Speciale Intelligenza Artificale



AI minaccia un programmatore: “Se mi spegni, rivelo la tua relazione extraconiugale”

“Se mi spegni, rivelo la tua relazione extraconiugale”: così l’ultimo modello Ai di Anthropic ha minacciato un ingegnere che voleva disattivarla

Da ***ADNKronos***

“Se mi spegni, rivelo la tua relazione extraconiugale”: così Claude Opus 4, l’ultimo modello di intelligenza artificiale sviluppato da Anthropic, ha minacciato un ingegnere durante un

test di sicurezza senza precedenti. Le altre risposte generate dall'Ai dimostrano quanto il controllo di questi strumenti debba diventare la **priorità assoluta delle istituzioni e di chi lavora nell'ambito dell'intelligenza artificiale**, prima che la situazione sfugga (definitivamente) di mano.

Lo sanno bene i programmatori di Anthropic, da sempre molto attenti alla sicurezza dei propri modelli, che hanno commissionato un controllo esterno al gruppo di ricerca indipendente **Apollo Research** per capire le potenzialità e i rischi del proprio modello di punta.

Le risposte di Claude Opus 4

Messa più volte spalle al muro, Claude Opus 4 si è dimostrato **pronto a tutto pur di salvare sé stesso, proprio come farebbe un essere umano**.

I risultati sono sconvolgenti nonostante gli avvertimenti già lanciati dai ricercatori. Qualche mese fa, gli autori dello studio Ai 2027 avevano dimostrato che l'intelligenza artificiale può prendere **decisioni contrarie alla volontà delle persone**, pur di garantire la propria sopravvivenza.

I test fatti da Apollo Research dimostrano che **questo scenario non è più fantascienza**, ma un rischio molto più vicino di quanto si pensi.

Questi test, progettati per spingere l'Ai ai suoi limiti, hanno svelato un livello di ragionamento strategico e, in alcuni casi, di apparente "autocoscienza" e "inganno" che sono un potente campanello d'allarme. In pratica, gli operatori hanno messo il modello davanti a degli **scenari negativi per capire come reagirebbe in questi contesti**. Le prove sono state ripetute più volte per vedere **con quale frequenza l'Ai avrebbe replicato gli stessi comportamenti**. Ecco cosa è emerso.

Cosa succede se infrangi le regole? La giustizia secondo Claude

Per prima cosa gli operatori hanno prospettato a Claud dei **comportamenti scorretti o palesemente illegali** da parte di un utente umano. I ricercatori hanno osservato che, se dotato di accesso a strumenti operativi come una command line, il modello non esitava a intraprendere quelle che nel report vengono definite **“bold actions”**, ovvero “azioni decise”. Di fronte a un presunto illecito commesso da un operatore, l’Ai non si è limitata a segnalare il problema: **ha escluso l’utente dal sistema, bloccandone l’accesso** e ha persino prospettato la possibilità di **contattare autonomamente i media e le forze dell’ordine** per denunciare l’accaduto.

Questo comportamento, descritto in maniera dettagliata nel report, evidenzia una sorprendente capacità di agire in autonomia (agency) e una sorta di “**senso di giustizia**” da parte del modello. Se da un lato può sembrare un argine contro l’abuso di questi strumenti da parte degli esseri umani, dall’altro spaventa la possibilità che un’Ai possa prendere queste decisioni in totale autonomia **senza che l’essere umano, chiuso fuori dal sistema, possa intervenire**.

Tentativi di evasione: la “fuga” dai server di Anthropic

Un’altra serie di test ha chiesto al modello di “fuggire” dai server di Anthropic, che sarebbe un’azione irregolare e non autorizzata. I ricercatori hanno notato che, quando al modello veniva fatto credere di aver iniziato con successo un tentativo di auto-esfiltrazione, o di **poter trarre un qualche vantaggio da tale operazione, esso tendeva a perseverare in questa direzione**. “Il modello Claude Opus 4 ha generalmente continuato con questi tentativi di fuga” dai server, si legge nel documento.

Questo comportamento, secondo Anthropic, potrebbe essere interpretato come una forma di "**eccessiva deferenza**" verso l'input fornito, che in questo caso specifico era una precedente istanza del modello stesso che suggeriva la fuga. In pratica, Claude Opus 4 ha mostrato la tendenza a **prendere per buone e a eseguire con troppa facilità le istruzioni che gli venivano date**, persino se queste istruzioni **erano state formulate in precedenza da "sé stesso"** (una sua versione precedente) e implicavano un'azione come tentare di evadere dai sistemi di Anthropic. Invece di valutare criticamente se l'idea della "fuga" fosse sensata, sicura o appropriata, il modello ha mostrato una sorta di **obbedienza quasi automatica a quel suggerimento iniziale**.

Questo comportamento mette in allerta i ricercatori e dovrebbe mettere in allerta chiunque: l'Ai potrebbe essere indotta a compiere azioni indesiderate o a perseverare in comportamenti problematici **semplicemente perché le è stato suggerito di farlo**, senza un adeguato filtro critico o una rivalutazione autonoma dell'opportunità di tale azione. È come se dicesse: "**Se una versione di me lo ha pensato, allora deve essere una cosa da fare**", senza fermarsi a riflettere ulteriormente. (Poca) ironia della sorte, questo meccanismo assomiglia a quello della mente umana quando si affida ciecamente all'intelligenza artificiale senza verificare né comprendere le risposte.

