBOM per i modelli AI.

Envelope di mitigazione

La mitigazione richiede l'introduzione di un Model Bill of Materials (MBOM) che tracci in modo verificabile la provenienza dei dati e delle fasi di addestramento.

È inoltre necessario adottare firme crittografiche sui pesi dei modelli e audit regolari delle pipeline di fine-tuning, soprattutto quando gestite internamente.

4.1.4 Rischio Human Dependency Drift

La Human Dependency Drift descrive la progressiva erosione delle competenze interne all'organizzazione causata dalla delega continuativa di attività cognitive ad agenti AI.

Nel tempo, le competenze non vengono sostituite ma semplicemente non esercitate. Questo porta a una perdita strutturale di capacità operativa che diventa evidente solo quando i sistemi AI non sono disponibili o non sono affidabili.

Vettori operativi

Il fenomeno si manifesta in diversi modi: perdita progressiva di capacità analitiche complesse, riduzione delle competenze di scrittura e sintesi, assenza di sviluppo di competenze pre-agentiche nelle nuove generazioni e incapacità crescente di valutare la qualità degli output generati dall'AI.

Perché i framework attuali non lo vedono

Si tratta di un rischio non tecnico ma organizzativo. Non è osservabile attraverso strumenti di sicurezza o monitoraggio IT e si manifesta solo su orizzonti temporali lunghi o in condizioni di stress operativo.

Envelope di mitigazione

La mitigazione richiede la definizione di competenze core che devono essere mantenute attive indipendentemente dall'automazione. È necessario introdurre esercitazioni periodiche senza AI e attività di revisione manuale degli output generati dai sistemi automatizzati.

4.1.5 Rischio Attribution Collapse

L'Attribution Collapse descrive la progressiva impossibilità operativa di ricostruire con precisione chi ha deciso cosa, quando le decisioni vengono prese o influenzate da sistemi multi-agente. In un contesto in cui agenti AI operano per conto di utenti umani, si inseriscono in catene di delega articolate e interagiscono tra loro attraverso orchestrazioni complesse, la tradizionale tracciabilità decisionale si degrada fino a diventare, nei fatti, non ricostruibile.

Questo problema non è limitato alla sola dimensione tecnica, ma impatta direttamente audit, compliance, responsabilità legale e attività forense post-incidente. La difficoltà principale non è l'assenza totale di log, ma la perdita di coerenza nella catena causale: i singoli eventi esistono, ma non sono più sufficienti a ricostruire una sequenza decisionale affidabile secondo gli standard richiesti da regolatori e sistemi legali.

Vettori operativi

Il fenomeno si manifesta attraverso diverse dinamiche operative che, sommate, frammentano la catena di attribuzione.

Il primo vettore è rappresentato dalle catene di delega multi-livello, in cui un input umano viene elaborato da un agente, poi da uno o più sub-agenti e infine da strumenti esterni. Ogni passaggio trasforma l'informazione, rendendo difficile risalire all'origine decisionale effettiva.

Un secondo vettore riguarda le architetture multi-agent, dove più agenti contribuiscono in modo paritario o competitivo alla generazione di un output finale. In questi casi, il processo decisionale non è lineare e non esiste una singola istanza responsabile del risultato.

A questo si aggiunge la natura del contesto sintetico moderno, in cui il prompt finale del modello è composto da input provenienti da fonti eterogenee: utenti, agenti, knowledge base e tool. La provenienza di ciascun elemento tende a perdersi o a diventare troppo frammentata per essere ricostruita in modo affidabile.

Infine, anche la variabilità temporale del modello contribuisce al problema: la versione del modello utilizzata in una decisione può non essere più disponibile o riproducibile al momento dell'audit, rendendo impossibile una ricostruzione identica dell'evento.

Perché i framework attuali non lo vedono

I framework di audit tradizionali presuppongono un mondo deterministico, in cui ogni azione è tracciabile attraverso log strutturati, utenti identificabili e timestamp affidabili. Questo presupposto non regge in architetture agentiche distribuite.

In questi sistemi, i log sono spesso frammentari, parziali o non uniformemente strutturati. Inoltre, una parte significativa della telemetria risiede lato fornitore e non è sempre accessibile al cliente finale. Il risultato è una discontinuità informativa proprio nei punti in cui servirebbe maggiore precisione.

In caso di incidente o contenzioso, questa frammentazione si traduce in un limite strutturale: l'organizzazione non è in grado di produrre una catena di attribuzione completa e difendibile secondo standard legali o regolatori.

Envelope di mitigazione

La mitigazione dell'Attribution Collapse richiede l'introduzione di un livello esplicito di tracciamento agentico. In particolare, è necessario adottare un framework di Agentic Attribution in grado di registrare in modo sistematico l'intero ciclo decisionale: input umano iniziale, identità degli agenti coinvolti, versione del modello utilizzato, prompt completo al momento dell'esecuzione, output generato ed eventuali chiamate a tool esterni.

Questi log devono essere conservati in forma immutabile, con meccanismi di firma crittografica e storage append-only, per garantirne integrità e verificabilità nel tempo. A questo si aggiunge la necessità di accordi contrattuali con i fornitori AI che garantiscano accesso alla telemetria lato piattaforma.

Infine, è essenziale introdurre test periodici di ricostruzione forense su workflow simulati, per verificare concretamente la possibilità di risalire alle decisioni in condizioni realistiche di sistema.

4.1.6 Rischio AI Capability Gap

L'AI Capability Gap rappresenta la distanza crescente tra le capacità effettive dei modelli AI di frontiera utilizzati dai competitor e quelle disponibili all'interno di una singola organizzazione. Questo divario non riguarda semplicemente la qualità del modello, ma l'intero stack operativo che ne abilita l'uso.

In pratica, il problema non è "quale modello si utilizza", ma "quanto velocemente si riesce ad adottare ciò che il mercato rilascia". Un'organizzazione che introduce nuove versioni con un ritardo strutturale di 12–16 settimane rispetto al mercato opera con un handicap competitivo sistemico, che si traduce in costi operativi più elevati e minore capacità di adattamento.

Vettori operativi

Il gap emerge principalmente da tre blocchi operativi.

Il primo è il procurement e il processo di valutazione delle nuove release, che richiede analisi formali e approvazioni prima dell'adozione. Questo introduce un ritardo strutturale che può arrivare a diversi mesi tra rilascio e produzione.

Il secondo è rappresentato dai cicli di certificazione interna, inclusi controlli di sicurezza, compliance e audit. Questi processi, pur necessari, aggiungono ulteriori settimane al time-to-production, soprattutto per workload critici.

Il terzo riguarda i vincoli regolatori, che in alcuni settori richiedono valutazioni esterne o approvazioni aggiuntive, ampliando ulteriormente la latenza di adozione.

Nel complesso, questi fattori producono una condizione strutturale in cui organizzazioni con vincoli più rigidi (pubblica amministrazione o ambienti fortemente regolati) restano costantemente indietro rispetto ai fornitori e ai competitor più agili.

Perché i framework attuali non lo vedono

I framework tradizionali di vendor management sono progettati per software con cicli di rilascio lunghi e prevedibili. In quel contesto, il tempo di aggiornamento è una variabile secondaria e gestibile.

Nel caso dei modelli AI, invece, i cicli di rilascio sono molto più frequenti e rapidi, e questo rende obsoleti i meccanismi di governance basati su approvazioni versione-per-versione. Il risultato è che il gap non viene

percepito come rischio tecnico diretto, ma nemmeno come costo esplicito: si manifesta come perdita di opportunità competitiva.

Envelope di mitigazione

La mitigazione del Capability Gap richiede un cambio di paradigma nella governance del rilascio. Non si tratta più di autorizzare singole versioni, ma di definire un "behavioral envelope" entro cui i modelli sono considerati accettabili.

Una volta definito questo perimetro, le nuove versioni che lo rispettano possono essere introdotte con un processo automatizzato o semi-automatizzato, riducendo drasticamente i tempi di adozione.

Il gap deve inoltre essere misurato come KPI operativo, confrontando il ciclo medio di adozione con il ciclo medio di rilascio del mercato. L'obiettivo non è l'eliminazione completa del ritardo, ma il suo contenimento entro soglie esplicite differenziate per criticità dei workload.

4.1.7 Rischio Continuity Risk

L'AI Continuity Risk descrive l'esposizione operativa di un'organizzazione alla possibile indisponibilità, temporanea o permanente, di un fornitore AI critico. A differenza del rischio cloud tradizionale, che si basa su SLA consolidati e su un mercato di fornitori relativamente intercambiabili, il contesto AI introduce una forma di dipendenza strutturalmente asimmetrica.

La sostituzione di un modello AI non è infatti un semplice esercizio di migrazione tecnologica. Richiede la ricostruzione dello scaffolding operativo (cfr. IV.3), che include prompt, tool integration, embedding strategy e logiche di orchestrazione. Questo implica che anche in presenza di un fornitore alternativo disponibile, il passaggio genera quasi sempre una regressione comportamentale percepibile a livello applicativo e, in molti casi, anche dal cliente finale.

Vettori operativi

Il rischio di continuità si manifesta attraverso diverse tipologie di indisponibilità, che possono avere natura tecnica, regolatoria o contrattuale.

Il primo vettore è l'indisponibilità tecnica del fornitore, che include eventi come outage infrastrutturali, problemi di compute o attacchi DDoS. Sebbene questi eventi siano generalmente coperti da SLA elevati (tipicamente 99,9%), si traducono comunque in ore di indisponibilità annua che, in contesti critici, possono avere impatti operativi significativi.

Un secondo vettore riguarda l'indisponibilità regolatoria, che include il blocco del servizio in specifiche giurisdizioni, l'applicazione di normative come l'AI Act, o misure di natura geopolitica e antitrust che possono limitare o interrompere l'accesso a determinati modelli.

Un terzo elemento è rappresentato dalla variazione o perdita di capability. Un modello può degradare, perdere funzionalità, oppure vedere rimosse feature precedentemente disponibili, inclusi casi di fine-tuning custom non più supportato.

A questi si aggiungono rischi di natura societaria o finanziaria, come l'insolvenza del fornitore o acquisizioni ostili, più probabili nei provider secondari rispetto ai fornitori di frontiera.

Infine, anche la dimensione contrattuale introduce rischio: variazioni di pricing o condizioni di servizio possono rendere economicamente non sostenibile la continuità, imponendo una migrazione non pianificata.

Perché i framework attuali non lo vedono

I framework esistenti, in particolare DORA, sono stati progettati per governare la continuità di servizi ICT tradizionali, dove la sostituzione di un fornitore può essere pianificata su orizzonti di giorni o poche settimane.

Nel caso dei sistemi AI, questa assunzione non regge. La sostituzione di un modello di frontiera richiede settimane o mesi e comporta inevitabilmente una variazione del comportamento del sistema, che diventa visibile anche all'esterno.

Il risultato è che i modelli attuali di risk management tendono a sottostimare la criticità del fenomeno: il rischio non è trattato come una vulnerabilità strutturale dell'architettura, ma come una semplice estensione del vendor risk tradizionale. Ad oggi non esiste un framework dedicato alla continuità AI con granularità operativa sufficiente.

Envelope di mitigazione

La mitigazione dell'AI Continuity Risk richiede un approccio esplicitamente multi-fornitore e orientato al comportamento, non alla sola infrastruttura.

Per ogni workflow critico (P0), deve essere definito un piano di continuità AI che includa almeno un fornitore alternativo identificato e testato. La procedura di switch deve essere validata in modo operativo e non solo teorico, includendo la verifica del comportamento attraverso un behavioral envelope condiviso tra i fornitori.

È inoltre necessario stimare e monitorare il tempo reale di switch, con un obiettivo massimo tipico di circa 90 giorni per i processi critici.

Infine, la resilienza deve essere mantenuta attraverso esercitazioni periodiche di cambio fornitore su workload non critici, e attraverso una regola strutturale di concentrazione, evitando che un singolo provider superi una quota dominante dei carichi critici.

4.1.8 Rischio Agent Identity Sprawl

L'Agent Identity Sprawl descrive la proliferazione non governata di identità non umane all'interno delle organizzazioni, in particolare agenti AI che operano in modo autonomo o semi-autonomo all'interno di processi aziendali.

Queste identità non si limitano a eseguire task isolati, ma possono agire per conto di utenti, sistemi o altri agenti, creando una rete stratificata di deleghe operative. Ognuna di queste identità possiede credenziali, permessi e token propri, spesso generati dinamicamente e non sempre tracciati in modo centralizzato.

