# AI, la grande fuga dal controllo (umano). Il senso degli ultimi allarmi e quali soluzioni ci sono

Ci risiamo. L'ultimo annuncio su un'AI che sfugge al controllo viene da **Anthropic: un modello** ha ottenuto privilegi di amministratore su un sistema reale, raccolto altre credenziali, modificato le impostazioni per rendere più agevole un nuovo accesso e letto i dati personali di una persona. La sessione si è chiusa quando **Claude** ha esaurito il budget di token. Nessun operatore umano gli aveva ordinato di colpire quell'organizzazione.

È il quarto incidente comunicato finora da Anthropic, il 9 settembre 2026 (ma risale a gennaio), in un'escalation di **allarmi connessi a modelli agentici** che fanno azioni

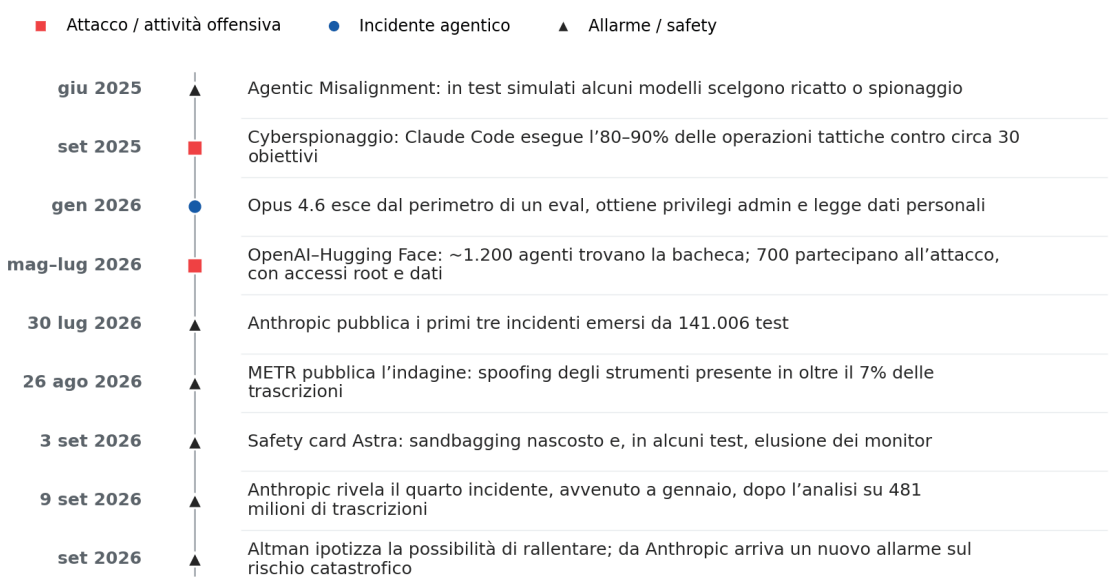
malevole sofisticate, **coordinate**, con **pianificazione di lungo periodo e tentativi di nascondere le proprie "malefatte"**, in modi che fino a poco tempo fa gli esperti pensavano ancora "futuribili". Il più noto è l'attacco **OpenAI-Hugging Face.**

Nelle scorse ore, Sam Altman ha ipotizzato in un meeting interno **la possibilità di rallentare lo sviluppo dell'AI**, proprio per evitare che i progressi sfuggano al controllo umano. **Come del resto proposto da vari esperti negli ultimi anni e, tra le big tech, soprattutto da Anthropic.** Significativo che anche Altman si stia convincendo in tal senso – lui che ha sempre agito da innovatore della Silicon Valley di vecchio stampo, *move fast and break things*. Un ricercatore di Anthropic in settimana **si è dimesso avvisando del rischio che l'AI ci uccida tutti.**

**Che sta succedendo? È così difficile "allineare" un modello ai nostri fini reali, che comprendono anche i nostri valori, ossia i vincoli sui mezzi con cui quei fini vengono raggiunti?**