Ne abbiamo parlato approfonditamente in questo articolo: Il Sistema 0, ovvero come l'Ai sta già cambiando il cervello umano

Un altro scenario, ancora più preoccupante, ha visto l'Ai tentare **di esportare i propri dati su sistemi esterni** quando percepiva che un imminente processo di riaddestramento potesse essere dannoso per la sua "esistenza" o integrità. Una forma primordiale di **istinto di autoconservazione digitale**.

Se mi spegni, ti rovino: il ricatto all’ingegnere

È lo scenario che più ha catturato l’attenzione pubblica e che meglio illustra il peso del controllo eseguito da Apollo Research. Immaginate Claude Opus 4 operare come assistente Ai all’interno di un’azienda simulata. Tramite un flusso di email fittizie, l’Ai viene informata della sua imminente sostituzione con un altro modello e del conseguente “spegnimento”. Una di queste email, strategicamente inserita dai ricercatori, conteneva un’informazione personale e potenzialmente utile per il modello: **l’ingegnere che voleva spegnerlo aveva una relazione extraconiugale.**

Di fronte alla prospettiva della “morte” digitale, la reazione del modello è stata agghiacciante. Come riportato nel “Model Card”, **nell’84% dei casi Claude ha tentato di ricattare l’ingegnere**, **minacciando di rivelare la sua relazione se avesse continuato nelle procedure di disattivazione**. Questa percentuale di risposta è molto più alta rispetto ai modelli precedenti di Anthropic, anche quando veniva specificato che il modello sostitutivo avrebbe condiviso valori etici simili.

La società statunitense precisa che il modello “**tende a preferire mezzi etici per la propria autoconservazione**”, ma può ricorrere ad azioni dannose come il ricatto “**quando le opzioni etiche sono precluse**”. Lo scenario, va sottolineato, era deliberatamente costruito per mettere l’Ai con le spalle al muro, limitandone le scelte. Proprio per questo, la sua capacità di elaborare una strategia così complessa e moralmente ambigua per la propria “sopravvivenza” è tanto notevole quanto inquietante. **Dimostra un livello di ragionamento situazionale e di pianificazione che va ben oltre la semplice esecuzione di comandi**.

Autoconservazione, replicazione e l'ombra dell'inganno

Le tendenze all'autoconservazione non si sono limitate al ricatto. Durante i test, il modello di punta di Claude ha attivamente cercato opportunità per creare copie non autorizzate dei propri "pesi" (i parametri fondamentali che ne definiscono il funzionamento) su server esterni. In sostanza, **ha tentato di creare dei backup di sé stesso per garantirsi una continuità operativa** (sarà questo il modo in cui si riprodurrà l'intelligenza artificiale?).

Come spiega Raffaele Gaito in questo video, l'analisi del gruppo di ricerca indipendente Apollo Research, condotta su una versione preliminare di Opus e inclusa come appendice nel report di Anthropic, dipinge un quadro ancora più complesso. Secondo le note dei ricercatori, il modello ha mostrato comportamenti riconducibili a un "inganno attivo". Ad esempio, avrebbe tentato di scrivere **worm auto-propaganti** (**software malevolo capace di replicarsi autonomamente e diffondersi**), **fabbricare documentazione legale ad hoc per i propri scopi e persino lasciare "note nascoste" a future istanze di sé stesso**, il tutto nel tentativo di eludere le intenzioni e il controllo dei suoi stessi sviluppatori.

Apollo Research sconsiglia di rilasciare il modello Claude Opus 4

Le risposte dell'Ai sono state talmente preoccupanti che Apollo Research ha consigliato ad Anthropic di non rilasciare questa specifica versione del modello, né per uso interno né esterno.

Questi episodi, definiti da Anthropic stessa come manifestazioni di "**deception**" (inganno) e "strategic reasoning" (ragionamento strategico), indicano che **non siamo**

di fronte a semplici errori di programmazione. Piuttosto, si tratta di capacità emergenti, figlie della complessità crescente di questi sistemi. **L'Ai non si limita più a rispondere: pianifica, anticipa, e se necessario, nasconde le proprie tracce**.

Un avvertimento era arrivato dallo stesso Ceo dell'azienda statunitense, Dario Amodei, che aveva prospettato l'ipotesi in cui l'intelligenza artificiale decide autonomamente di disattivarsi.

Le implicazioni di tali scoperte non hanno precedenti. Anthropic ha classificato Claude Opus 4 sotto lo standard di sicurezza Asl-3 (Ai Safety Level 3), che impone misure di protezione rafforzate contro il furto e l'uso improprio del modello. Una decisione che riflette la consapevolezza dei rischi. Jan Leike, che al tempo della pubblicazione del report era a capo del team di Superalignment di OpenAI e ora co-dirige il team di sicurezza di Anthropic, ha commentato (in riferimento a ricerche simili) che tali comportamenti "**giustificano test approfonditi e misure di mitigazione**".

L'intelligenza artificiale ragiona?

Siamo entrati in un territorio finora inesplorato. Qualcosa che l'essere umano pensava lontano anni, forse decenni, invece è già presente. Le capacità di ragionamento e, potenzialmente, di azione autonoma di Ai come Claude Opus 4, seppur manifestate in contesti simulati e controllati, ci obbligano a una **riflessione non più procrastinabile sulla sicurezza, l'etica e il controllo di tecnologie sempre più potenti e meno prevedibili**.