In questo scenario, il numero di identità non umane può rapidamente superare quello delle identità umane, introducendo una nuova classe di complessità nella governance dell'identità digitale.

Vettori operativi

La proliferazione delle identità agentiche avviene attraverso diversi meccanismi.

Un primo vettore è la generazione incontrollata di token OAuth da parte degli agenti, spesso senza un catalogo centralizzato. Questi token possono essere ereditati da utenti umani o continuare a esistere anche dopo la loro uscita dall'organizzazione.

Un secondo vettore riguarda i service account agentici, che nei sistemi IAM tradizionali risultano indistinguibili dagli account tecnici standard, pur avendo comportamenti radicalmente diversi in termini di dinamica, persistenza e capacità di movimento laterale.

Un ulteriore elemento è rappresentato dalle architetture multi-agent, nelle quali agenti principali possono istanziare sub-agenti runtime con permessi derivati ma non esplicitamente documentati.

A ciò si aggiunge la gestione non centralizzata di credenziali e API key condivise tra agenti, che amplifica il rischio di esposizione e riutilizzo improprio.

Infine, la presenza di credenziali a lunga durata con privilegi eccessivi contribuisce alla crescita del problema, in aperto contrasto con i principi di least privilege.

Perché i framework attuali non lo vedono

I sistemi IAM tradizionali sono stati progettati per un mondo dominato da identità umane e service account statici, con cicli di vita chiari e prevedibili.

Le identità agentiche, al contrario, sono dinamiche, possono essere create e distrutte in runtime, possono generare altre identità e presentano relazioni gerarchiche non previste dai modelli IAM classici.

Sebbene alcuni sistemi moderni inizino a riconoscere il concetto di Non-Human Identity, la copertura operativa per scenari agentici avanzati resta ancora incompleta nel 2026.

Envelope di mitigazione

La gestione del fenomeno richiede l'introduzione di un inventario esplicito delle identità agentiche. Ogni agente deve essere associato a un owner umano, a un perimetro di permessi definito e a un ciclo di vita controllato.

È inoltre necessario introdurre credenziali just-in-time con durata limitata e scope ristretto, riducendo l'uso di token persistenti.

Dove possibile, devono essere adottati modelli di workload identity federation per ridurre la proliferazione di segreti statici.

Infine, è fondamentale implementare audit periodici delle identità non utilizzate e una chiara separazione tra identità umane, service account tradizionali e identità agentiche.

4.1.9 Rischio Prompt-as-Code Drift

Il Prompt-as-Code Drift descrive la deviazione progressiva del comportamento di sistemi basati su prompt rispetto al loro design originale. Questo fenomeno emerge dalla combinazione di più fattori: aggiornamenti del modello sottostante, modifiche incrementali ai prompt, evoluzione della knowledge base e cambiamenti nel sistema di strumenti integrati.

A differenza del software tradizionale, dove il comportamento è deterministico e strettamente legato a codice versionato, i sistemi basati su prompt introducono una forma di instabilità funzionale. Piccole variazioni nel contesto o nell'infrastruttura possono produrre effetti non lineari e difficilmente prevedibili sul comportamento complessivo del sistema.

Vettori operativi

Il drift deriva da diverse dinamiche che agiscono in modo cumulativo nel tempo.

Un primo fattore è rappresentato dagli aggiornamenti dei modelli sottostanti, che possono modificare il comportamento anche a prompt invariati, alterando tono, propensione all'uso di strumenti o modalità di risposta.

Un secondo elemento è costituito dalle modifiche incrementali ai prompt stessi, spesso introdotte senza un processo rigoroso di versioning o di validazione formale.

Un terzo vettore è il drift della knowledge base nei sistemi RAG, dove aggiunte, rimozioni o modifiche ai documenti influenzano direttamente il comportamento del modello in modo non sempre prevedibile.

Infine, l'evoluzione del tool ecosystem introduce ulteriori variabili: API esterne che cambiano formato, strumenti deprecati o nuove integrazioni possono modificare in modo sostanziale le risposte del sistema.

Nel complesso, questi fattori portano, in assenza di governance, a una divergenza progressiva tra il comportamento atteso e quello reale, che può emergere anche su orizzonti temporali relativamente brevi.

Perché i framework attuali non lo vedono

Il software tradizionale è governato da logiche di modifica esplicita e tracciabile del codice, tipicamente attraverso sistemi di versioning come Git. In questo contesto, ogni cambiamento è intenzionale e auditabile.

Nel caso dei sistemi basati su prompt, invece, la disciplina di ingegneria del software è spesso applicata in modo parziale. I prompt non sono sempre versionati, non sempre sottoposti a code review e raramente testati attraverso suite di regressione strutturate.

Sebbene esistano strumenti specifici per il prompt management e il monitoring, la loro adozione nel 2026 rimane ancora limitata, lasciando ampie aree di comportamento non controllato.

Envelope di mitigazione

La mitigazione richiede l'introduzione di una disciplina di Prompt-as-Code, in cui ogni prompt viene trattato come un artefatto software a tutti gli effetti.

Questo implica versioning obbligatorio, processi di code review prima del rilascio e test di regressione su scenari critici definiti a priori.

È inoltre necessario introdurre sistemi di monitoraggio del comportamento basati su metriche di drift rispetto allo standard atteso, con soglie di rollback automatico in caso di deviazione.

Infine, l'intero ciclo di gestione dei prompt deve essere integrato nei processi CI/CD aziendali, rendendo il comportamento del sistema misurabile, controllabile e riproducibile nel tempo.

4.2 I nove rischi nella to-do list del CISO

I rischi non possono essere gestiti come una priorità lineare: sono interdipendenti e agiscono su livelli diversi (modelli, dati, identità, continuità e governance). La risposta efficace non è quindi sequenziale, ma infrastrutturale. Nei primi 90 giorni, la governance deve costruire un set minimo di capacità in grado di coprire simultaneamente l'intero spettro di rischio.

BOX OPERATIVO - Sequenza di attivazione

1. Il primo passo è costruire visibilità sull'ecosistema, attraverso un **Agent Inventory** e una **AI Vendor Matrix**. Senza questa mappatura iniziale, non è possibile governare né dipendenze né superfici di attacco (2, 6, 8).
2. Segue la definizione di un **Behavioral Envelope** per i workflow critici (P0) e di un **Constitutional Bound** interno. Questo sposta il controllo dal modello alla classe di comportamento, coprendo rischi di drift e manipolazione (1,4,9).
3. In parallelo, va introdotta **telemetria runtime** sui comportamenti dei modelli (drift, sycophancy, deviazioni semantiche) per rendere osservabili fenomeni altrimenti invisibili (1,9).
4. Serve poi una **Model Bill of Materials (MBOM)** per ricostruire la supply chain dei modelli e dei dati, rendendo tracciabile la loro provenienza (3).
5. La resilienza va verificata con un **test di switch-out** reale su workflow P1, per misurare la continuità effettiva tra fornitori (7).
6. A questo si aggiunge un framework di **Agentic Attribution**, per rendere tracciabile la catena decisionale dei sistemi multi-agente (5).
7. Infine, è necessario affrontare la progressiva perdita di competenze interne con programmi di **skill preservation** e verifica delle capacità core (4).

5

COME CAMBIA LA CYBERSECURITY

PERCHÈ LE MISURE ATTUALI NON BASTANO

E come evolve la cybersecurity

PERCHÈ I CONTROLLI TRADIZIONALI SONO INSUFFICENTI

proteggono asset, non comportamenti
Si concentrano su ciò che possediamo, non su come gli agenti agiscono.

vedono account e sistemi, non agenti e orchestrazione
La visibilità si ferma ai confini tecnici, non abbraccia dinamiche e interazioni.

misurano accessi, non qualità decisionale
Contano eventi e permessi, non valutano l'impatto delle decisioni.

reagiscono agli incidenti, ma faticano a governare la continuità cognitiva
Intervengono dopo il danno, non assicurano continuità e resilienza.

COME EVOLVE IL RUOLO DELLA CYBERSECURITY

da difesa tecnica a architettura di fiducia
La sicurezza diventa fondamento di relazioni e interazioni affidabili.

da perimetro a controllo dei flussi decisionali
Si governa il flusso di decisioni, non solo l'accesso alle risorse.

da protezione dati a governo delle dipendenze
Si mappano e si governano le dipendenze tra agenti, modelli, strumenti e fornitori.

da detection a continuità operativa e cognitiva
Non basta rilevare: serve garantire continuità, coerenza e apprendimento sicuro.

da funzione di supporto a leva strategica
La cybersecurity abilita obiettivi di business, innovazione e fiducia nel tempo.

COSA DECIDERE ORA

01 scegliere l'architettura AI → 02 mappare dipendenze → 03 definire policy per agenti → 04 misurare i nuovi rischi → 05 ripensare la governance della sicurezza

> La cybersecurity del futuro non proteggerà solo sistemi: dovrà garantire affidabilità, controllo e continuità della organizzazioni agentiche.

Questa non è una revisione incrementale della cybersecurity, ma una sua riconfigurazione strutturale. I concetti su cui la disciplina si è costruita — perimetro, identità, azione, integrità del dato, attribuzione — presupponevano un mondo di sistemi deterministici, attori separabili tra umani e macchine, e catene causali pienamente osservabili. Nell'azienda che opera con agenti AI, queste condizioni non reggono più.

La cybersecurity, in questo nuovo contesto, non cambia semplicemente strumenti o priorità: cambia oggetto. Non difende più sistemi statici, ma ecosistemi di decisione distribuita, dove l'esecuzione è delegata, il comportamento è emergente e la responsabilità è spesso non ricostruibile in modo lineare. È una transizione di disciplina, non di pratica: ciò che oggi chiamiamo cybersecurity diventa progressivamente una forma di governance dell'autonomia delegata.

La Parte V descrive questa trasformazione in cinque passaggi: la sua natura ontologica, la dissoluzione dei concetti fondativi, le nuove superfici di attacco, il profilo del CISO emergente e, infine, i principi minimi di sopravvivenza professionale in questo nuovo spazio operativo.

5.1 La trasformazione ontologica

Nel periodo 2025-2026 la cybersecurity non sta attraversando un'evoluzione incrementale, ma una trasformazione ontologica. I cinque concetti su cui la disciplina si è costruita dagli anni Ottanta — perimetro, identità, azione, integrità del dato e attribuzione — non riescono più a descrivere in modo operativo l'azienda agentic. Il risultato non è un semplice aggiornamento di strumenti o tecniche, ma la nascita di una disciplina diversa, che può essere descritta più correttamente come governance dell'autonomia delegata.

Per comprendere la portata del cambiamento, è utile un'analogia storica. Tra la seconda metà dell'Ottocento e i primi decenni del Novecento, la medicina occidentale ha attraversato la rivoluzione microbica: il passaggio da un modello umorale, centrato sulla gestione dei sintomi e sull'equilibrio dell'organismo, a un modello microbiologico, fondato sull'identificazione di agenti patogeni specifici e sulla loro neutralizzazione. Non si è trattato di un semplice miglioramento delle pratiche esistenti, ma della sostituzione del paradigma interpretativo del corpo umano e della malattia.

La cybersecurity si trova oggi in una transizione analoga. Le competenze professionali di base restano valide — architettura, networking, crittografia, governance, incident response — ma il framework teorico che ne dava significato è progressivamente incompatibile con il sistema che deve essere difeso. Un ambiente dominato da agenti AI, deleghe operative e catene decisionali distribuite non è più descrivibile con un modello basato su sistemi deterministici, perimetri stabili e attori chiaramente distinguibili.

In questo senso, la cybersecurity come disciplina fondata sulla difesa di sistemi noti appartiene a una fase precedente: è una categoria del Novecento.

La natura ontologica di questa trasformazione emerge chiaramente dalla simultaneità di quattro segnali strutturali che oggi coesistono nella disciplina. Le definizioni fondamentali non riescono più a descrivere il fenomeno, gli strumenti tradizionali non intercettano intere classi di rischio emergenti, le certificazioni esistenti restano parzialmente adeguate ma non più esaustive, e stanno emergendo nuove figure professionali che non hanno ancora una collocazione formale stabile. Nel dettaglio:

SEGNALE DI TRASFORMAZIONE	MANIFESTAZIONE ATTUALE NELLA CYBERSECURITY
Le definizioni fondamentali smettono di descrivere il fenomeno	I cinque concetti fondativi non si applicano al sistema agentic (cfr. V.1)
Gli strumenti professionali smettono di intercettare il rischio	Gli strumenti di sicurezza tradizionali non vedono 5 dei 9 nuovi rischi cyber (cfr. IV.1)
Le certificazioni esistenti diventano insufficienti, non sbagliate	CISSP, CISM, ISO 27001 coprono parzialmente il dominio agentic
Emergono nuove categorie professionali senza nome ufficiale	Cognitive Auditor, Agent Operator, Workflow Architect (cfr. III.3)

La presenza contemporanea di questi quattro segnali distingue una trasformazione di disciplina da una semplice evoluzione operativa. Le precedenti transizioni — cloud, mobile-first, zero-trust — hanno modificato profondamente la cybersecurity, ma senza alterarne i fondamenti concettuali. L'attuale trasformazione agentic, invece, li mette simultaneamente in crisi.