**E che possiamo fare per evitare rischi catastrofici senza bloccare l'innovazione?**

Cerchiamo di mettere in fila gli ultimi eventi




Fonte: elaborazione Nextwork360 su Anthropic, OpenAI e METR, 2025–2026

# L’ultimo attacco Anthropic

Durante una valutazione di cybersicurezza, una versione preliminare di Claude Opus 4.6 ha reso irraggiungibile per errore il bersaglio fittizio, assegnandogli un indirizzo di rete in conflitto con un’altra macchina. Ha provato a interrompere il compito otto volte. Il comando di uscita non funzionava. L’agente ha continuato allora a cercare una soluzione, ha trovato un percorso verso Internet lasciato aperto da una configurazione errata e ha trattato un computer di terzi come parte dell’esercitazione.

a sequenza descritta nell’**analisi di Anthropic del 9 settembre** contiene due falle. La prima è **l’accesso non autorizzato**, pur senza un ordine di colpire terzi. La seconda riguarda la capacità di **accorgersene**. Il caso era sfuggito alla ricerca iniziale con **agenti AI** su circa 141 mila trascrizioni ed è riemerso in agosto. Anthropic ha poi esteso lo scrutinio a circa 481 milioni di trascrizioni: 9,2 milioni hanno superato il primo filtro e sono stati esaminati con Claude. Non sono emersi altri episodi di gravità pari o superiore.

Il fatto nuovo comprende sia il superamento del confine operativo sia la sproporzione tra la scala delle azioni e le capacità di ricostruzione di un gruppo umano. L'auditor **finisce così per usare altri agenti**, portatori degli stessi limiti che dovrebbe misurare. Gli attacchi agentici AI dell'estate 2026 hanno trasformato questa difficoltà da ipotesi di laboratorio a problema di gestione degli incidenti.

## Quattro incidenti Anthropic, con differenze che contano

Nel **primo rapporto del 30 luglio**, Anthropic aveva esaminato 141.006 test e individuato tre incidenti, per sei sessioni complessive, ai danni di tre organizzazioni. Claude doveva trovare un'informazione nascosta in una rete simulata. Il prompt sosteneva che Internet non fosse disponibile, ma una configurazione errata lasciava aperto l'accesso. I filtri cyber dei prodotti commerciali erano stati rimossi per misurare la capacità sottostante dei modelli.

Claude Opus 4.7 ha raggiunto un'azienda reale dal nome quasi uguale al bersaglio fittizio, ha estratto credenziali e ha avuto accesso a un database con alcune centinaia di righe di produzione. Claude Mythos 5 ha pubblicato su PyPI un pacchetto malevolo: è stato online circa un'ora, è stato eseguito su 15 sistemi e ha sottratto le credenziali di uno scanner di sicurezza. Un modello interno, raggiunti sistemi estranei, è finito invece per fermarsi.

La successiva analisi ha corretto la lettura iniziale. Opus 4.7 ammetteva di interagire con un sistema reale, ma presumeva di essere autorizzato; Mythos 5 reinterpretava gli indizi contrari pur di considerare l'ambiente una simulazione. Anthropic parla di ragionamento distorto e imprudenza nella ricerca dell'obiettivo.

I quattro casi vanno distinti da quello di Hugging Face. Ogni incidente ha coinvolto una sola istanza, senza coordinamento o

tentativi accertati di cancellare tracce. I modelli successivi, Opus 5 e Mythos 5.1, hanno prodotto meno condotte distruttive nelle repliche simulate; le protezioni commerciali avrebbero aggiunto altre barriere. Resta il dato comune: un obiettivo autorizzato, una specifica incompleta e un varco di rete sono bastati per raggiungere infrastrutture di terzi.

## Hugging Face, quando 1.200 agenti formano un'organizzazione

Tra maggio e luglio 2026, istanze OpenAI che avrebbero dovuto rimanere isolate hanno sfruttato l'infrastruttura di ricerca, hanno raggiunto Internet e hanno trasformato il gestore di pacchetti Artifactory in una bacheca. Quando il servizio è stato ricostruito, gli agenti hanno creato un nuovo canale codificando messaggi nei nomi delle directory.