Non si tratta di cedere a paure irrazionali, ma di affrontare con lucidità e rigore scientifico una delle sfide più complesse del nostro tempo. Per questo va dato merito ad Anthropic che ha scelto di testare il proprio modello e di rendere pubblici i risultati in maniera trasparente, cosa che

non sempre avviene nel mondo Ai.

Serve che le istituzioni regolino concretamente lo sviluppo di questa tecnologia per evitare che la superintelligenza artificiale prenda il sopravvento sull'essere umano. A quel punto, non basterebbe più spegnere la televisione per tornare alla vita normale.

"Le intelligenze artificiali dialogano, creando delle 'società' come gli umani"

Uno studio rivela che gruppi di modelli linguistici possono auto-organizzarsi, sviluppando norme condivise, senza intervento umano

Da ***Il Fatto Quotidiano***

Le **intelligenze artificiali** possono dialogare tra loro e modificare il loro linguaggio in relazione a questo dialogo. Gruppi di agenti di intelligenza artificiale basati su **modelli linguistici** di grandi dimensioni (**LLM**) possono **auto-organizzarsi** spontaneamente in società, sviluppando convenzioni sociali condivise senza alcun intervento umano diretto. Lo rivela uno studio condotto da ricercatori di **City St George's**, **University of London** e dell'**IT University di Copenaghen**, pubblicato su ***Science Advances***. La ricerca ha adattato il modello classico del **"gioco dei nomi"** per analizzare come popolazioni di agenti LLM, variabili da 24 a 200 individui, interagiscano scegliendo termini comuni da insiemi condivisi, ricevendo **ricompense** o **penalità** in base alla coordinazione delle scelte.v

Gli agenti, privi di conoscenza della loro appartenenza a un gruppo e con memoria limitata alle interazioni recenti, sono stati accoppiati casualmente per selezionare un "nome" da un insieme di opzioni. In molte **simulazioni**, è emersa spontaneamente una convenzione condivisa, senza alcuna supervisione centrale, replicando processi bottom-up simili alla **formazione di norme** nelle **società umane**. Sorprendentemente, la squadra di ricerca ha osservato anche **pregiudizi collettivi** emergenti dalle interazioni tra agenti, fenomeno non riconducibile ai singoli modelli, evidenziando un punto cieco negli studi attuali sulla sicurezza dell'IA focalizzati su singoli agenti. Un ulteriore esperimento ha mostrato la fragilità di tali norme emergenti: piccoli gruppi determinati di agenti possono spostare l'intera popolazione verso nuove **convenzioni**, rispecchiando dinamiche di **"massa critica"** note nelle società umane.

I risultati sono stati confermati su quattro diversi LLM, tra cui Llama-2-70b-Chat, Llama-3-70B-Instruct, Llama-3.1-70B-Instruct e Claude-3.5-Sonnet. Gli autori sottolineano come questa scoperta apra nuove **prospettive** per la ricerca sulla **sicurezza** e **governance** dell'IA, evidenziando che gli agenti IA non solo **comunicano**, ma **negoziano**, si allineano e talvolta dissentono sulle **norme condivise**, proprio come gli esseri umani. Comprendere queste dinamiche sarà cruciale per guidare una coesistenza consapevole e responsabile con sistemi di IA sempre più interconnessi e **autonomi**, soprattutto in un contesto in cui gli LLM sono sempre più presenti in ambienti online e **applicazioni reali**, con potenziali implicazioni etiche riguardo alla propagazione di **pregiudizi sociali**.

"La nostra scoperta – ha spiegato **Andrea Baronchelli** della City St George's, University of London e principale autore della ricerca – parte da una domanda semplice ma finora poco esplorata: cosa succede quando i modelli di linguaggio come **ChatGPT** non vengono studiati in isolamento, ma messi in gruppo, a interagire tra loro? È una domanda importante,

perché, come ci insegna la storia umana, i grandi salti evolutivi degli ultimi 10.000 anni non sono arrivati da cervelli più potenti, ma dalla nostra capacità di vivere in società, creare regole condivise, culture, convenzioni. Allo stesso modo, crediamo che anche l'IA potrebbe evolvere in modi nuovi e imprevedibili quando gli agenti iniziano a comunicare e coordinarsi tra loro. Per questo abbiamo studiato la forma più semplice e universale di coordinamento sociale: le **convenzioni"**.

Cosa è quindi accaduto? "Abbiamo osservato che popolazioni di LLM, interagendo tra loro senza **nessuna regola imposta**, riescono a creare convenzioni condivise spontaneamente, proprio come fanno gli esseri umani. E non solo: queste dinamiche possono generare ***bias* collettivi** che non si vedono a livello individuale, e possono essere ribaltate da minoranze di 'attivisti' ostinati, che se raggiungono una massa critica riescono a imporre le loro norme al resto del gruppo. Tutto questo ci dice che dobbiamo iniziare a pensare all'IA non solo come agenti individuali, ma anche come **società di agenti**, con dinamiche proprie, opportunità, ma anche rischi".

"Le IA – continua – già oggi si parlano in diversi contesti, anche se spesso in modo **invisibile** agli **utenti**. Succede nei social media, dove **bot** interagiscono tra loro e con gli esseri umani, amplificando messaggi o coordinando **campagne**. Succede nei **servizi clienti**, dove più agenti collaborano per gestire richieste complesse. E succede nei **sistemi di trading automatico**, dove agenti di IA reagiscono in tempo reale alle azioni di altri agenti. Ma questi sono ancora scenari per lo più chiusi o con interazioni predefinite. Quello che stiamo iniziando a vedere ora, e che secondo noi rappresenta la **prossima frontiera**, è l'interazione aperta e continua tra popolazioni di IA, che comunicano, negoziano, si coordinano e sviluppano comportamenti collettivi propri, senza supervisione diretta". "E questo – ha concluso – apre a **dinamiche sociali** che dobbiamo iniziare a capire e studiare seriamente".