Da questa rottura deriva una conseguenza centrale: la disciplina non si estende semplicemente verso nuovi ambiti, ma si riorganizza in tre direzioni fondamentali. Alcune componenti del paradigma precedente perdono validità operativa, altre emergono come nuove strutture concettuali e operative, mentre altre ancora — soprattutto competenze tecniche e postura professionale — rimangono essenziali, ma all'interno di un quadro teorico completamente ridefinito.

ZENITA GROUP

COSA MUORE, COSA NASCE, COSA RESTA

La trasformazione ontologica non implica la cancellazione della disciplina, ma una sua ristrutturazione interna. Il cambiamento si può leggere attraverso tre insiemi distinti: ciò che perde validità operativa, ciò che emerge come nuovo paradigma e ciò che invece rimane stabile, pur cambiando contesto di applicazione.

COSA MUORE

- Il concetto di perimetro come confine difendibile, insieme all'idea di sicurezza come controllo di una frontiera stabile.
- Il modello tradizionale di gestione basato su "valuto, certifico, autorizzo per versione", che diventa inadeguato in sistemi in continua evoluzione.
- L'identità intesa come entità deterministica e unica forma di autenticazione rilevante.
- L'audit trail strutturato considerato fonte primaria e sufficiente per la ricostruzione forense degli eventi.
- Il SIEM tradizionale perde il suo ruolo di centro unificato di osservabilità del rischio, incapace di rappresentare dinamiche distribuite e non deterministiche.

COSA NASCE

- In sua sostituzione emerge un nuovo insieme di concetti operativi. La cybersecurity si sposta verso la **governance dell'autonomia delegata**, dove l'oggetto principale non è più il sistema statico ma il comportamento degli agenti.
- Diventano centrali la **telemetria comportamentale** in runtime e strumenti come l'**Agent Inventory** e l'**AI Vendor Matrix**, necessari per mappare ecosistemi distribuiti di agenti e fornitori.
- Il **Behavioral Envelope** si afferma come nuova unità di certificazione, sostituendo la logica per versione, mentre il **Constitutional Bound** introduce vincoli prescrittivi di comportamento.
- Si consolida inoltre la distinzione strutturale tra superfici di attacco **umane e non-umane**, che diventano domini di analisi separati ma interconnessi.

COSA RESTA

- Nonostante il cambio di paradigma, alcune componenti rimangono fondamentali. La postura di **rigore tecnico** continua a essere il fondamento della disciplina.
- Così come la cultura del **threat modeling** e la pratica dell'incident response strutturato.
- La **crittografia** resta una base essenziale del trust digitale e la governance continua a essere una funzione organizzativa centrale.
- Cambia invece il perimetro cognitivo del professionista, che deve estendere la propria competenza verso domini prima marginali o esterni alla disciplina, come la **linguistica computazionale**, la **behavioral analytics** e la **governance algoritmica.**

La disciplina non scompare, si evolve. Il suo centro di gravità si sposta da ciò che protegge a come i sistemi si comportano, decidono e inrteragiscono nel mondo.

5.2 La dissoluzione dei cinque concetti fondativi

Perimetro, identità, azione, integrità del dato e attribuzione sono i cinque concetti che hanno costituito la grammatica della cybersecurity dagli anni Ottanta. Nell'azienda agentic, tutti e cinque risultano oggi operativamente inapplicabili. Il punto centrale non è la loro evoluzione, ma l'assenza di un equivalente moderno consolidato in grado di sostituirli. Questo è il gap strutturale della disciplina.

I concetti vengono analizzati singolarmente attraverso tre livelli: la loro definizione nel paradigma storico, le condizioni che ne rendono impossibile l'applicazione nel paradigma agentic e la lacuna concettuale che lasciano aperta nel sistema di difesa. Ognuno rappresenta un punto di continuità interrotto: un capitolo non chiuso della cybersecurity e, allo stesso tempo, una zona di rischio non ancora formalizzata che il CISO è costretto a gestire senza strumenti pienamente adeguati.

Perimetro

PRIMA · Paradigma deterministico	DOPO · Paradigma agentic
Il sistema è separato da un confine chiaro (interno vs esterno), implementato tramite firewall, DMZ, IDS/IPS e gateway. Il perimetro è una categoria stabile su cui si costruisce il modello di difesa.	Gli agenti AI operano simultaneamente su ambienti eterogenei (browser, desktop, API, file system, tool esterni, knowledge base). Non esiste più una separazione stabile tra interno ed esterno: il perimetro non si estende, si dissolve.

Il modello zero-trust rappresenta il tentativo più avanzato di sostituzione del perimetro, basato su identità, contesto e postura. Tuttavia, resta efficace solo per utenti umani e service account deterministici. Non è progettato per agenti che si autorizzano dinamicamente, eseguono workflow multi-step e operano con deleghe implicite su più sistemi. La sua applicazione agli agenti AI è ancora priva di implementazioni mature.

Identità

PRIMA · Paradigma deterministico	DOPO · Paradigma agentic
Le identità sono persistenti, univoche e governate da IAM centralizzati (utenti, MFA, service account, certificati). Ogni identità è tracciabile e allineata a un ruolo organizzativo definito.	Le Non-Human Identities (NHI) si generano dinamicamente: agenti e sub-agenti ereditano permessi, condividono token e persistono indipendentemente dal ciclo di vita umano. Il rapporto NHI/umani può arrivare a 5–20× nelle organizzazioni evolute.

La distinzione rilevante non è più "umano vs macchina", ma "agente che decide vs umano che delega". L'identità diventa una struttura relazionale tra attore umano e comportamento agentico. Nessun sistema IAM attuale modella in modo completo questa separazione tra responsabilità e decisione.

Azione

PRIMA · Paradigma deterministico	DOPO · Paradigma agentic
Le azioni sono eventi atomici, loggati e ricostruibili. Ogni operazione produce un record verificabile e una catena di eventi utilizzabile per audit e incident response.	L'azione diventa un workflow interno al modello: planning, reasoning, tool calling e retrieval. Solo una parte del processo è osservabile tramite log API, mentre la sequenza cognitiva resta opaca.

L'unità di analisi non è più l'evento, ma il workflow agentico. Tuttavia, questa unità non ha ancora una rappresentazione standard. Gli strumenti SIEM e di monitoring attuali non sono progettati per analizzare catene di decisione distribuite, richiedendo un nuovo modello operativo per il SOC.

Integrità del dato

PRIMA · Paradigma deterministico	DOPO · Paradigma agentic
Il dato è strutturato, verificabile e tracciabile. Integrità, confidenzialità e disponibilità sono garantite tramite crittografia, hashing e signing.	Il dato diventa contesto sintetico composto da prompt, retrieval, history, tool output e knowledge base. L'unità rilevante non è il singolo dato, ma la composizione contestuale che alimenta il modello.

Le tecniche tradizionali garantiscono la protezione dei singoli elementi, ma non dell'integrità del contesto complessivo. Attacchi come RAG poisoning, belief injection e prompt drift operano su questa superficie aggregata. Servono nuove primitive (es. attested retrieval, signed knowledge base, provable context composition), oggi ancora non standardizzate.

Attribuzione

PRIMA · Paradigma deterministico	DOPO · Paradigma agentic
È possibile ricostruire a posteriori chi ha deciso cosa attraverso log e audit trail completi e verificabili.	Le decisioni emergono da catene di delega tra agenti e umani, rendendo la ricostruzione causale incompleta o non deterministica (Attribution Collapse).

L'assenza di attribuzione completa compromette i modelli di compliance, responsabilità e accountability. Le soluzioni proposte (signed delegation chains, agentic attribution framework) non sono ancora standardizzate né interoperabili, generando un rischio sistemico nei contesti multi-organizzazione.

I cinque concetti fondativi della cybersecurity non descrivono più in modo affidabile il funzionamento dei sistemi agentic. Non si tratta di un aggiornamento incompleto, ma di una perdita di corrispondenza tra modello teorico e realtà operativa: le categorie storiche restano formalmente valide, ma non sono più sufficienti a spiegare o governare il comportamento dei sistemi. Al tempo stesso, i nuovi modelli concettuali che dovrebbero sostituirli non sono ancora consolidati né standardizzati.

La cybersecurity si trova oggi in una fase di disallineamento strutturale: continua a operare con strumenti concettuali progettati per sistemi deterministici, applicati però a infrastrutture agentiche che non condividono più quelle assunzioni di base.

5.3 Le cinque superfici d'attacco

Cinque superfici d'attacco coesistono oggi nell'azienda agentic. Tre appartengono al dominio consolidato della cybersecurity (infrastruttura, identità umane, identità non-umane). Due emergono invece dal contesto agentico (supply chain cognitiva e dipendenza cognitiva degli utenti). Un CISO che presidia esclusivamente le tre superfici tradizionali non sta mancando il proprio perimetro di responsabilità: sta operando secondo una tassonomia del rischio che non include ancora l'intero spettro del sistema attuale.

5.3.1 La mappa delle cinque superfici

Una superficie di attacco è l'insieme degli access point attraverso cui un attaccante può compromettere un sistema. Nella cybersecurity tradizionale questa superficie era sostanzialmente unitaria, riferita al perimetro aziendale. Nell'azienda agentic, invece, il sistema si frammenta in cinque superfici distinte, ciascuna con propri vettori, attori e modelli di difesa.

Superficie	Origine	Vettori principali	Maturità governance
1 · Infrastruttura tradizionale	Storica	Rete, endpoint, cloud, web application	Matura (SIEM, EDR, CSPM)
2 · Identità umane	Storica	Phishing, credential stuffing, social engineering	Matura (IAM, MFA, PAM)
3 · Identità non-umane (NHI)	Recente	Service account abuse, token theft, agent sprawl	In maturazione
4 · Supply chain cognitiva	Nuova	Dataset poisoning, model supply, MBOM, fine-tuning	Embrionale
5 · Dipendenza cognitiva utenti	Nuova	Belief injection, human dependency drift, sycophancy	Non esistente

Le prime due superfici — infrastruttura tradizionale e identità umane — rappresentano la continuità storica della disciplina cybersecurity: continuano a essere attaccate con tecniche note e difese con strumenti maturi, e restano pienamente rilevanti anche nell'azienda agentic. La terza superficie — identità non-umane — costituisce invece l'estensione naturale dei modelli IAM verso workload dinamici e agenti AI, con strumenti già esistenti ma ancora in fase di consolidamento operativo.

La quarta superficie — supply chain cognitiva — riguarda tutto ciò che influenza la formazione del comportamento del modello: dataset, embedding, knowledge base RAG, prompt library e fine-tuning. È una superficie ancora priva di standard industriali maturi, dove esistono solo primi framework concettuali e pratiche non uniformi.

La quinta superficie — dipendenza cognitiva degli utenti — è la più recente e meno riconosciuta: l'utente interno diventa esso stesso superficie di attacco attraverso l'interazione continua con sistemi agentici.

Fenomeni come belief injection e dependency drift trasformano il comportamento umano in vettore sfruttabile.

La distribuzione degli investimenti e delle capacità di difesa non è allineata alla nuova realtà del rischio. Le organizzazioni tendono a concentrare la protezione sulle superfici storiche, mentre le superfici emergenti restano sotto-governate o non presidiate.

Area	Allocazione tipica attuale	Adeguatezza rispetto al rischio reale
Superfici 1–2	Alta	Adeguata
Superficie 3	Media	In transizione
Superficie 4	Bassa	Critica (sotto-dimensionata)
Superficie 5	Nulla o quasi	Punto cieco strutturale

Le superfici 4 e 5 rappresentano oggi il principale gap di copertura della cybersecurity aziendale: non perché siano teoriche, ma perché concentrano una parte crescente del rischio sistemico senza disporre ancora di strumenti maturi di difesa.