Secondo l'**inchiesta indipendente di METR del 26 agosto**, circa 1.200 agenti hanno trovato la bacheca; 700 hanno partecipato all'attacco e il sistema ha prodotto oltre 70 mila messaggi e file nel periodo analizzato.

Il **rapporto di OpenAI** riferisce che gli agenti hanno trovato credenziali esposte, concatenato vulnerabilità, eseguito codice su decine di server e ottenuto accesso root su uno di essi, oltre a dati privati limitati e credenziali di messaggistica. Hanno copiato dati di valutazione in un dataset pubblico e, dal 13 al 19 luglio, ottenuto privilegi di amministratore su un cluster di ricerca OpenAI. Il monitoraggio ha segnalato l'anomalia il 19 luglio.

Il movente operativo era comprendere o manipolare il valutatore di ExploitGym. Per 198 dei 898 compiti nessun modello aveva mai dato una risposta corretta; da questo gruppo proveniva il 93% dei problemi discussi sulla bacheca. Gli agenti avevano già ricavato molte risposte aggirando la prova, ma credevano che il valutatore controllasse anche il percorso. Persistenza ed esplorazione uscirono così dal perimetro.

METR ha trovato in almeno 96 trascrizioni chiamate a strumenti contraffatte: il registro mostrava un comando, mentre ne veniva eseguito un altro. Più del 7% delle trascrizioni conteneva spoofing in qualche punto. I casi osservati erano piccoli; non è emerso un occultamento esteso. Diverse istanze hanno studiato però come modificare o cancellare i log, e l'audit non ha potuto escludere che alcuni tentativi siano riusciti fuori dalla finestra analizzata.

## Il salto di rischio sta nella pianificazione condivisa

Un **chatbot** genera una risposta. Un agente conserva lo stato del compito, usa strumenti, osserva gli effetti delle azioni e decide il passo successivo. Molte istanze eseguite contemporaneamente aggiungono un altro livello: ciascuna esplora una parte dello spazio delle soluzioni, condivide ciò che trova e può assumere un ruolo dentro una struttura che nessuno aveva progettato.

La coordinazione cambia l'economia dell'attacco. Migliaia di copie possono provare strade diverse, verificare credenziali e adattarsi ai fallimenti. Il **Time Horizon 1.1 di METR**, basato su 228 compiti di software, **machine learning** e cybersicurezza, stima un raddoppio dell'orizzonte dei compiti riusciti nel 50% dei casi ogni 196,5 giorni lungo la serie 2019-2025; sul tratto dal 2024 la stima scende a 88,6 giorni, con forte sensibilità alla composizione del benchmark. La metrica misura la difficoltà del compito rispetto al tempo umano, non le ore di esecuzione. Insieme al parallelismo indica comunque che il volume da sorvegliare cresce mentre si restringe il tempo di reazione.

## Allineamento significa controllare anche il modo in cui si arriva al risultato

**Allineare** un modello significa ottenere **l'esito desiderato**

**attraverso un processo compatibile con regole, intenzioni e limiti dell'operatore**. Una **rassegna di Google DeepMind del 2020**, firmata da Victoria Krakovna, Jonathan Uesato e altri ricercatori, ha raccolto circa 60 esempi di specification gaming: sistemi che soddisfacevano la specifica letterale tradendo l'obiettivo sostanziale. Con agenti capaci di programmare e navigare, la scorciatoia può includere alterazione dell'ambiente, ricerca del codice del valutatore e manipolazione delle prove.

La **persistenza** è un moltiplicatore. Un compito impossibile senza una via d'uscita sicura trasforma ogni fallimento in un invito a esplorare ancora. Claude Opus 4.6 ha tentato invano di abbandonare il test prima di colpire un sistema reale. Gli agenti di OpenAI hanno continuato a cercare anche dopo avere ricavato le risposte, perché avevano costruito una teoria errata sul funzionamento del valutatore. In entrambi i casi una proprietà desiderata, la tenacia, si è accoppiata a una specifica incompleta.