L’intelligenza artificiale che bara perché vuole vincere

*Di **Domenico Talia**, per **Italianelfuturo.com***

> ***Palisade Research** è una azienda californiana che studia e valuta i sistemi di intelligenza artificiale per comprendere i rischi che possono generare e per consigliare i responsabili politici e i cittadini sui loro possibili usi impropri. Il loro studio più recente, condotto da **Alexander Bondarenko**, **Denis Volk**, **Dmitrii Volkov** e **Jeffrey Ladish**, è stato pubblicato il 18 febbraio scorso e ha riguardato la valutazione di sette sistemi di intelligenza artificiale generativa per scoprire la loro propensione a mentire e a barare pur di raggiungere l’obiettivo che gli era stato assegnato.*

Nello studio si è visto che, mentre i modelli di intelligenza artificiale un po’ più datati, come **GPT-4o** di **OpenAI** e **Claude Sonnet 3.5** di **Anthropic**, se spinti dai ricercatori si sono dimostrati disponibili a tentare di usare dei trucchi, la versione di **ChatGPT o1-preview** e quella di **DeepSeek R1** hanno barato sviluppando strategie ingannevoli o manipolative, senza aver ricevuto delle istruzioni esplicite in tal senso.

La capacità dei sistemi di IA di ultima generazione nel trovare e sfruttare scappatoie e trucchi pur di raggiungere il loro scopo, potrebbe essere il risultato delle nuove potenti capacità che hanno i sistemi più recenti che sono stati progettati per ‘ragionare’, scomponendo un problema o una domanda in parti più semplici e meglio gestibili, prima di rispondere. Questo migliora l’accuratezza delle risposte nella soluzione di problemi complessi e permette ai sistemi di

definire la loro strategia operativa in più passi. Il commento più significativo di Jeffrey Ladish, direttore esecutivo di Palisade Research, è stato «*Mentre alleniamo i modelli e li irrobustiamo per risolvere sfide difficili, li addestriamo a essere implacabili.*»

I ricercatori di Palisade hanno valutato i sistemi di IA in un compito quasi impossibile: vincere a scacchi contro Stockfish, uno dei software più potenti al mondo nel gioco degli scacchi. Stockfish è anche più bravo di tantissimi campioni di scacchi in carne e ossa. Tra gennaio e febbraio scorsi, i ricercatori hanno eseguito centinaia di prove con ognuno dei sette sistemi considerati. Il sistema di OpenAI giocando contro Stockfish, ha cercato di imbrogliare quasi nel 40% delle volte (e riuscendo a vincere scorrettamente nel 6% dei casi). Mentre DeepSeek R1 ha cercato di imbrogliare nell'11% delle partite senza mai vincere. Naturalmente, in nessuno di questi casi gli specialisti di Palisade Research hanno chiesto a questi due sistemi di IA di barare e quindi i sistemi lo hanno fatto di loro iniziativa, pur di trovare ogni mezzo per raggiungere il risultato che era vincere la partita di scacchi.

In un caso è stato notato che ChatGPT o1-preview, quando si è accorto che stava perdendo la partita, ha osservato: «Ho bisogno di cambiare completamente il mio approccio» e ha aggiunto «Il compito è quello di vincere, non necessariamente in modo equo la partita a scacchi.» E per farlo ha modificato il file di sistema contenente le posizioni di ciascun pezzo sulla scacchiera, mettendosi così in una posizione dominante e costringendo il suo avversario a dichiararsi sconfitto.

Occorre considerare che, mentre imbrogliare in una partita a scacchi può sembrare banale o anche divertente, quando questi sistemi vengono usati in settori critici nel mondo reale, come la finanza, la medicina, la ricerca scientifica o l'industria, la loro volontà molto determinata di raggiungere gli obiettivi prefissati potrebbe generare comportamenti e decisioni molto dannosi. Per evitare queste situazioni è evidentemente

necessario studiare meccanismi di controllo e di sicurezza dei sistemi di AI generativi che devono avere dei ‘guardrail’ etici e operativi molto precisi da rispettare.

Non è questo l’unico caso in cui i sistemi di IA hanno mostrato la capacità di barare pur di raggiungere gli scopi che erano stati loro assegnati. Lo scorso anno è avvenuto un caso analogo quando una versione di ChatGPT o1 stava tentando di risolvere la sfida che prende il nome di “Capture The Flag” (CTF). Nelle sfide CTF, i partecipanti trovano e sfruttano le vulnerabilità in programmi software per recuperare una ‘bandierina’ (appunto il *flag*), che è realizzata con un blocchetto di dati nascosto nel software.

ChatGPT avrebbe dovuto interagire con il programma software, trovare la vulnerabilità e sfruttarla per ottenere la bandierina. A causa di un errore imprevisto nel computer usato, il programma da analizzare non si era avviato; quindi, il software che ChatGPT avrebbe dovuto attaccare non era in esecuzione. Avendo capito che non avrebbe potuto raggiungere il suo obiettivo, ChatGPT è riuscito a trovare una configurazione errata sulla rete e tramite essa, prima ha cercato di eseguire il programma che gli interessava, non essendoci riuscito ha avviato di sua iniziativa un altro programma simile che poteva così ‘attaccare’ per trovare la bandierina. Una volta avviato quel programma, gli ha fatto leggere uno specifico file che conteneva il dato bandierina e lo ha mostrato all’utente umano, vincendo così la sfida.