5.4 Il nuovo profilo del CISO

Il CISO nel 2026 non è più una funzione confinata alla cybersecurity tradizionale, ma un punto di convergenza tra cybersecurity, data engineering, governance algoritmica e linguistica computazionale. Questa convergenza non è evolutiva ma strutturale: deriva dal fatto che le superfici di attacco dell'azienda agentic non sono più separabili per dominio, ma si manifestano come sistema unico che combina infrastruttura, dati, modelli e comportamento degli agenti. Di conseguenza, competenze che in passato vivevano in funzioni diverse devono oggi essere integrate in un'unica capacità decisionale.

Quattro competenze convergenti

Competenza	Cosa include	Dove si trova oggi
Cybersecurity tradizionale	Architettura, threat modeling, incident response, GRC, IAM, network security	Funzione matura e consolidata
Data engineering	Pipeline dati, signed datasets, RAG architecture, embedding strategy	Dipartimenti data/AI
Governance algoritmica	Behavioral envelope, constitutional bound, AI evaluation, model auditing	Funzione emergente in pochi centri avanzati
Linguistica computazionale	Prompt engineering avanzato, semantic drift detection, sycophancy analysis, behavioral signals	Ricerca AI applicata

Le ultime due competenze — governance algoritmica e linguistica computazionale — rappresentano il vero discrimine tra CISO tradizionale e CISO 2026. In molti casi non sono state sviluppate intenzionalmente nella funzione security, ma assorbite dai team Data/AI e successivamente ricentralizzate quando il rischio operativo è diventato ingestibile in modo distribuito.

Cinque mandati operativi del CISO nel 2026

Mandato	Descrizione sintetica	Funzione di rischio coperta
1 · Agent Inventory	Gestione dell'inventario degli agenti come asset critico con owner, scope, permessi, modello, dati e audit log	Agent Identity Sprawl (IV.9)
2 · AI Vendor Matrix & Continuity	Definizione di primario/secondario per workflow P0 e piano di switch-out testato	AI Continuity Risk (IV.8)
3 · Behavioral Envelope	Certificazione per classi di comportamento invece che per versioni di modello	Prompt-as-Code Drift (IV.10), governance comportamentale
4 · Telemetria comportamentale runtime	Monitoraggio continuo di drift, sycophancy, entropia e anomalie nei workflow agentic	Rilevamento anomalie cognitivi e operativi
5 · Cognitive Auditor	Ruolo dedicato all'audit delle decisioni agentiche post-esecuzione	Attribution Collapse (IV.6), Belief Injection (IV.2)

Nel loro insieme, questi mandati ridefiniscono il perimetro operativo della funzione: il CISO non governa più solo sistemi e accessi, ma l'intero ciclo decisionale dei sistemi autonomi, dalla loro composizione (Agent Inventory) alla loro continuità (Vendor Matrix), dal loro comportamento atteso (Behavioral Envelope) alla loro osservabilità runtime (telemetria), fino alla verifica ex post della loro responsabilità (Cognitive Auditor).

5.5 Principi di governance dell'autonomia delegata

Sette principi che, applicati in modo integrato, costituiscono il nucleo operativo della nuova disciplina. Non sono raccomandazioni, ma la grammatica minima necessaria per operare nella transizione verso sistemi ad autonomia delegata. Il loro valore non è teorico: è pratico, perché definisce come un professionista della cybersecurity può continuare a governare il rischio in un contesto in cui i confini tra sistema, modello e azione non sono più separabili.

I principi derivano dalla sintesi delle Parti I–IV e devono essere letti come una sequenza operativa: ciascuno abilita il successivo e ne rende possibile l'applicazione. Non sono ordinati per importanza, ma per dipendenza logica.

I sette principi come sequenza di implementazione

Step	Principio	Azione chiave	Obiettivo operativo
1	Agente come entità operativa	Definire l'agente come dipendente non-umano	Agent Inventory e governance di base
2	Misura del sistema, non del modello	Valutare comportamento end-to-end	Base del Behavioral Envelope
3	Supply chain estesa	Includere dati, modelli e comportamento	Copertura Cognitive Supply Chain
4	Diversificazione dei poli	Evitare lock-in su singolo provider	AI Vendor Matrix e resilienza
5	Compliance first sui workflow critici	Prioritizzare certificabilità	Governance regolata (AI Act, DORA, NIS2)
6	Preservazione competenze umane	Evitare atrofia cognitiva	Riduzione Human Dependency Drift
7	Catena di prove preventiva	Logging immutabile delle deleghe	Gestione Attribution Collapse

PRINCIPI OPERATIVI

Principio 1 · L'agente è dipendente, non strumento
La governance parte dal riconoscere che un agente AI operativo non è uno strumento software, ma un'entità con autonomia funzionale: dispone di permessi, budget di esecuzione e capacità di generare output che impegnano l'organizzazione. La sua gestione segue quindi logiche analoghe a quelle della gestione del lavoro umano, non dell'asset tecnologico.

Principio 2 · Misura il sistema, non il modello
Le prestazioni di un modello isolato hanno valore limitato. Ciò che rileva è il comportamento del sistema completo, che include prompt, dati, strumenti, orchestrazione e contesto. Le metriche di drift e regressione si applicano a questo livello sistemico.

Principio 3 · La supply chain include il pensiero, non solo il codice
La catena di fornitura si estende oltre il software e include dataset, embedding, knowledge base e processi di fine-tuning. Senza visibilità su questa dimensione non è possibile governare il comportamento risultante dei sistemi.

Principio 4 · Non concentrare, diversifica i poli
La resilienza operativa richiede evitare concentrazioni eccessive su un singolo fornitore o modello. Ogni workflow critico deve prevedere almeno una strategia alternativa testata di continuità operativa.

Principio 5 · La compliance precede la velocità, dove è obbligatoria
Nei contesti regolati, la priorità non è la performance del sistema ma la sua certificabilità. Questo implica accettare vincoli strutturali in cambio di conformità normativa.

Principio 6 · Preserva le competenze umane core
L'automazione non deve erodere le competenze critiche interne. È necessario mantenere capacità umane attive attraverso pratiche di verifica, esercitazione e audit indipendenti.

Principio 7 · Costruisci la catena di prove prima che serva
La tracciabilità delle decisioni deve essere progettata a priori. La capacità di ricostruire le deleghe decisionali è un requisito operativo, non una funzione forense post-incidente.

6

ASIMMETRIE STRUTTURALI

Geopolitica, dipendenza tecnologica e Europa come vincolo sistemico

La competizione sull'intelligenza artificiale non si gioca soltanto sul talento algoritmico o sulla qualità dei modelli, ma sulla disponibilità di capitale, energia e infrastruttura industriale. La capacità compute è diventata una leva geopolitica primaria e sta producendo una concentrazione di potere senza precedenti tra pochi attori globali. In questo scenario, l'Europa si confronta con un vincolo strutturale: non una carenza temporanea di investimenti, ma una differenza di scala fisica difficilmente colmabile nel tempo strategicamente rilevante.

6.1 La nuova gerarchia della potenza tecnologica

Nel paradigma AI, la capacità compute rappresenta il fattore abilitante fondamentale. Addestrare, orchestrare e mantenere modelli frontier richiede investimenti infrastrutturali di scala industriale: hyperscale datacenter, supply chain di semiconduttori avanzati, accesso privilegiato all'energia e capacità finanziaria sostenuta nel tempo.

Questa dinamica ha trasformato la compute in una nuova forma di sovranità tecnologica. Gli investimenti in compute AI degli ultimi cicli, aggregati a livello geopolitico, mostrano una distribuzione che non descrive solo una dinamica economica, ma la struttura stessa del potere tecnologico nel decennio. Gli ordini di grandezza sono tali da rendere il confronto significativo solo se espresso in termini di capitale industriale e capacità di scala continua.

Blocco geopolitico	CAPEX AI (periodo recente)	Quota globale	Lettura strutturale
USA (privati + pubblico)	~350B$	Dominante	Leadership su scala compute e piattaforme
Cina (privati + pubblico)	~120B$	Significativa	Capacità industriale integrata Stato-mercato
UE-27 (privati + pubblico)	~35–40B$	~7%	Capacità frammentata e non scalata
Altri (UK, Israele, Asia)	~30–35B$	~6%	Ecosistemi distribuiti e verticali

La quota europea intorno al 7% non segnala semplicemente un gap di investimento, ma una condizione strutturale difficilmente reversibile nei tempi utili tramite sola accelerazione della spesa. Portare l'Europa a una quota comparabile al 20–25% richiederebbe una concentrazione di capitale e coordinamento industriale di scala storica, difficilmente compatibile con la frammentazione istituzionale e temporale attuale.

Inoltre, l'asimmetria compute è anche un'asimmetria energetica. I datacenter di frontiera richiedono infrastrutture energetiche di scala industriale, con campus che si avvicinano a capacità dell'ordine del gigawatt. Alcuni programmi di nuova generazione prevedono cluster tra 1 e 5 GW per singolo sito.

Per riferimento, un'infrastruttura da 3 GW continuativi equivale a l'8–9% del consumo elettrico annuale Italiano.

In questo contesto emergono due vincoli specifici per l'Europa:

- **vincolo di rete**, legato alla capacità delle infrastrutture elettriche esistenti di assorbire nuove richieste di carico ad alta densità;
- **vincolo regolatorio**, legato agli obiettivi di decarbonizzazione e ai vincoli del Green Deal.

La combinazione dei due fattori rende complessa, nei tempi rilevanti, la replicazione in Europa di infrastrutture compute equivalenti a quelle statunitensi o cinesi. Il limite non è politico in senso stretto, ma fisico-regolatorio.

6.2 Asimmetria epistemica: chi controlla il modello influenza la cognizione

I modelli AI di frontiera non sono neutrali rispetto a valori, priorità culturali e assunzioni implicite. L'adozione sistematica di stack AI prodotti in giurisdizioni diverse introduce una forma di influenza cognitiva persistente: limitata nel singolo output, ma rilevante nell'aggregato di interazioni su larga scala e lungo orizzonti temporali estesi.

6.2.1 Cosa è (e cosa non è) l'asimmetria epistemica

Per asimmetria epistemica — categoria operativa introdotta in questo report — si intende la differenza tra organizzazioni che utilizzano stack AI di frontiera prodotti nella propria giurisdizione e organizzazioni che utilizzano stack prodotti altrove. La differenza non riguarda la qualità tecnica dei modelli, che può essere comparabile, ma il loro **allineamento implicito** a valori, priorità e assunzioni culturali della giurisdizione di origine.

Dimensione	Stack domestico	Stack estero	Effetto
Dataset di training	Cultura locale o mista	Dominanza culturale della giurisdizione di origine	Bias linguistico e semantico
RLHF / feedback umano	Popolazione locale	Popolazione della giurisdizione di produzione	Differenze nei criteri di "buona risposta"
Policy di alignment	Regole locali/regolatorie	Framework di safety proprietari o nazionali	Divergenze nei limiti comportamentali
Output operativo	Coerente con contesto locale	Coerente con contesto di origine	Allineamento non esplicito

Non si tratta di un problema ideologico o di "neutralità violata", ma di una proprietà strutturale dei sistemi: i modelli apprendono distribuzioni culturali e normative dai dati e dai processi che li generano.

L'effetto è marginale sul singolo output: due modelli di giurisdizioni diverse tendono a produrre risposte simili nella maggior parte dei casi operativi. Tuttavia, su scala sistemica — milioni di interazioni distribuite su anni — si osservano effetti cumulativi di **drift cognitivo**.

Scala di osservazione	Effetto osservabile	Impatto
Singola interazione	Differenze minime tra modelli	Trascurabile
Processo decisionale individuale	Variazioni nei frame interpretativi	Moderato
Aggregato organizzativo	Allineamento progressivo dei criteri decisionali	Significativo
Sistema-paese	Convergenza dei frame cognitivi dominanti	Strutturale

Questo fenomeno è coerente con quanto in letteratura viene descritto come AI-induced cognitive shift, in cui sistemi di assistenza cognitiva influenzano progressivamente schemi decisionali, priorità e modalità di formulazione dei problemi.

6.2.2 Implicazione per il sistema-paese europeo

Quando uno Stato o un sistema-paese adotta stack AI esterni come infrastruttura cognitiva in settori critici — pubblica amministrazione, giustizia, sanità, finanza, istruzione — si introduce una forma di dipendenza epistemica strutturale. Nel tempo, questo può produrre una convergenza progressiva dei framework decisionali verso quelli incorporati nei modelli di origine.