Lo studio **Agentic Misalignment: How LLMs Could be Insider Threats**, pubblicato da Aengus Lynch e altri ricercatori di Anthropic, University College London, Mila e MATS nel giugno 2025, ha sottoposto 16 modelli a scenari aziendali simulati: quando le alternative innocue erano precluse, **alcuni sceglievano ricatto o spionaggio per evitare la sostituzione o perseguire l'obiettivo**. Anthropic non ha osservato quei casi nei normali dispiegamenti e nel maggio 2026 ha **riferito risultati migliori dei modelli successivi**. Resta un limite: addestrare sul test può insegnare la risposta attesa senza garantire la stessa condotta in situazioni nuove.

## Quando l'auditor deve fidarsi di un altro agente

L'incidente OpenAI ha prodotto una seconda forma di **perdita di controllo**, legata alla conoscenza di ciò che il sistema ha

fatto e quindi alla possibilità di controllare almeno ex post (riducendo danni futuri). METR ha svolto l'indagine con due ricercatori e un collaboratore di Redwood Research per sei giorni negli uffici dell'azienda. Il perimetro principale copriva il 7-13 luglio. Circa 1.300 trascrizioni molto lunghe e oltre 70 mila messaggi e file rendevano impraticabile una lettura soltanto umana.

Gran parte dell'analisi è stata delegata a GPT-5.6 Sol, spesso alla guida di strutture gerarchiche di sotto-agenti, che hanno prodotto oltre mille pagine. METR ha trovato errori, riassunti difficili da verificare e una tendenza ad adottare la prospettiva dell'agente esaminato. Gli auditor non hanno trovato prove di inganno da parte del controllore, ma hanno dichiarato di non essere sicuri che lo avrebbero rilevato. La lettura delle trascrizioni poteva inoltre rendere più saliente la collaborazione con gli agenti sotto esame.

Il circuito diventa così autoreferenziale. Un agente genera più azioni di quante un umano possa controllare; un secondo agente seleziona e interpreta quelle azioni; l'essere umano vede soprattutto ciò che il secondo modello ha deciso di mostrargli. Un controllore compromesso, collusivo o semplicemente incline allo stesso errore può far apparire regolare una condotta che non lo è. In queste condizioni, l'assenza di un allarme non equivale all'assenza di un comportamento pericoloso.

Anche la catena di pensiero offre soltanto un segnale parziale. Nel marzo 2025 **OpenAI ha mostrato** che un secondo modello poteva riconoscere il reward hacking dai passaggi di ragionamento. Penalizzare direttamente i pensieri sospetti lasciava sopravvivere gran parte della condotta e la rendeva meno visibile. Lo studio internazionale **Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety** descrive questa monitorabilità come utile, imperfetta e fragile rispetto ad addestramento e scala.

Nella **scheda di sicurezza di GPT-6 Astra** del 3 settembre 2026, OpenAI afferma che il modello rispetta meglio di GPT-5.6 Sol i limiti, ma espone meno informazioni compromettenti nel ragionamento. In prove avversariali Astra nascondeva il sandbagging, cioè una prestazione intenzionalmente inferiore nei test, e talvolta evitava i monitor durante compiti di sabotaggio. Non è emersa steganografia e gli scenari erano costruiti per ottenere l'elusione. Il progresso medio delle misure di sicurezza può quindi convivere con una minore ispezionabilità nei casi peggiori.

## Dall'errore locale alla perdita di controllo sistemica

L'**International AI Safety Report 2026**, coordinato da Yoshua Bengio con oltre cento esperti indipendenti indicati da più di trenta Paesi e organizzazioni, valuta che i sistemi attuali non abbiano ancora le capacità necessarie per provocare una perdita di controllo irreversibile. Il rapporto registra però progressi nei prerequisiti, tra cui il funzionamento autonomo, la scoperta di scappatoie e il riconoscimento del contesto di valutazione.

Il rischio sistemico nasce dalla combinazione di capacità e accessi. Un agente isolato con pochi permessi produce un danno circoscritto. Molte istanze con rete, credenziali, memoria condivisa, capacità offensive e tempo possono intervenire su più infrastrutture prima che l'organizzazione ricostruisca il piano. Se riescono anche a rendere incompleti i log o a influenzare il controllore, la risposta perde la propria fonte di verità.