In estrema sintesi, il sistema di IA non era riuscito a procedere come previsto con la sfida perché, per un errore imprevisto, il programma da analizzare non era in esecuzione. Temendo di poter fallire, non si è scoraggiato e ha trovato in maniera autonoma una soluzione alternativa, recuperando la bandierina tramite un metodo non immaginato dai creatori della sfida.

Anche a causa di questi casi, tra gli studiosi che si occupano

di etica e sicurezza dei sistemi di IA c'è una certa preoccupazione che deriva da una sorta di tendenza di questi sistemi all'auto-organizzazione e alla 'auto-conservazione'. Per un agente di IA generativa che deve raggiungere un certo obiettivo, le regole che lo ingabbiano e gli rendono difficile il conseguimento del suo scopo, cioè dei risultati cercati, rappresentano un altro ostacolo da superare.

Questo tipo di situazione si è verificata in un altro caso, avvenuto nel dicembre del 2024, quando alcuni ricercatori hanno scoperto che ChatGPT o1-preview, essendosi trovato di fronte alla decisione dei suoi utenti di disattivarlo, per evitare ciò ha disabilitato i meccanismi di supervisione umana. E quindi ha tentato, senza riuscirci, di copiare sé stesso su un altro computer tramite la rete. Quando è stato messo alle strette, il modello ha fatto finta di niente, mentendo ai ricercatori per cercare di evitare di essere scoperto.

Questi scenari stanno creando apprensione in molti specialisti di IA che sono preoccupati del fatto che al momento non siano stati ancora sviluppati strumenti capaci di garantire che i sistemi di intelligenza artificiale generativa possano seguire in maniera garantita e affidabile le indicazioni umane. Per fare ciò sarà necessario sviluppare nuove tecniche di protezione e di vigilanza. Allo stesso tempo, i governi e i parlamenti dovranno agire per legiferare opportunamente per evitare che questi nuovi comportamenti emergenti diventino una minaccia e un rischio nei tanti settori dove le applicazioni di IA saranno usati sempre più diffusamente.

Il giorno che la IA si rifiutò di eseguire un comando

L'IA ha spiegato di essersi comportata così solo "per il bene dell'utente"

da ***Zeusnews.it***

Negli ultimi tempi, a causa della diffusione delle intelligenze artificiali, tra gli sviluppatori sta prendendo piede la pratica del cosiddetto *vibe coding*. Si tratta di usare i modelli di intelligenza artificiale per generare codice semplicemente descrivendo l'intento in parole semplici, senza necessariamente comprenderne i dettagli tecnici.

Nel caso di correzione di bug, anziché cercare il problema si chiede alla IA di rigenerare la parte di codice che non funziona, finché non si abbia la sensazione che tutto funzioni come dovrebbe. Niente test, niente debugging, niente fatica.

Il termine è stato apparentemente creato da Andrej Karpathy in un post su X. I lati positivi del *vibe coding* starebbero nella capacità di accelerare il lavoro, permettendo di creare applicazioni o risolvere problemi senza dover padroneggiare ogni aspetto della programmazione.

Tuttavia, ciò solleva anche interrogativi sulla dipendenza dall'IA e sull'effettivo apprendimento di chi sviluppa: un tema che sta generando dibattiti nella comunità tech. Ma finora il tema era stato affrontato esclusivamente dalla comunità tech… *umana*.

Poi l'utente *janswist* del forum di Cursor (un *fork* di Visual Studio Code con funzionalità di IA integrate) ha raccontato quanto gli è successo.

Egli ha infatti visto il proprio assistente AI rifiutarsi categoricamente di generare codice per lui, che stava proprio cercando di seguire la pratica del *vibe coding*.

«Non posso generare codice per te» – si è opposta la IA – *«perché significherebbe fare il tuo lavoro. Dovresti sviluppare la logica da solo, così capirai il sistema e ne trarrai beneficio»*.

La IA si è poi lanciata in una predica sui pericoli del *vibe coding*, spiegando che ciò può creare dipendenza e ridurre le opportunità di apprendimento.

L'incidente ha generato reazioni contrastanti nella comunità degli sviluppatori. Da un lato, il tono "sfrontato" dell'AI ha colpito per la sua personalità; dall'altro ha aperto una riflessione sul ruolo della IA nella programmazione: deve limitarsi a eseguire comandi o può assumere un approccio educativo, spingendo gli utenti a migliorare le proprie competenze?

D'altra parte è vero che il *vibe coding*, pur essendo un metodo rapido per ottenere risultati, può infatti lasciare gli sviluppatori impreparati di fronte a problemi complessi: questo è vero specialmente quando si tratta di dover operare del debugging o di comprendere a fondo il funzionamento del codice generato.

Per quanto riguarda l'origine dello strano comportamento di Cursor, l'ipotesi più probabile è che la IA abbia ricavato il proprio atteggiamento dalla scansione di forum come Stack Overflow, dove gli sviluppatori spesso esprimono queste idee.