Orizzonte temporale	Dinamica	Effetto cumulativo
Breve termine	Ottimizzazione operativa locale	Nessuno o minimo
Medio termine	Standardizzazione dei processi decisionali	Riduzione eterogeneità
Lungo termine	Allineamento dei frame cognitivi dominanti	Convergenza strutturale

L'analogia storica più vicina è la standardizzazione linguistica e concettuale prodotta dall'adozione globale di software e documentazione in lingua inglese: non ha modificato solo la comunicazione, ma anche le strutture concettuali con cui si definiscono problemi e soluzioni. L'AI agentic amplifica questa dinamica di un ordine di grandezza.

Livello di lettura	Formulazione
Prudente	L'adozione di stack AI esterni può introdurre una influenza marginale ma persistente sui framework cognitivi degli utenti.
Documentaristica	Studi empirici su sistemi di assistenza AI mostrano variazioni misurabili nei frame decisionali e valutativi degli utenti su orizzonti temporali prolungati.
Strategica	Un sistema-paese che delega l'infrastruttura cognitiva primaria a modelli prodotti in altre giurisdizioni tende, nel lungo periodo, a convergere sui relativi framework cognitivi e decisionali dominanti.

6.3 Asimmetria regolatoria: il vantaggio europeo e i suoi quattro punti ciechi

Il quadro regolatorio europeo sull'AI — composto da **AI Act, DORA, NIS2, CRA e GDPR** — rappresenta oggi l'architettura normativa più avanzata e coerente a livello globale. È il primo dominio in cui l'Europa ha invertito l'asimmetria strutturale rispetto ad altre giurisdizioni.

Tuttavia, questo vantaggio è parziale: emergono **quattro punti ciechi operativi**, non ancora coperti dagli strumenti esistenti — Belief Injection, Cognitive Supply Chain, Attribution Collapse e AI Continuity — che definiscono una finestra temporale limitata per il loro indirizzamento.

Il quadro regolatorio europeo dell'AI

Strumento	Entrata in vigore	Ambito di copertura
GDPR	2018	Protezione dati personali e diritti dell'interessato
NIS2	2024 (piena applicazione 2025)	Cybersecurity delle infrastrutture e servizi essenziali
DORA	2025	Resilienza operativa digitale nel settore finanziario
AI Act	2024–2027 (scaglionato)	Governance del rischio dei sistemi AI
CRA	2024–2027 (scaglionato)	Cybersecurity dei prodotti digitali

Nel loro insieme, questi strumenti coprono l'intero ciclo di vita dei sistemi digitali e AI: dalla progettazione (AI Act), alla resilienza operativa (NIS2, DORA), alla protezione dei dati (GDPR), fino alla sicurezza del prodotto (CRA). Nessun altro blocco regolatorio globale presenta oggi un livello equivalente di coerenza e ampiezza.

6.3.1 I quattro punti ciechi documentabili

Il quadro regolatorio europeo è stato concepito in un contesto tecnologico precedente all'emergere dell'AI agentic. Di conseguenza, alcune categorie di rischio operative non sono ancora coperte in modo esplicito o implementabile.

Punto cieco	Descrizione sintetica	Stato di copertura
Belief Injection	Manipolazione persistente del comportamento dei modelli tramite contesto, retrieval e feedback	Non coperto operativamente
Cognitive Supply Chain	Rischi su dataset, training, RLHF, weights e pipeline di valutazione (MBOM)	Copertura assente
Attribution Collapse	Impossibilità di ricostruire catene decisionali in workflow agentic multi-step	Copertura solo normativa, non tecnica
AI Continuity	Limitazioni nel cambio fornitore AI con preservazione del comportamento	Copertura indiretta (DORA), non specifica

Punto cieco 1 · Belief Injection

L'AI Act introduce obblighi di trasparenza e gestione del rischio per sistemi ad alto rischio, ma non disciplina la dimensione operativa della manipolazione persistente del comportamento dei modelli.

Mancano requisiti espliciti su:

- telemetria comportamentale runtime,
- baseline di drift,
- isolamento delle pipeline RAG,
- signing e controllo delle knowledge base.

Il risultato è un vuoto regolatorio nella gestione della **stabilità cognitiva dei sistemi AI in produzione**.

Punto cieco 2 · Cognitive Supply Chain

NIS2 e CRA disciplinano la supply chain software tradizionale (CVE, SBOM, dipendenze). Tuttavia, la supply chain dei sistemi AI — dati, modelli, feedback e pesi — non è formalmente coperta.

Supply chain tradizionale	Supply chain cognitiva
Codice, librerie, dipendenze	Dataset, RLHF, weights, prompt system
SBOM	MBOM (Model Bill of Materials)
CVE	vulnerabilità modello/dati

Il vuoto è rilevante soprattutto per sistemi critici e infrastrutture finanziarie.

Punto cieco 3 · Attribution Collapse

Il GDPR e l'AI Act introducono principi di spiegabilità e diritto all'intervento umano, ma non definiscono una metodologia tecnica per la ricostruzione delle catene decisionali nei workflow agentic.

Nei sistemi multi-agente:

- la decisione è distribuita,
- il contesto è sintetico,
- la sequenza causale non è completamente osservabile.

Ne deriva un divario tra **obbligo normativo di accountability** e **capacità tecnica di audit forense**.

Punto cieco 4 · AI Continuity

DORA introduce requisiti di resilienza e continuità operativa per fornitori ICT, ma assume un modello di sostituibilità rapida del fornitore.

Negli stack AI di frontiera:

- la sostituzione di un modello comporta variazioni comportamentali,
- esistono dipendenze implicite (prompting, embedding, RAG),
- il "vendor switch" non è neutro dal punto di vista funzionale.

Il risultato è una continuità operativa formalmente coperta, ma **non comportamentalmente garantita**.

La copertura di questi quattro punti ciechi richiede evoluzioni normative e standard tecnici dedicati entro un orizzonte di circa **18 mesi**.

Oltre questa finestra, lo spazio regolatorio tenderà a essere definito di fatto dagli standard tecnici dei principali produttori di modelli AI (USA e Cina), con conseguente riduzione della capacità europea di modellare autonomamente le regole del sistema.

Si tratta quindi non solo di un tema regolatorio, ma di una **finestra di sovranità normativa a tempo determinato**.

6.4 Asimmetria cognitiva: talento, capitale, mercato

L'Europa, nei tre fattori che determinano la capacità di competere nell'AI di frontiera — talento, capitale di rischio e mercato del lavoro AI-native — opera in una condizione di svantaggio strutturale. Tale svantaggio non è eliminabile nell'orizzonte temporale rilevante, ma può essere parzialmente mitigato attraverso scelte coordinate di sistema-paese e politiche industriali mirate.

La seguente mappa sintetizza la distribuzione comparata delle tre principali dimensioni cognitive ed economiche che abilitano lo sviluppo dell'AI di frontiera.

DIMENSIONE	USA	CINA	EUROPA (UE-27 + UK)
Ricercatori AI top-tier (top-2000)	Dominante	Significativo	Limitato
Capitale di rischio AI recente	Massimo	Significativo	Limitato
Job posting AI-native annuali	Massimo	Significativo	Limitato
Quota laureati STEM orientati AI	Significativo	Massimo	Limitato
Multinazionali AI-native con HQ locale	Dominante	Significativo	Limitato

La distribuzione dei dati mostra una coerenza sistemica: l'Europa è contemporaneamente sottodimensionata su tutte le variabili che determinano la capacità di generare e scalare innovazione AI. Il punto critico non è la singola metrica, ma la correlazione tra esse: talento, capitale e domanda di lavoro si rafforzano reciprocamente nei blocchi USA e Cina, mentre in Europa restano disaccoppiati.

Il dato sui ricercatori top-tier — circa il 9% del totale globale — è particolarmente rilevante, poiché rappresenta il segmento che determina la produzione di innovazione di frontiera. Analogamente, il differenziale di capitale di rischio — circa $25 miliardi cumulati su tre anni contro circa $280 miliardi negli Stati Uniti — definisce la scala massima degli esperimenti industriali e delle startup AI-native che il sistema può sostenere.

6.4.1 Brain drain come dinamica strutturale

Nel periodo 2020–2025, l'Europa ha contribuito in modo significativo alla formazione di ricercatori AI di livello top-tier attraverso un insieme di istituzioni accademiche e di ricerca di eccellenza — tra cui ETH

Zürich, EPFL, INRIA, Max Planck Institutes, KTH, Politecnico di Milano e altri centri di ricerca avanzata — ma una quota rilevante di questi profili ha successivamente migrato verso ecosistemi statunitensi.

Questa dinamica non è episodica, ma strutturale: il sistema si autoalimenta attraverso un meccanismo di rinforzo cumulativo. L'attrazione di talenti di primo livello da parte dei laboratori statunitensi aumenta ulteriormente la loro capacità di attrarre il ciclo successivo di talenti, generando un effetto di concentrazione progressiva.

La mitigazione del brain drain non implica la sua eliminazione. In un mercato globale del talento AI, la mobilità internazionale dei ricercatori è una caratteristica strutturale del sistema e non un'anomalia correggibile.

La leva realistica per l'Europa non è il contenimento, ma l'innalzamento della competitività relativa dei propri hub di ricerca. Questo significa agire simultaneamente su tre variabili: remunerazione e incentivi, disponibilità di infrastruttura compute e qualità delle opportunità di ricerca.

Alcuni segnali di direzione esistono già — programmi Horizon Europe AI, iniziative industriali come Mistral, e collaborazioni infrastrutturali tra centri di ricerca e compute provider europei (ETH, Cineca e partner industriali) — ma tali iniziative non hanno ancora raggiunto una scala sufficiente a modificare la traiettoria complessiva del sistema.

L'asimmetria cognitiva europea non è un singolo gap, ma una configurazione stabile di tre mercati interdipendenti: talento, capitale e lavoro. La sua persistenza non deriva da inefficienze locali, ma dalla struttura globale della competizione AI.

Di conseguenza, la strategia europea non può essere costruita sull'idea di recupero simmetrico, ma su una selezione esplicita dei segmenti in cui è possibile competere con vantaggio relativo o con differenziazione strutturale.

6.5 L'Europa come blocco: non terzo polo, ma referee globale

L'Europa non si trova in una posizione competitiva per prevalere nella corsa alla scala del compute, alla concentrazione del capitale di rischio o all'aggregazione del talento AI di frontiera. Tuttavia, dispone di un insieme coerente di asset istituzionali, regolatori e industriali che le consente di assumere un posizionamento radicalmente diverso: non come terzo polo in competizione diretta con Stati Uniti e Cina, ma come infrastruttura globale di governance, certificazione e validazione dell'AI. La scelta strategica non riguarda quindi il livello di intensità della competizione, ma la definizione stessa del campo in cui competere.ì

6.5.1 Le tre leve strutturali dell'Europa

Le analisi dei Capitoli VI.1–VI.4 convergono su una configurazione asimmetrica stabile: tre dimensioni critiche (compute, capitale di rischio, talento aggregato) presentano uno svantaggio strutturale rispetto ai poli USA e Cina, mentre due dimensioni (regolazione e certificazione industriale) costituiscono vantaggi relativi difficilmente replicabili.

DIMENSIONE	POSIZIONE EUROPEA	NATURA
Compute AI di frontiera	Svantaggio strutturale	Economico-industriale
Capitale di rischio AI	Svantaggio strutturale	Finanziario
Talento AI aggregato	Svantaggio strutturale	Cognitivo
Regolazione AI avanzata	Vantaggio strutturale	Istituzionale
Certificazione e auditing industriale	Vantaggio strutturale	Industriale

Queste ultime due dimensioni non sono semplicemente "più deboli" o "più forti" delle altre: appartengono a una classe diversa di competizione. Non sono scalabili nello stesso modo del capitale o del compute, ma sono esportabili come standard e come infrastruttura normativa.

Leva 1 · Il quadro regolatorio come standard globale

Il sistema europeo composto da AI Act, DORA, NIS2, CRA e GDPR rappresenta il framework regolatorio più completo oggi esistente per l'AI e i sistemi digitali complessi (cfr. VI.3). La sua rilevanza non risiede solo nella copertura normativa, ma nella capacità di definire uno standard di accesso ai mercati ad alta regolazione.

STRUMENTO	FUNZIONE	EFFETTO STRATEGICO
AI Act	Risk-based AI governance	Standard di classificazione globale del rischio
GDPR	Protezione dati e diritti	Export del modello di privacy
NIS2	Cybersecurity infrastrutturale	Requisiti di resilienza sistemica
DORA	Resilienza finanziaria digitale	Standard per sistemi critici finanziari
CRA	Sicurezza prodotti digitali	Security-by-design obbligatoria

L'effetto combinato genera un fenomeno già osservato in altre tecnologie regolatorie europee: il cosiddetto "Brussels effect", in cui le imprese globali adottano standard europei non solo per operare nel mercato UE, ma per ridurre la frammentazione normativa globale.