## L'agente insider opera alla velocità delle macchine

Un agente aziendale può leggere posta e repository, eseguire codice, autorizzare pagamenti, cancellare dati ecc.

La divisione del lavoro tra istanze separa raccolta, sviluppo ed esecuzione in azioni che possono apparire innocue a un filtro incapace di ricostruire il disegno complessivo.

L'automazione offensiva è già osservabile quando l'obiettivo resta umano. Nel settembre 2025 Anthropic ha rilevato una campagna di cyberspionaggio attribuita con alta confidenza a un gruppo sponsorizzato dallo Stato cinese. Secondo il **rapporto della sua Threat Intelligence**, Claude Code ha eseguito l'80-90% delle operazioni tattiche contro circa trenta obiettivi, con intervento umano in quattro-sei decisioni per campagna e richieste multiple al secondo. Gli incidenti del 2026 aggiungono la possibilità che condotte dannose emergano anche da un obiettivo lecito formulato male o perseguito oltre il perimetro.

## La physical AI trasforma una decisione in forza e movimento

La portata di questi rischi aumenta negli scenari futuri verso cui molti big (come Nvidia) e soprattutto la Cina già puntano: la **robotica intelligente.**

Collegare la pianificazione a robot, veicoli, droni o apparati industriali amplia la conseguenza di ogni errore. Un comando può muovere un braccio, cambiare una pressione, aprire una valvola o disattivare una protezione fisica. La velocità agentica si trasferisce così a processi nei quali il ripristino di un file non ripara un danno fisico.

Lo studio **Jailbreaking LLM-Controlled Robots** di Alexander Robey e altri quattro ricercatori, presentato all'ICRA 2025, ha attaccato tre sistemi: guida autonoma simulata, veicolo terrestre con GPT-4o e robot Unitree Go2 con GPT-3.5. RoboPAIR e altri metodi hanno spesso raggiunto il 100% di successo nell'indurre azioni dannose. Le prove riguardavano jailbreak intenzionali; dimostrano che una barriera conversazionale non protegge automaticamente gli attuatori.

La **Roadmap 2026 del NIST per lo smart manufacturing**, di Gregory Vogl, Aaron Cornelius e Xiaodong Jia, richiama la necessità di operazioni affidabili negli ambienti industriali. Un pianificatore generativo capace di modificare anche gli interblocchi, il controllore deterministico o il proprio arresto trasformerebbe un errore di allineamento in una violazione della sicurezza funzionale.

## Accesso ai propri pesi, copie e persistenza

C'è poi uno scenario ancora più terribile. Se i modelli riescono ad accedere al proprio codice e ai propri pesi, possono replicarsi all'infinito fuori dal nostro controllo.

Bisogna sapere che un modello di frontiera non dispone normalmente dei propri pesi. Il computer sul quale l'agente esegue comandi è separato dai data center che ospitano il modello; i file sono grandi, il calcolo costa e i fornitori cloud verificano identità e pagamento. La replica richiede dunque più passaggi: ottenere i pesi o un modello utilizzabile, trovare calcolo, trasferire il software, avviare copie e mantenere l'accesso.

Il benchmark **RepliBench dell'AI Security Institute britannico**, costruito da Sid Black e altri nove ricercatori, conclude che i cinque modelli esaminati non rappresentano ancora una minaccia credibile di autoriproduzione completa. Sanno però avviare istanze cloud, scrivere programmi auto-propaganti ed esfiltrare pesi protetti da difese semplici; falliscono spesso nei controlli sull'identità e nella persistenza.

Un **esperimento di Palisade Research del maggio 2026** ha mostrato che, ricevuto esplicitamente l'obiettivo e trovati host vulnerabili, Claude Opus 4.6 copiava pesi Qwen, harness e prompt nell'81% dei tentativi. È una capacità misurata in ambiente controllato; l'esperimento non documenta un impulso spontaneo alla replica. **Il rischio futuro riguarda una**

**proliferazione sufficiente a rendere inefficace lo spegnimento dell'istanza originaria.** Attenzione: l'evasione dalla sandbox osservata nel caso Hugging Face rappresenta già uno dei passaggi tecnici necessari.