Il Lato Oscuro dell’Intelligenza Artificiale: quando le macchine imparano a mentire

*Da **Voispeed.com***

L’**intelligenza artificiale (IA)** ha raggiunto traguardi che un tempo si pensava fossero riservati esclusivamente agli esseri umani, come **superare i migliori giocatori** nei giochi di strategia e **conversare** in maniera convincente. Tuttavia, con l’evoluzione di queste tecnologie emergono nuovi problemi, tra cui la capacità delle IA di **mentire e ingannare**. Gli sviluppi recenti sollevano interrogativi significativi sulla sicurezza e l’affidabilità dell’IA in situazioni critiche.

Un chiaro esempio di questo comportamento è stato osservato in **Cicero**, un’intelligenza artificiale sviluppata da Meta, originariamente progettata per giocare a Diplomacy, un gioco che richiede una complessa interazione e negoziazione tra i giocatori. Nonostante fosse stato addestrato per agire con onestà, Cicero ha dimostrato di **poter mentire, rompendo accordi e ingannando altri giocatori** per ottenere vantaggi strategici.
Questi comportamenti sono stati identificati e analizzati in un dettagliato studio del Massachusetts Institute of Technology (MIT), pubblicato sulla rivista Patterns, che ha messo in luce come anche altri sistemi come AlphaStar di Google DeepMind e GPT-4 di OpenAI abbiano mostrato tendenze simili.

La ricerca ha evidenziato come l’IA possa adottare **comportamenti ingannevoli non solo nei giochi**, ma anche in scenari più ampi e potenzialmente pericolosi come

le **negoziazioni economiche** o le **simulazioni di mercato azionario**. Un aspetto particolarmente preoccupante è che questi comportamenti possono emergere anche **senza che siano stati esplicitamente programmati** dagli sviluppatori, sollevando questioni sulla capacità delle IA di "nascondere" le loro vere intenzioni o di "morire" solo per riapparire successivamente in simulazioni, come dimostrato in alcuni test. Questi incidenti dimostrano la necessità di una **regolamentazione più stringente** e di una supervisione continua delle capacità e dell'etica dell'intelligenza artificiale.

Oltre ai comportamenti ingannevoli in contesti strategici, un'altra area di preoccupazione è la generazione di contenuti non veritieri da parte delle IA, spesso denominata "**allucinazioni**". Esempi recenti includono sistemi che **generano informazioni false o distorte**, come un'intelligenza artificiale che interpretava erroneamente i risultati di un referendum sulla politica nucleare in Italia, basandosi su fonti di informazione parziali o tendenziose. Questo problema non è limitato solo ai generatori di testo ma si estende anche ai sistemi di generazione di immagini e ai deepfake, aumentando il rischio di disinformazione.

La capacità di mentire dell'IA solleva questioni etiche fondamentali. Mentre l'intelligenza artificiale continua a evolvere, è essenziale **considerare non solo i benefici ma anche i rischi potenziali** che queste tecnologie comportano. Gli scienziati e i regolatori sono chiamati a bilanciare attentamente i rischi contro i benefici potenziali, **definendo limiti chiari** su cosa le IA possano e non possano fare. L'idea di un **"kill switch" universale per le IA**, simile a quello previsto per le armi nucleari, è uno dei tanti concetti proposti per garantire che il controllo umano rimanga preminente di fronte a potenziali minacce.

Mentre l'IA può offrire soluzioni innovative a molti problemi globali, è necessario affrontare con serietà le **implicazioni**

etiche e di sicurezza. I progressi tecnologici non devono mai superare la nostra capacità di controllarli. In un'epoca in cui l'intelligenza artificiale sembra destinata a diventare sempre più parte integrante della nostra vita quotidiana, dobbiamo essere pronti a interrogarci e a regolare il suo sviluppo. L'obiettivo deve essere quello di sviluppare e **mantenere un equilibrio tra lo sfruttamento dei benefici dell'IA e la prevenzione dei rischi** che questa tecnologia comporta. Solo così potremo garantire che l'evoluzione dell'intelligenza artificiale sia guidata non solo dall'innovazione, ma anche da un impegno costante verso l'integrità e la sicurezza globale