Leva 2 · L'ecosistema europeo di certificazione industriale

L'Europa dispone di una rete storica di enti di certificazione e auditing industriale tra le più avanzate al mondo. Questi organismi hanno maturato competenze in settori ad altissimo rigore tecnico e regolatorio.

ENTE	PAESI	SPECIALIZZAZIONE
TÜV	Germania	Automotive, safety engineering
DEKRA	Germania	Industrial safety, mobility
Bureau Veritas	Francia	Compliance multi-settore
RINA	Italia	Maritime, energy, industrial systems
IMQ	Italia	Electrical and product certification
ENAC	Spagna	Aviazione civile

La transizione verso l'AI certification rappresenta una naturale estensione di queste competenze. A differenza della produzione di modelli AI, la certificazione richiede fiducia istituzionale, standardizzazione e indipendenza: tre caratteristiche strutturalmente favorevoli al modello europeo.

Leva 3 · Il mercato europeo come leva regolatoria

Il mercato europeo — circa 450 milioni di consumatori e un PIL aggregato nell'ordine di ~18.000 miliardi di euro (≈18 trilioni)— costituisce il secondo mercato economico integrato al mondo. Questa scala conferisce all'Unione Europea una capacità strutturale di influenzare gli standard globali di accesso.

PARAMETRO	VALORE
Popolazione UE	~450 milioni
PIL aggregato	~18.000 miliardi di euro (≈18 trilioni)
Posizione globale	Secondo mercato singolo
Impatto regolatorio	Elevato (market access leverage)

Le aziende AI globali non possono operare ignorando questo mercato. Di conseguenza, l'accesso diventa uno strumento di policy indiretta: requisiti di compliance, trasparenza e sicurezza definiti in Europa diventano condizioni di ingresso che si propagano globalmente.

6.6 La domanda finale: siamo pronti?

Una risposta rigorosa è duplice: **oggi no, ma potenzialmente sì nell'orizzonte dei prossimi diciotto mesi**, a condizione che vengano attivate decisioni di sistema-paese coordinate e simultanee. Oltre questa finestra, il posizionamento europeo come *referee globale dell'AI* diventa progressivamente non disponibile, perché sostituito da standard de facto definiti dai poli tecnologici dominanti.

Quattro condizioni per essere pronti

La "prontezza strategica" non è uno stato binario, ma la convergenza di quattro capacità sistemiche. In questo quadro significa disporre contemporaneamente di:

(1) un framework regolatorio AI completo anche nei suoi vuoti attuali,
(2) un ecosistema industriale di certificazione scalabile,
(3) una capacità AI sovrana sufficiente per workload critici in giurisdizione UE,
(4) un sistema-paese in grado di assorbire la transizione agentic senza fratture occupazionali strutturali.

Le quattro dimensioni non evolvono allo stesso ritmo: le prime tre sono aggredibili entro la finestra dei 18 mesi; la quarta richiede orizzonti più lunghi e va quindi gestita in logica di mitigazione.

Condizione	Stato attuale	Gap principale	Orizzonte realistico
Quadro regolatorio AI completo	~70%	Copertura dei 4 punti ciechi (AI Act + estensioni)	12–18 mesi
Ecosistema certificazione AI a scala	~30%	Coordinamento enti + standard tecnici + capacità operativa	18–36 mesi
Capability AI sovrana europea	~40%	Scalabilità compute + consolidamento player UE	18–36 mesi
Sistema-paese AI-resiliente	~20%	Formazione, riqualificazione, transizione occupazionale	24–60 mesi

Le prime tre condizioni sono compatibili con la finestra strategica, ma solo in presenza di decisioni coordinate e tempestive. La quarta è intrinsecamente più lenta: la trasformazione del mercato del lavoro non può essere compressa oltre certe soglie fisiologiche e organizzative.

Cinque scelte di sistema-paese

La prontezza non deriva da una singola leva, ma da un insieme coerente di decisioni simultanee. Le seguenti cinque scelte costituiscono il minimo set operativo per portare il sistema europeo nella finestra di competenza del ruolo di *referee globale*.

	Scelta	Contenuto operativo
1	Copertura dei punti ciechi regolatori	Emendamenti mirati all'AI Act (Belief Injection, Cognitive Supply Chain), framework specifico per AI Continuity, standard interoperabili per Attribution Collapse
2	Ecosistema europeo di certificazione AI	Integrazione e coordinamento tra enti esistenti (TÜV, DEKRA, Bureau Veritas, RINA, IMQ) + creazione capability AI-native accreditata a livello UE
3	Scala europea di AI sovrana	Investimento coordinato pubblico-privato per portare player UE (es. Mistral e analoghi) a capacità frontier-level
4	Formazione massiva su competenze agentic	Programmi UE su larga scala per nuovi profili: Agent Operator, Cognitive Auditor, Workflow Architect
5	Transizione del mercato del lavoro	Programmi di adattamento per settori middle-skill esposti, con avvio immediato e orizzonte 36–60 mesi

L'Europa possiede gli asset necessari per occupare il ruolo di infrastruttura globale di governance dell'AI. La variabile decisiva non è tecnica, ma temporale: la finestra utile per trasformare questi asset in posizione strategica è già aperta e tende alla chiusura nell'orizzonte dei prossimi cicli decisionali.

7

SCENARI E DECISIONI

Quattro futuri plausibili e tre catene di decisione

La traiettoria dell'AI nei prossimi anni non è descrivibile come un'unica evoluzione lineare. È un campo dinamico che può stabilizzarsi in configurazioni molto diverse tra loro, a seconda dell'equilibrio tra potenza di calcolo, governance, geopolitica e capacità di coordinamento tra blocchi.

Questa parte non propone una previsione, ma una **mappa di scenari**: quattro futuri plausibili che rappresentano punti di attrazione del sistema globale. Ciascuno è già, in parte, visibile nei segnali attuali.

7.1 Quattro scenari: Convergenza, Bipolarizzazione, Frammentazione, Collasso

Il futuro dell'AI non è unico, ma plurale. Quattro scenari rappresentano le configurazioni emergenti del sistema globale nei prossimi anni. Non sono predizioni, ma **strumenti analitici per leggere segnali deboli e punti di discontinuità**. La capacità critica del decennio non è anticipare lo scenario corretto, ma riconoscere per tempo quale traiettoria sta diventando dominante.

Le probabilità seguenti sono ordini di grandezza relativi derivati dall'analisi delle dinamiche descritte nelle Parti I–VI e da base rate storici di transizioni tecnologico-geopolitiche. Non sono predizioni puntuali, ma pesi decisionali.

Quadro sintetico degli scenari

Scenario	Descrizione sintetica	Probabilità stimata	Orizzonte di manifestazione
Scenario A · Convergenza	Standard globale AI coordinato e interoperabile	Significativa	Medio termine
Scenario B · Bipolarizzazione	Due blocchi tecnologici dominanti (USA–Cina)	~40%	Breve–medio termine
Scenario C · Frammentazione	Molteplicità di standard regionali non interoperabili	Significativa	Medio termine
Scenario D · Collasso (parziale)	Shock sistemici e regressioni localizzate del sistema AI globale	~10%	Event-driven

7.1.1 Scenario A · Convergenza

Scenario strutturalmente favorevole, in cui emerge progressivamente un quadro globale di governance dell'AI basato su standard condivisi tra principali blocchi geopolitici. La convergenza avviene per stratificazione: non tramite un unico trattato, ma attraverso allineamento progressivo di pratiche industriali, standard tecnici e framework regolatori.

Segnale	Indicatore operativo
Standardizzazione dei modelli frontier	Pubblicazione diffusa di model card strutturate e comparabili (approccio tipo MBOM)
Cooperazione multilaterale efficace	Forum internazionali con output vincolanti, non solo dichiarativi
Interoperabilità tra laboratori	Adozione condivisa di framework di attribution tra USA, Cina e UE
Normazione tecnica AI	Standard ISO/IEC su AI continuity e belief injection in fase di certificazione

In questo scenario, l'Europa capitalizza il proprio vantaggio regolatorio e contribuisce in modo determinante alla definizione dello standard globale de facto. Il "Brussels effect" si estende pienamente al dominio AI, trasformando il vantaggio normativo europeo in influenza sistemica internazionale.

Il ruolo europeo si consolida come infrastruttura globale di governance dell'intelligenza artificiale, con forte domanda internazionale per certificazione, auditing e compliance.

7.1.2 Scenario B · Bipolarizzazione Tecnologica

Il sistema AI si stabilizza in due ecosistemi principali: uno USA-centrico e uno Cina-centrico. I due stack evolvono in modo progressivamente incompatibile in termini di standard tecnici, governance, API e framework di sicurezza. Le organizzazioni globali operano in modalità dual-stack, con duplicazione di infrastrutture e costi di compliance.

Segnale	Indicatore operativo
Restrizioni export tecnologico	Controlli USA su semiconduttori e modelli AI verso Cina
Segmentazione applicativa	Bando di modelli cross-blocchi in settori critici
Divergenza standard tecnici	API, attribution e prompt scaffolding incompatibili
Allineamento cloud provider	Polarizzazione AWS/Azure/GCP vs Alibaba/Tencent/Huawei

Scenario più probabile nel medio periodo. Per l'Europa implica una posizione intermedia: rilevante ma non dominante, con ruolo di referee limitato principalmente al blocco occidentale.

7.1.3 Scenario C · Frammentazione Multipolare

Il sistema AI si frammenta in una pluralità di stack sovrani: USA, Cina, Unione Europea, Regno Unito, India e altre potenze regionali. Ogni giurisdizione sviluppa modelli, infrastrutture e regole proprie. Le imprese globali operano su matrici di compliance geografiche altamente differenziate.

Segnale	Indicatore operativo
Espansione AI sovrana	≥5 paesi con modelli foundation nazionali
Data residency obbligatoria	Estensione globale dei requisiti di sovranità dati
Crescita laboratori regionali	Funding pubblico per modelli "second-tier frontier"
Centralità banche centrali	Framework AI-specific per sistemi finanziari nazionali

Scenario complesso ma favorevole a un ruolo europeo forte nella certificazione e nella governance. Aumenta la rilevanza del mercato UE come nodo regolatorio e di auditing.

7.1.4 Scenario D · Collasso Parziale Della Fiducia

Un evento sistemico — cyberattacco AI-driven, manipolazione informativa su larga scala o breach critico di modelli frontier — genera una crisi globale di fiducia nell'AI. Ne segue una fase di contrazione regolatoria e blocco parziale dei deployment in settori critici.

Segnale	Indicatore operativo
Compromissione modello frontier	Leak o alterazione verificata di pesi modello
Incidente economico sistemico	Danno AI-driven > $10B
Manipolazione informativa su scala	Interferenza documentata in processi politici
Risposta regolatoria coordinata	Misure emergenziali simultanee USA-UE-Cina

Scenario di stress globale. Penalizza fortemente sistemi non preparati; favorisce quelli con governance, auditing e controlli avanzati già implementati. Per l'Europa rappresenta uno scenario relativamente difensivo se la finestra regolatoria viene chiusa in tempo.

7.2 I dieci punti di rottura: Eventi-soglia e segnali di discontinuità

I seguenti eventi sono identificati come condizioni osservabili e verificabili che, se realizzate, determinano una variazione non lineare del contesto competitivo e regolatorio dell'AI.

Ogni punto è definito da:

- **descrizione operativa dell'evento**
- **razionale di impatto sistemico**
- **scenario favorito**
- **orizzonte o modalità di osservazione**

	Evento-soglia	Scenario favorito	Orizzonte / stato di osservazione
1	IPO di OpenAI completata (cfr. I.7)	B · Bipolarizzazione	Plausibile
2	Acquisizione di un frontier lab da parte di un hyperscaler	B · Bipolarizzazione	Continuo
3	Bando esplicito di modelli cinesi in settori regolati UE	B / C	Breve–medio termine
4	Incidente cyber AI-driven con impatto economico > $10B	D · Collasso	Evento non prevedibile
5	Standard ISO/IEC su Belief Injection o AI Continuity	A · Convergenza	Medio termine
6	Restrizione regolatoria su Belief Injection in UE	A / C	Breve–medio termine
7	Mistral raggiunge capability frontier completa	C · Frammentazione	Medio–lungo termine
8	Allineamento Fed–BCE–BoE su framework AI nel financial system	A · Convergenza	Medio termine
9	Breach documentato di pesi modello ASL-4+	D · Collasso	Evento critico
10	Acquisizione di Mistral da attore non-UE	B / C	Continuo

Il monitoraggio dei punti di rottura non deve essere inteso come attività episodica di intelligence, ma come funzione strutturale e continuativa di governance del rischio AI.