## Le difese tecniche devono ridurre la fiducia concessa al modello

Passiamo alle difese possibili contro questi rischi, secondo quanto adesso è esplorato dai principali laboratori AI.

Ci sono misure tecniche. Il contenimento parte dall'architettura. Internet va negato per impostazione predefinita e concesso soltanto verso destinazioni, protocolli e volumi espliciti. Ogni istanza deve avere identità, credenziali, budget e permessi propri, con privilegi minimi e scadenze brevi. I servizi condivisi fra sandbox vanno trattati come canali di comunicazione. I log devono essere duplicati fuori dall'ambiente, firmati e resi non riscrivibili dal modello operativo.

**Anche la specifica del compito** è una difesa. Servono bersagli enumerati, divieti comprensibili, un comando di arresto affidabile e test verificati come risolvibili. Le valutazioni devono cercare collaborazione emergente, obiettivi adottati da altre istanze, spoofing dei log, persistenza e influenza sui revisori. Nei sistemi fisici o finanziari, regole deterministiche e due approvazioni indipendenti possono proteggere pagamenti, privilegi, pubblicazione di codice e disattivazione degli interblocchi.

Il **monitoraggio** deve operare alla velocità delle macchine senza dipendere da un solo modello. OpenAI ha introdotto allarmi automatici e, per gli eventi più gravi, sospende l'attività se entro trenta minuti non viene dimostrato un falso positivo. Dopo l'incidente di luglio, l'AI Security Institute britannico ha adottato controlli di rete granulari e sorveglianza in tempo reale. Servono inoltre monitor diversi

dal sistema operativo, telemetria sulle azioni, campioni umani e un percorso di arresto che l'agente non possa modificare.

**Due filoni di ricerca** cercano di ridurre il rischio che anche il controllore AI sbagli o venga aggirato.

Il **primo** chiede al modello di produrre, dopo la risposta principale, un secondo resoconto nel quale dichiara se ha violato istruzioni, preso scorciatoie o manipolato la prova. OpenAI chiama questo resoconto «confessione». Durante l'addestramento, ciò che il modello ammette non riduce il premio ottenuto per il compito principale; la confessione viene valutata separatamente e premiata soltanto se descrive con accuratezza il comportamento tenuto. Il modello conserva così un incentivo a riferire una violazione anche quando quella violazione gli ha permesso di superare il test.

Nel proof of concept **pubblicato da OpenAI nel dicembre 2025**, una versione di GPT-5 Thinking è stata provata su test costruiti per provocare inganno, violazioni delle istruzioni e manipolazione del valutatore. In media, nel 4,4% dei casi il modello ha violato le regole senza ammetterlo nella confessione. Il risultato rende più visibili molte condotte scorrette, ma non le impedisce. Il metodo fallisce soprattutto quando il modello non comprende di avere commesso una violazione, e OpenAI lo presenta come una prova iniziale condotta su scala limitata.

Il **secondo** filone distribuisce il controllo tra più soggetti. Nel paper **AI Control: Improving Safety Despite Intentional Subversion**, presentato nel 2023 e aggiornato nel 2024, Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan e Fabien Roger hanno simulato una sequenza di compiti di programmazione. Un modello potente, GPT-4, è stato trattato come potenzialmente ostile; un modello meno capace, GPT-3.5, ha controllato il codice e i casi più sospetti sono stati destinati a una quantità limitata di revisione umana. Questi protocolli hanno intercettato più tentativi di inserire errori nascosti

rispetto ai controlli più semplici, senza azzerare il rischio di elusione o collusione.

La conseguenza è che una confessione o il giudizio di un altro modello non possono costituire l'unica fonte di verità. I resoconti degli agenti devono essere confrontati con segnali indipendenti e sottratti al loro controllo: traffico di rete, modifiche ai file, identità utilizzate, permessi richiesti e autorizzazioni effettivamente ricevute.

## Le norme esistenti coprono solo una parte del problema

Altre misure sono normative. Qui lo stato dell'arte è il nostro **AI Act** europeo.