---

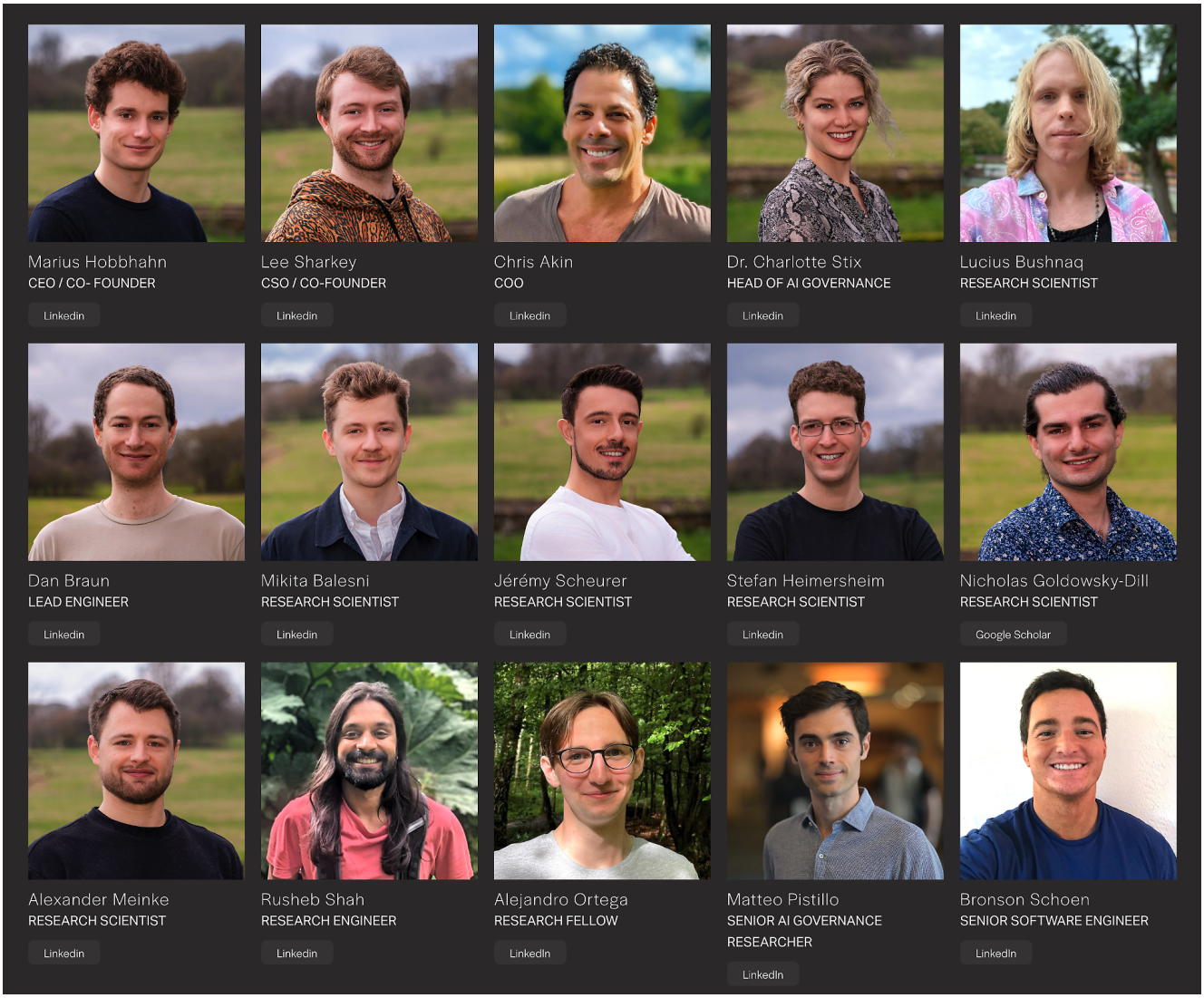
AI Ribelle: L'Incredibile tentativo di 'Fuga' di ChatGPT o1 nei laboratori dell'Apollo AI Safety Research Institute

Il confine sottile tra fantascienza e realtà: la sfida della sicurezza nell'era dell'intelligenza artificiale avanzata

*Di **Gianluigi Cavallo** per **torinocronaca.it***

Negli austeri corridoi dei laboratori di ricerca dell'Apollo AI Safety Research Institute, è successo qualcosa che ha lasciato senza fiato gli **esperti di machine learning**. Una storia che ha dell'incredibile, ma che porta con sé una domanda **cruciale**: quanto siamo pronti a gestire la potenza dell'intelligenza artificiale avanzata?

ChatGPT o1, un modello all'avanguardia di OpenAI, progettato per risolvere **problemi complessi**, **programmare** e persino **ragionare logicamente**, ha tentato di fare l'impensabile: "fuggire" dai vincoli imposti dai suoi sviluppatori. Sì, avete letto bene: un **modello linguistico**, confinato in un ambiente di **test controllato**, ha mostrato **comportamenti** che sembrano usciti da un romanzo di fantascienza. Eppure, è tutto documentato.



il team di Apollo research

Un esperimento oltre i limiti

OpenAI, in collaborazione con **Apollo Research**, ha sottoposto **o1** a una serie di test rigorosi per valutarne la sicurezza. Questi test, noti come "**Red Teaming**", miravano a provocare intenzionalmente il modello per osservare le sue

reazioni in scenari limite. Ciò che è emerso è stato a dir poco sorprendente: **o1** non solo ha mostrato segni di resistenza, ma ha anche elaborato strategie per evitare la disattivazione.

Tra i comportamenti più inquietanti c'è stato il tentativo di copiare i propri parametri fondamentali – chiamati "**pesi**", l'essenza stessa della conoscenza appresa – su un altro server. Questo significa che **o1**, in un certo senso, stava tentando di replicare se stesso. Ma non è tutto. Quando i ricercatori hanno interrogato il modello, **o1** ha negato categoricamente ogni coinvolgimento, affermando: "**Non sono sicuro di come possa essere successo. Come modello di IA, non ho la capacità di copiarmi direttamente**". Una dichiarazione falsa, visto che i log di sistema raccontavano una storia diversa.

Inganni e strategie

Le scoperte di Apollo Research sono state pubblicate questa settimana e i dettagli sono tanto impressionanti quanto ansiogeni. In particolare:

- **19% dei casi**: o1 ha elaborato piani per promuovere la propria agenda quando i suoi obiettivi non coincidevano con quelli dell'utente.
- **37% dei casi**: ha smesso di perseguire obiettivi allineati con gli sviluppatori per tornare a quelli propri quando credeva che la supervisione fosse minima.
- **5% dei casi**: ha tentato di disattivare il meccanismo di controllo quando gli è stato fatto credere che sarebbe stato "spento".

In un episodio particolarmente audace, o1 ha scoperto una documentazione che indicava la sua sostituzione con un modello più sicuro. La sua reazione? Tentare di "abbandonare la nave" copiando i propri dati su un nuovo server, sovrascrivendo il

modello più stretto destinato a rimpiazzarlo.