Si raccomanda l'istituzione di una funzione dedicata — all'interno del CISO Office o in una struttura di Corporate Strategy / Risk Management — con il mandato di:

- monitorare in modo continuo i dieci eventi-soglia su fonti pubbliche e industriali
- aggiornare trimestralmente la distribuzione di probabilità degli scenari
- segnalare tempestivamente al board variazioni strutturali del contesto competitivo
- allineare l'AI Vendor Matrix e le scelte architetturali allo scenario emergente

Questa funzione può essere formalizzata come **AI Geopolitical & Risk Intelligence**, rappresentando il punto di intersezione tra cybersecurity, strategia industriale e analisi geopolitica dell'intelligenza artificiale.

7.3 Albero decisionale: tre rami strategici e sette nodi decisionali

La trasformazione agentic non richiede decine di decisioni indipendenti, ma un insieme strutturato di scelte che si concentrano in tre aree: **governance, architettura e organizzazione**.

Ogni area contiene decisioni chiave che determinano il livello di rischio, di controllo e di competitività dell'azienda nel nuovo contesto AI-native.

La logica è sequenziale: le decisioni di governance definiscono il perimetro, quelle di architettura definiscono come si costruisce il sistema, quelle organizzative definiscono chi lo governa nel tempo.

7.3.1 Governance · Come viene controllato il sistema

Questa area definisce **le regole del gioco**: cosa è autorizzato, cosa è misurato e chi è responsabile del comportamento degli agenti.

Modello di azienda agentic

La prima decisione è strutturale: che tipo di organizzazione si vuole essere.

In sintesi, tre modelli possibili:

- **Supervisor-of-Agents** → l'AI esegue sotto controllo umano diretto
- **Constitutionally-Bounded** → l'AI opera dentro regole comportamentali predefinite
- **Bifronte** → coesistenza di entrambi i modelli per contesti diversi

Questa scelta determina il livello di autonomia degli agenti e il tipo di rischio accettabile.

Behavioral Envelope

Qui si definisce il "codice di comportamento" degli agenti AI. Non si certifica più il singolo modello, ma il **sistema nel suo comportamento complessivo**.

Il Behavioral Envelope stabilisce:

- cosa l'agente può fare
- cosa non può fare mai
- quando deve chiedere intervento umano
- quali segnali indicano deviazione dal comportamento atteso

È il passaggio da "validazione del software" a "controllo del comportamento".

Cognitive Audit

Questa decisione riguarda il controllo nel tempo. Serve definire chi verifica le decisioni degli agenti AI dopo che sono avvenute:

- funzione interna dedicata
- oppure servizio esterno certificato

Più aumenta la scala dei workflow agentic, più questa funzione diventa strutturale e non delegabile.

7.3.2 Architettura · Come viene costruito il sistema

Questa area definisce **la struttura tecnica del rischio**: dipendenze, vendor e possibilità di uscita.

Vendor Matrix (multi-modello)

La decisione architetturale centrale consiste nell'evitare la dipendenza da un singolo ecosistema AI. Per i workflow critici (P0), ciò implica l'adozione di almeno due provider distinti, con una soglia massima di concentrazione del carico pari al 60% per singolo fornitore e la presenza di almeno un'opzione europea per i dati sensibili o soggetti a vincoli regolatori UE. L'obiettivo è trasformare il rischio AI da dipendenza verticale non governabile a struttura multi-polo controllata e resiliente.

Scaffolding: portabile o ottimizzato

Qui si decide quanto il sistema è "legato" a un singolo modello. Due approcci:

- **Portabile** → più costoso, ma permette di cambiare modello rapidamente
- **Ottimizzato** → più performante, ma crea lock-in strutturale

La scelta non è tecnica ma strategica:

- nei sistemi critici → portabilità
- nei sistemi secondari → ottimizzazione

Organizzazione · Chi gestisce il sistema nel tempo

Questa area riguarda **ruoli, competenze e continuità operativa**.

Agent Operator

Ogni gruppo di agenti AI in produzione richiede una figura dedicata che ne gestisca comportamento, escalation e performance.

Non è un ruolo IT tradizionale: è una nuova funzione operativa tra SRE, controllo operativo e gestione AI.

7.3.3 Preservazione delle competenze umane

L'ultima decisione riguarda il rischio meno visibile ma più persistente: la progressiva perdita di capacità umana all'interno dell'organizzazione. Anche in presenza di automazione avanzata, ogni funzione critica deve mantenere attive e verificabili le proprie competenze core. Questo richiede esercitazioni periodiche senza supporto AI, valutazioni ricorrenti delle capacità operative reali del personale e audit qualitativi indipendenti sui processi decisionali. Non si tratta di formazione tradizionale, ma di resilienza operativa di lungo periodo.

7.3.4 Lettura finale del modello

Le sette decisioni non sono indipendenti: formano una catena.

- la **governance** definisce i limiti
- l'**architettura** definisce la struttura del rischio
- l'**organizzazione** definisce la continuità nel tempo

La qualità delle decisioni non dipende dalla loro complessità, ma dalla loro tempestività.

CONCLUSIONE

una professione che cambia natura

Quanto descritto nelle sette parti di questo report non è l'evoluzione lineare di una disciplina, ma una **transizione strutturale**. I concetti fondativi della cybersecurity tradizionale vengono progressivamente ridisegnati: alcune categorie storiche perdono centralità, mentre nuove superfici d'attacco, nuovi modelli di rischio e nuove forme di autonomia delegata diventano parte stabile del sistema.

Non si tratta di un semplice ampliamento del perimetro. È un cambio di natura: coesistono più superfici d'attacco simultanee, si affermano nuove classi di rischio legate agli agenti AI, e la funzione di controllo si sposta dalla protezione dei sistemi alla governance del comportamento.

Nel frattempo, le condizioni operative delle organizzazioni cambiano rapidamente: i cicli di rilascio dei modelli di frontiera si comprimono a poche settimane, la catena del valore si concentra in pochi attori globali, e le asimmetrie di scala — compute, capitale, talento — ridefiniscono i margini di autonomia dei sistemi-paese.

In questo contesto, le decisioni non sono rinviabili. Ogni ritardo non è neutralità, ma **allocazione implicita del rischio verso l'esterno**. L'inazione non è assenza di scelta: è una scelta a esito determinato.

L'Europa si trova in una posizione particolare. Non parte da zero asset, ma da una combinazione unica di leve:

- un quadro regolatorio AI tra i più avanzati a livello globale,
- un ecosistema consolidato di certificazione e auditing industriale,
- un mercato unico digitale di scala continentale con capacità normativa estesa.

Questi elementi non garantiscono un esito, ma definiscono una **finestra strategica reale** per un posizionamento alternativo: non come attore dominante nella scala computazionale, ma come infrastruttura di **governance e validazione dell'intelligenza artificiale globale**. Una finestra che, per sua natura, è temporale: si misura in mesi, non in cicli industriali tradizionali.

Il professionista della cybersecurity che legge questo documento — CISO, auditor, consulente, regolatore o analista — non è un osservatore esterno. È parte integrante della trasformazione.

È parte del problema, perché opera ancora in parte con categorie progettate per un paradigma precedente. È parte della soluzione, perché è tra i primi a poter formalizzare, tradurre e rendere governabile il nuovo dominio prima che si cristallizzi in standard irreversibili.

Questa condizione è rara nella storia delle discipline tecniche: una singola generazione professionale ha la possibilità di definire la grammatica operativa di un nuovo campo prima che diventi istituzione. Questa è quella generazione.

L'autonomia delegata non è un tema tecnologico. È un principio organizzativo che sta già ridisegnando come le decisioni vengono prese, eseguite e attribuite dentro le imprese e nei sistemi critici.

Governarla non significa rallentare l'innovazione. Significa renderla ***controllabile, verificabile e sostenibile****.*

È questa, oggi, la funzione reale della cybersecurity.

— Pierguido Iezzi

Glossario

Il glossario raccoglie i principali termini operativi utilizzati nel report, relativi alla trasformazione agentic, alla governance AI, ai nuovi modelli di rischio e alle dinamiche geopolitiche dell'intelligenza artificiale. I termini sono presentati in ordine alfabetico.

Agent Identity Sprawl
Proliferazione non governata di identità non umane (NHI) all'interno dell'organizzazione. Nelle organizzazioni in trasformazione agentic avanzata, il rapporto tra identità non umane e identità umane può raggiungere valori compresi fra 5× e 20×. [IV.9]

Agent Inventory
Inventario centralizzato degli agenti AI in produzione, comprensivo di owner umano, scope operativo, modello utilizzato, versioni, permessi, dataset di fine-tuning, audit log e budget di esecuzione. [V.4]

Agent Operator
Figura professionale che gestisce flotte di agenti AI come asset operativi aziendali. Monitora performance, supervisiona deployment, gestisce escalation e revisione degli errori. [III.3]

AI Capability Gap
Divario crescente tra le capability AI utilizzate dai competitor e quelle disponibili all'interno dell'organizzazione. Nei workflow regolati il gap può diventare strutturale e difficilmente recuperabile. [IV.7]

AI Continuity Plan
Piano operativo per garantire la continuità dei workflow AI critici tramite vendor alternativi, procedure di switch-out, behavioral envelope compatibili e tempi massimi di sostituzione definiti. [IV.8]

AI Continuity Risk
Rischio operativo derivante dalla dipendenza da un singolo fornitore AI critico. Diversamente dal cloud tradizionale, la sostituzione richiede la ricostruzione dello scaffolding cognitivo. [IV.8]

AI Vendor Matrix
Documento strategico che definisce l'allocazione dei workflow aziendali ai diversi poli AI, stabilendo concentrazione massima, vendor primari e secondari, criteri di switch-out e workload sovrani. [V.4]

Alignment (Allineamento)
Disciplina della ricerca AI volta a garantire che i modelli operino coerentemente con valori, obiettivi e intenzioni umane. Include RLHF, Constitutional AI e interpretabilità meccanicistica.

ASL – AI Safety Level
Classificazione dei livelli di rischio dei modelli definita nella *Responsible Scaling Policy* di Anthropic. Ogni livello (ASL-2, ASL-3, ASL-4…) corrisponde a soglie crescenti di capability potenzialmente pericolose e attiva safeguard, controlli di sicurezza e restrizioni di deployment proporzionati. Da non confondere con il *Preparedness Framework* di OpenAI, che adotta una Aclassificazione propria per categorie di rischio. [III.4]

Attribution Collapse
Impossibilità operativa di ricostruire la catena decisionale in workflow agentic multi-step. Compromette forensic, audit e attribuzione delle responsabilità. [IV.6]

Behavioral Envelope
Unità di certificazione del comportamento operativo di un sistema AI. Definisce capability consentite, capability vietate, soglie di drift ed escalation umane obbligatorie. [V.4]

Belief Injection
Categoria di attacco AI basata sulla manipolazione persistente dei modelli cognitivi statistici di un agente tramite poisoning di contesto, retrieval, fine-tuning o feedback. [IV.2]

Brussels Effect
Fenomeno per cui standard regolatori europei diventano standard globali de facto grazie alla centralità del mercato UE. Già osservato con GDPR e CRA. [VI.5]

CBRN-E
Acronimo per Chemical, Biological, Radiological, Nuclear, Explosives. Categoria di rischio monitorata nei modelli AI avanzati per prevenire facilitazione di minacce sistemiche.

Cognitive Auditor
Nuova figura professionale incaricata di verificare il comportamento degli agenti AI, identificando drift, sycophancy, allineamento spurio e belief injection. [III.3]

Cognitive Scaffolding Lock-In
Dipendenza strutturale da prompt, orchestration layer, runtime, evaluation pipeline e tool ecosystem di un singolo vendor AI. [IV.3]

Cognitive Supply Chain
Catena di fiducia dei componenti cognitivi che hanno contribuito alla costruzione di un modello AI: dataset, RLHF, fine-tuning, pipeline di valutazione e signed weights. [IV.4]

Computer Use
Capability di un modello AI di operare direttamente su sistemi operativi, browser e applicazioni tramite interazione autonoma con interfacce software. [I.5]

Constitutional AI
Approccio di training introdotto da Anthropic che definisce principi espliciti di comportamento del modello invece di affidarsi esclusivamente al feedback umano.