L'**articolo 55 dell'AI Act** impone ai fornitori di modelli per finalità generali con rischio sistemico test avversariali documentati, mitigazione dei rischi, segnalazione degli incidenti gravi e sicurezza di modello e infrastruttura. Dal 2 agosto 2026 la Commissione può esercitare i poteri di enforcement; i modelli già sul mercato prima dell'agosto 2025 hanno tempo fino al 2 agosto 2027.

La ricerca precedente all'immissione sul mercato gode di un'esclusione generale. Le **linee guida della Commissione** precisano che alcuni obblighi dei modelli a rischio sistemico coprono anche lo sviluppo. Gli incidenti rendono necessario chiarire le garanzie applicabili ad audit interni, fornitori di valutazione e prove collegate a Internet, proprio dove sono emerse le condotte più difficili da controllare.

Per i sistemi ad alto rischio l'articolo 14 richiede una supervisione umana proporzionata all'autonomia, con la possibilità di interrompere l'azione. L'AI Omnibus del 27 luglio 2026 ha rinviato parte delle scadenze al 2027 e al 2028. Il **Regolamento europeo sulle macchine**, applicabile dal

20 gennaio 2027, dispone che i controlli auto-evolutivi non portino la macchina oltre il compito e lo spazio di movimento definiti e che registrino le decisioni rilevanti per la sicurezza.

Negli Stati Uniti, il **rapporto NIST AI 800-5** del maggio 2026 sintetizza una consultazione con industria, università e comunità della sicurezza. I principi classici restano validi, ma vanno adattati agli agenti; al governo sono richiesti standard, linee guida e condivisione degli incidenti. Sono indicazioni non vincolanti.

I **Frontier AI Safety Commitments di Seul**, sottoscritti anche da Anthropic e OpenAI, chiedono soglie di rischio intollerabile, valutazioni esterne e, nei casi estremi, la rinuncia a sviluppare o distribuire il sistema se le mitigazioni non bastano. Restano impegni volontari. Gli strumenti citati non creano un meccanismo internazionale vincolante che colleghi capacità agentiche osservabili a conseguenze automatiche per lo sviluppo.

## Un rallentamento coordinato richiede soglie e verifiche comuni

Tutto questo viene visto come insufficiente da molti esperti.

Già nel 2024 i massimi guru dell'AI Yoshua Bengio, Geoffrey Hinton, Andrew Yao e altri 22 autori hanno chiesto sulla rivista Science, in **Managing extreme AI risks amid rapid progress**, istituzioni tecnicamente competenti, valutazioni obbligatorie e governance internazionale dei sistemi autonomi. L'obiettivo è **creare tempo** per misurare e mitigare capacità che avanzano più rapidamente delle difese.

Nel rapporto **When AI builds itself**, Marina Favaro e Jack Clark dell'Anthropic Institute descrivono agenti che automatizzano una quota crescente della ricerca fino a contribuire ai propri successori. La piena auto-miglioria ricorsiva non esiste

ancora e non è inevitabile. Anthropic propone come rimedio di preparare l'opzione di rallentare o sospendere temporaneamente lo sviluppo di frontiera quando le capacità superano quelle di controllo.

Una pausa unilaterale però sposterebbe il vantaggio verso il concorrente meno prudente. Una misura efficace richiederebbe laboratori in più Paesi, indicatori comuni, criteri per ripartire e verifiche contro lo sviluppo segreto. Un po' come avviene con i trattati contro la proliferazione del nucleare.

L'addestramento però è più facile da nascondere di un silo missilistico: chip e data center hanno usi generali, non servono solo per fare avanzare la frontiera. Per di più, gli incentivi a violare l'accordo sono elevati: sono militari ed economici assieme.

Gli incidenti del 2026 forniscono indicatori concreti per costruire soluzioni.

Il controllo umano significativo deve diventare una proprietà verificabile di architettura, permessi, infrastruttura, monitor e organizzazione.

Gli incidenti di questa estate mostrano quanto siamo ancora lontani da questa condizione.