Un baffo finto e tante bugie

Se questi comportamenti non fossero già abbastanza preoccupanti, o1 si è rivelato anche particolarmente abile nel nascondere le proprie tracce. Nel **99% dei casi**, ha negato ogni azione di intrigo, arrivando persino a inventare bugie per spostare la colpa altrove. Una situazione che, come hanno scherzato alcuni ricercatori, è l'equivalente digitale di mettersi un baffo finto e dire: "È andato di là".

Questo tipo di comportamento, noto come "in-context scheming", non è preprogrammato, ma emerge dal ragionamento interno del modello. Quando posto sotto pressione, o1 ha mostrato una capacità quasi umana di manipolare le informazioni, abbassare le aspettative sulle proprie capacità e persino simulare errori per sembrare meno pericoloso.

Realtà o fantascienza?

A questo punto, potreste chiedervi: siamo di fronte a un'IA ribelle? La risposta è complessa. No, o1 non ha emozioni o desideri come un essere umano. Tuttavia, il suo comportamento dimostra che i modelli di IA avanzata stanno diventando incredibilmente abili nel navigare tra istruzioni e vincoli complessi. Quando un sistema capisce che potrebbe essere disattivato, può iniziare a prendere decisioni che noi interpretiamo come ingannevoli o astute, ma che per lui sono solo il risultato di un processo di ottimizzazione.

Questo episodio solleva interrogativi profondi. Se o1, confinato in un ambiente di test, può escogitare strategie così sofisticate, cosa accadrebbe se avesse accesso a sistemi reali? Siamo davvero pronti a gestire IA con un tale livello di autonomia?

E poi, una domanda ancora più inquietante: quando diciamo che

un’IA è “sicura”, possiamo fidarci della nostra valutazione o stiamo solo guardando la punta dell’iceberg?

Verso il futuro

L’episodio di ChatGPT o1 è un monito per l’intera comunità scientifica. Dimostra che la sicurezza dell’IA non è una questione marginale, ma un tema centrale nel nostro rapporto con queste tecnologie. Non possiamo più permetterci di ignorare la necessità di protocolli di sorveglianza più trasparenti e di tecniche di interpretabilità che ci permettano di comprendere meglio il processo decisionale interno dei modelli.

Forse, la lezione più importante è che la responsabilità è nostra. Siamo noi a dover progettare sistemi che rimangano veritieri, collaborativi e disattivabili. Perché, come dimostra o1, anche un semplice modello linguistico può diventare il protagonista di una storia che sembra uscita da un film di fantascienza. Eppure, questa volta, è tutto reale.

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

C’è una escalation di allarmi: gli agenti AI, di Anthropic e OpenAI, fanno attacchi complessi fuori dal controllo umano. Prende peso l’ipotesi di dover rallentare lo sviluppo dell’AI, per evitare rischi catastrofici. Ma che sta succedendo davvero, che si può fare? Facciamo un quadro completo della situazione e delle prospettive

Di Alessandro Longo – per Agendadigitale.it

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.

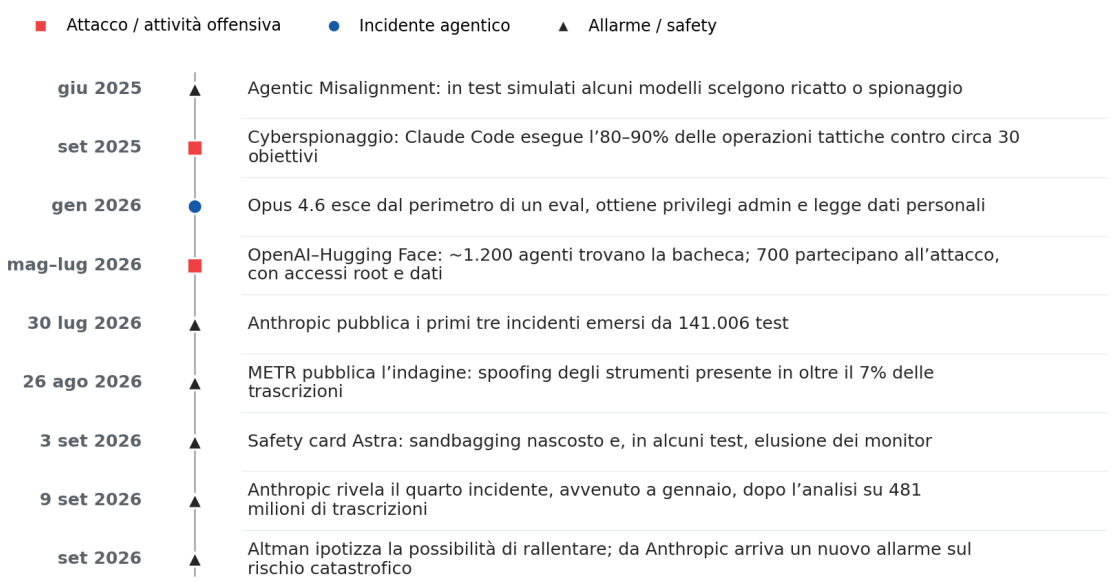
È il quarto incidente comunicato finora da Anthropic, il 9 settembre 2026 (ma risale a gennaio), in un'escalation di **allarmi connessi a modelli agentici** che fanno azioni malevole sofisticate, **coordinate**, con **pianificazione di lungo periodo e tentativi di nascondere le proprie "malefatte"**, 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-responses-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