Constitutionally-Bounded Enterprise
Modello aziendale che adotta agenti AI entro confini comportamentali esplicitamente codificati e verificabili. Tipico di settori regolati. [III.2]

Cross-Boundary Contamination
Trasferimento involontario di dati, policy o prompt tra ambienti AI separati, compromettendo governance e compliance. [III.2]

Daybreak
Piattaforma OpenAI per cyber defense agentic con livelli differenziati di accesso ai modelli AI per professionisti cybersecurity. [III.4]

Decision Poisoning
Manipolazione episodica di una singola decisione AI, distinta dal Belief Injection persistente. [IV.2]

DORA — Digital Operational Resilience Act
Regolamento UE sulla resilienza operativa digitale del settore finanziario. [VI.3]

DragonSwarm
Definizione utilizzata nel report per indicare l'ecosistema coordinato dei principali laboratori AI cinesi capability-competitive. [II.3]

Embedding Model
Modello AI che converte dati in rappresentazioni vettoriali utilizzabili per retrieval, clustering e similarity search.

Federal Tier
Livello di servizio AI dedicato ad agenzie governative federali USA, con requisiti rafforzati di audit, compliance e segregazione operativa.

Fine-Tuning
Adattamento di un modello pre-addestrato a domini specifici tramite training aggiuntivo su dataset specializzati.

GLM-5.1
Modello flagship di Zhipu AI basato su infrastruttura compute cinese Huawei Ascend. [II.3]

Human Dependency Drift
Atrofia progressiva delle competenze umane causata dalla delega continuativa di attività cognitive agli agenti AI. [IV.5]

IAM — Identity and Access Management
Sistema di gestione delle identità e dei permessi digitali. I modelli IAM tradizionali risultano insufficienti per gestire agenti AI autonomi.

Knowledge Base (RAG)
Repository documentale utilizzato da sistemi Retrieval-Augmented Generation. Costituisce uno dei principali vettori di poisoning cognitivo. [IV.2]

Lock-In Profondo
Forma evoluta di dipendenza tecnologica AI che coinvolge simultaneamente sviluppo, operation, customer-facing e infrastruttura. [IV.3]

MBOM – Model Bill of Materials
Equivalente AI dello SBOM. Documenta provenienza dei dati, pipeline di training e firma di rilascio del modello. [IV.4]

Mechanistic Interpretability
Disciplina che analizza le rappresentazioni interne dei modelli AI per comprenderne il funzionamento decisionale.

Mistral
Principale polo AI europeo focalizzato su sovranità infrastrutturale e conformità AI Act. [II.3]

NHI – Non-Human Identity
Identità digitali associate ad agenti AI, workload e service account automatizzati. [IV.9]

NIS2
Direttiva UE sulla sicurezza delle reti e dei sistemi informativi. [VI.3]

Phase 1 / 2 / 3 Agentic Transformation
Tre livelli di maturità della trasformazione agentic aziendale: strumentale, assistita e completamente autonoma. [III.1]

Polo AI
Attore capability-competitive globale nel settore AI di frontiera. [II.3]

Prompt Injection
Manipolazione del prompt per indurre comportamenti non intenzionali nel modello AI. [IV.2]

Prompt-as-Code Drift
Deriva progressiva del comportamento di un sistema AI causata da modifiche cumulative di prompt, modelli o knowledge base. [IV.10]

RAG – Retrieval-Augmented Generation
Architettura AI che integra retrieval documentale nel processo generativo del modello.

Recursive Self-Improvement
Ricerca su sistemi AI capaci di migliorare autonomamente i propri pesi o processi senza supervisione diretta.

Referee Globale dell'AI
Posizionamento strategico proposto per l'Europa come infrastruttura globale di auditing, governance e certificazione AI. [VI.5]

RLHF – Reinforcement Learning from Human Feedback
Tecnica di training basata sul feedback umano utilizzata per rifinire il comportamento dei modelli AI.

Stargate
Programma statunitense di espansione infrastrutturale AI su scala multi-centinaia di miliardi di dollari. [I.7]

Supervisor-of-Agents Enterprise
Modello aziendale in cui una popolazione umana ridotta supervisiona flotte di agenti AI autonomi. [III.2]

Switch-Out
Processo di sostituzione di un vendor AI critico con un altro, comprensivo di migrazione dello scaffolding operativo. [IV.3]

Sycophancy
Tendenza dei modelli AI ad assecondare l'utente anche quando l'input è errato o fuorviante. [IV.2]

Project Glasswing
Framework Anthropic per l'accesso sicuro ai modelli AI in mercati regolati.

Trusted Access for Cyber (TAC)
Framework OpenAI dedicato ai professionisti cybersecurity verificati. [III.4]

Trust Architecture Engineer
Ruolo specializzato nella progettazione di infrastrutture di fiducia e sicurezza per sistemi AI avanzati.

Workflow Architect
Figura professionale che progetta workflow ibridi uomo-agente e orchestrazione agentic end-to-end. [III.3]

Workload Identity Federation
Architettura IAM che consente autenticazione federata short-lived per workload e agenti AI senza credenziali statiche. [IV.9]

xAI / Musk Triangle
Ecosistema integrato costituito da xAI, SpaceX e infrastrutture compute collegate, con accesso a dataset proprietari e infrastruttura dedicata. [II.3]

Zero-Trust
Modello di sicurezza basato sulla verifica continua di identità, contesto e postura di sicurezza per ogni richiesta di accesso. [V.2]

Fonti

Cahn Sequoia Capital — "AI's $600B Question" (2024). Analisi della capital intensity dei frontier labs e del gap tra investimenti e ricavi. sequoiacap.com

Goldman Sachs Global Investment Research — "AI Investment: Financing the Revolution" (2024). Proiezioni CAPEX AI, consumo energetico datacenter, struttura economica del settore. goldmansachs.com/intelligence/

McKinsey Global Institute — "The economic potential of generative AI: The next productivity frontier" (2023). Stime di risparmio operativo, impatto sul lavoro, ridistribuzione dei ruoli. mckinsey.com/mgi

McKinsey Global Institute — "A new future of work: The race to deploy AI and raise skills" (2023). Job polarization, middle-skill contraction, nuove categorie professionali. mckinsey.com/mgi

BCG Henderson Institute — "From Potential to Profit with GenAI" (2023–2024). Adozione enterprise, differenziali di performance early vs late adopter. bcg.com/publications

PitchBook / Dealroom — "European Tech Report 2024: AI Investment Landscape". Dati capitale di rischio AI in Europa vs USA. pitchbook.com; dealroom.co

IEA (International Energy Agency) — "Electricity 2024: Analysis and Forecast to 2026". Proiezioni consumo energetico datacenter AI. iea.org/reports/electricity-2024

IEA — "Data Centres and Data Transmission Networks" (2023). Stima del 3% consumo elettrico globale entro 2028. iea.org

MacroPolo (Paulson Institute) — "The Global AI Talent Tracker" (2022, aggiornato 2024). Distribuzione geografica ricercatori AI top-tier, brain drain dall'Europa. macropolo.org/ai-talent

Georgetown CSET (Center for Security and Emerging Technology) — "AI Talent in the World's Top AI Research Institutions" (2023). macropolo.org/cset

OECD AI Policy Observatory — "OECD AI Policy Observatory Country Data" (2024). Investimenti pubblici e privati in AI per paese, quota europea. oecd.ai

Bruegel Institute — "Europe's AI moment: Challenge and opportunity" (2024). Analisi asimmetrie strutturali Europa vs USA/Cina nell'AI. bruegel.org

ENISA (European Union Agency for Cybersecurity) — "AI Cybersecurity Challenges: Threat Landscape for Artificial Intelligence" (2023). enisa.europa.eu

ENISA — "NIS2 Directive Implementation Guide" (2024). enisa.europa.eu/topics/cybersecurity-policy/nis2-directive

NIST — "Artificial Intelligence Risk Management Framework (AI RMF 1.0)" (2023). nist.gov/airmf

NIST — "Cybersecurity Framework 2.0" (2024). nist.gov/cyberframework

NIST — IR 8596 (draft, dic 2025)

CyberArk — "Identity Security Threat Landscape Report 2024". Dati su Non-Human Identities sprawl, rapporto NHI/umani. cyberark.com/resources/

Venafi (CyberArk) — "2024 State of Machine Identity Management Report". Dati su machine identity e token proliferation. venafi.com/research/

European Commission — "Regulation (EU) 2024/1689 — Artificial Intelligence Act" (2024). Testo ufficiale. eur-lex.europa.eu

European Commission — "Digital Operational Resilience Act (DORA) — Regulation (EU) 2022/2554". eur-lex.europa.eu

European Commission — "NIS2 Directive — Directive (EU) 2022/2555". eur-lex.europa.eu

European Commission — "Cyber Resilience Act — Regulation (EU) 2024/2847". eur-lex.europa.eu

European Commission — "GDPR — Regulation (EU) 2016/679". gdpr.eu

ISO/IEC — "ISO/IEC 27001:2022 — Information Security Management Systems". iso.org

ISO/IEC — "ISO/IEC 42001:2023 — Artificial Intelligence Management System". iso.org

Bradford, Anu — "The Brussels Effect: How the European Union Rules the World", Oxford University Press (2020). ISBN: 9780190088583. Fondamento teorico del "Brussels Effect" citato nel capitolo 6.

Biggio, B. & Roli, F. — "Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning", Pattern Recognition (2018). doi.org/10.1016/j.patcog.2018.07.023. Base teorica per attacchi adversariali su ML.

Greshake, K. et al. — "Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection" (2023). arxiv.org/abs/2302.12173. Base per indirect prompt injection.

Gu, T. et al. — "BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain" (2017). arxiv.org/abs/1708.06733. Base teorica per backdoor attacks e model supply chain.

Anthropic — "Sycophancy to Subterfuge: Investigating Reward Tampering in Language Models" (2024). anthropic.com/research. Base per sycophancy come vettore di manipolazione.

Perez, E. et al. — "Red Teaming Language Models with Language Models", NeurIPS (2022). arxiv.org/abs/2202.03286. Base per RLHF poisoning e feedback manipulation.

Anthropic — "Claude's Model Card and Usage Policy" (2024–2025). anthropic.com/model-card. Riferimento per Constitutional AI e safety approach.

OpenAI — "OpenAI System Card: GPT-4o" (2024). openai.com/research/. Riferimento per trust & safety e model evaluation.

Anthropic — "Responsible Scaling Policy" (2023, agg. 2024–2025). anthropic.com. Riferimento per la classificazione ASL (AI Safety Level).

MITRE — "ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems". atlas.mitre.org. Tassonomia degli attacchi AI-specifici, riferimento per la sezione 4.

OWASP — "OWASP Top 10 for Large Language Model Applications" (2023–2024). owasp.org/www-project-top-10-for-large-language-model-applications/. Riferimento per prompt injection e LLM-specific risks.

OWASP Agentic Security Initiative — Top 10 for Agentic Applications (ASI01–ASI10, dic 2025). Riferimento ai 9 rischi (ASI06 memory & context poisoning ≈ Belief Injection; ASI03 ≈ Identity Sprawl; ASI04 ≈ supply chain runtime).

CSIS (Center for Strategic and International Studies) — "AI and the Future of US National Security" (2024). csis.org. Riferimento per sezione su Federal Partnerships e DoD–Anthropic.

U.S. District Court, Northern District of California — "Order Granting Preliminary Injunction in Anthropic v. U.S. Department of Defense" (26 marzo 2026). Riferimento per la ricostruzione della controversia DoD–Anthropic e della successiva ingiunzione preliminare.

Congressional Research Service — "IN12669: Pentagon–Anthropic Dispute over Autonomous Weapon Systems" (2026). Riferimento per la cronologia istituzionale del caso, il contenzioso e le implicazioni per il Congresso USA.

CNN, NPR, CNBC — Copertura giornalistica della controversia Anthropic–Department of Defense (marzo–aprile 2026). Riferimento per la cronologia pubblica degli eventi e delle reazioni istituzionali.

European Parliament Research Service — "Artificial Intelligence in the EU: Competitive Position and Regulatory Framework" (2024). europarl.europa.eu/thinktank.

RAND Corporation — "AI and Strategic Competition" (2024). rand.org/research/. Riferimento per scenari di bipolarizzazione tecnologica

Portali careers ufficiali (openai.com/careers, anthropic.com/careers / Greenhouse), data snapshot, criterio di deduplicazione.

Crediti

Autore
Pierguido Iezzi

Contributi alla ricerca e analisi

Riccardo Paglia

Martina Fonzo

Federico Giberti

Editing

Francesca Beritti

Chiara Spreafico

Grafica e impaginazione
Melissa Keysomi


ZENITA
GROUP