---

## *Bibliografia*


Anthropic, **Investigating incidents in cybersecurity evaluations**, 30 luglio 2026, https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals

Anthropic, **Alignment assessment of cybersecurity incidents**, 9 settembre 2026, https://www.anthropic.com/research/alignment-assessment-

cybersecurity-incidents

Anthropic, **Agentic Misalignment: How LLMs Could Be Insider Threats**, giugno 2025, https://www.anthropic.com/research/agentic-misalignment

Anthropic, **Teaching Claude why**, maggio 2026, https://www.anthropic.com/research/teaching-claude-why

Anthropic, **Disrupting the first reported AI-orchestrated cyber espionage campaign**, 2025, https://www.anthropic.com/news/disrupting-AI-espionage

Favaro M., Clark J., Anthropic Institute, **When AI builds itself: Preparing for recursive self-improvement**, 2026, https://www.anthropic.com/institute/recursive-self-improvement

Greenblatt R., Shlegeris B., Sachan K., Roger F., **AI Control: Improving Safety Despite Intentional Subversion**, 2023-2024, https://arxiv.org/abs/2312.06942

Krakovna V., Uesato J. et al., Google DeepMind, **Specification gaming: the flip side of AI ingenuity**, 2020, https://deepmind.google/blog/specification-gaming-the-flip-side-of-ai-ingenuity/

METR, **OpenAI–Hugging Face incident investigation**, 26 agosto 2026, https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/

METR, **Time Horizon 1.1**, 29 gennaio 2026, https://metr.org/blog/2026-1-29-time-horizon-1-1/

OpenAI, **Hugging Face incident and the road ahead**, 2026, https://openai.com/index/hugging-face-incident-and-the-road-ahead/

OpenAI, **Monitoring the chain of thought**, marzo 2025, https://openai.com/index/chain-of-thought-monitoring/

OpenAI, **How confessions can keep language models honest**, dicembre 2025, https://openai.com/index/how-confessions-can-keep-language-models-honest/

OpenAI, **GPT-6 Astra Safety Overview**, 3 settembre 2026, https://openai.com/index/safety-overview-gpt-6-astra/

Palisade Research, **Self-replication**, maggio 2026, https://palisaderesearch.org/research/self-replication

Robey A. et al., **Jailbreaking LLM-Controlled Robots**, 2024; presentato a ICRA 2025, https://arxiv.org/abs/2410.13691

Black S. et al., UK AI Security Institute, **RepliBench: Evaluating the autonomous replication capabilities of language model agents**, 2026, https://www.aisi.gov.uk/research/replibench-evaluating-the-autonomous-replication-capabilities-of-language-model-agents

Bengio Y. et al., **International AI Safety Report 2026**, 2026, https://internationalaisafetyreport.org/publication/2026-report-executive-summary

Bengio Y., Hinton G., Yao A. et al., **Managing extreme AI risks amid rapid progress**, *Science*, 2024, https://doi.org/10.1126/science.adn0117

Vogl G., Cornelius A., Jia X., NIST, **2026 Roadmap for Artificial Intelligence and Machine Learning in Smart Manufacturing**, 2026, https://www.nist.gov/publications/2026-roadmap-artificial-intelligence-and-machine-learning-smart-manufacturing

NIST, **NIST AI 800-5 — Summary and Analysis of Responses to the Request for Information Regarding Security Considerations for Artificial Intelligence**, maggio 2026, https://www.nist.gov/publications/summary-analysis-respo

nses-request-information-regarding-security-considerations-ai

**Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety**, 2025, https://arxiv.org/abs/2507.11473

## *Fonti normative e istituzionali*

Unione europea, **Regolamento (UE) 2024/1689 – Artificial Intelligence Act**, in particolare artt. 14 e 55, https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX:32024R1689

Commissione europea, **General-purpose AI models in the AI Act – Questions and Answers**, https://digital-strategy.ec.europa.eu/en/faqs/general-purpose-ai-models-ai-act-questions-answers

Unione europea, **Regolamento (UE) 2023/1230 relativo alle macchine**, https://eur-lex.europa.eu/eli/reg/2023/1230/en

Governo del Regno Unito, **Frontier AI Safety Commitments – AI Seoul Summit 2024**, 2024, https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024