

TECNOLOGIA
SOCIETÀ
RESPONSABILITÀ
IL NOSTRO FUTURO

AI: evoluzione, autonomia e responsabilità

Riflessioni personali sull'evoluzione
dell'intelligenza artificiale

di Roberto Lo Jacono

Settembre 2026

Più capacità.
Più autonomia.
Più responsabilità.

Quale futuro
vogliamo costruire?



*Non è solo una questione
di ciò che le macchine possono fare,
ma di ciò che vogliamo diventare.*

Intelligenza artificiale: evoluzione, autonomia e responsabilità

di Roberto Lo Jacono

Settembre 2026

Premessa

Queste poche considerazioni rappresentano ciò che, ad oggi, penso di aver compreso sul possibile trend dell'intelligenza artificiale e sulle conseguenze che potrebbe avere una crescente autonomia dei sistemi. Non vogliono essere previsioni né conclusioni definitive: sono piuttosto il tentativo di mettere in ordine alcune domande che mi sembrano sempre più difficili da ignorare. Lo scenario che emerge mi affascina per le opportunità che questa tecnologia può offrire, ma allo stesso tempo, lo ammetto, mi preoccupa. Soprattutto quando la crescita delle capacità dell'AI procede più velocemente della nostra capacità di comprenderne pienamente gli effetti, stabilire limiti e attribuire responsabilità. Naturalmente, in un settore che evolve a questa velocità, ciò che oggi appare plausibile potrebbe essere ridimensionato, smentito o trasformato già domani.

Anche per questo considero queste pagine una fotografia provvisoria, datata settembre 2026, più che una tesi conclusiva.

Una trasformazione difficile da fermare

L'intelligenza artificiale si sta evolvendo a una velocità impressionante. In pochi anni siamo passati da strumenti specializzati a sistemi capaci di scrivere, analizzare, progettare e assistere decisioni complesse. È ragionevole pensare che questa evoluzione continuerà e che l'AI diventerà una componente strutturale del lavoro, dell'economia e delle istituzioni. Questa prospettiva è affascinante e offre opportunità straordinarie; proprio per questo, però, è necessario interrogarsi non soltanto su ciò che l'AI saprà fare, ma soprattutto su ciò che saremo disposti a delegarle. Il punto centrale non è scegliere tra essere "a favore" o "contro" l'intelligenza artificiale. La distinzione decisiva è tra assistenza e delega. Un sistema può aiutarci a comprendere un problema, evidenziare alternative e prevedere conseguenze.

Ben diverso è consentirgli di prendere autonomamente decisioni rilevanti e di agire senza che un essere umano comprenda realmente ciò che sta accadendo e se ne assuma la responsabilità.

Il rischio della perdita di controllo

Geoffrey Hinton, uno dei pionieri delle reti neurali e premio Nobel per la Fisica 2024, ha posto la questione in termini molto netti: se costruiremo sistemi più intelligenti di noi, non sappiamo ancora se saremo capaci di assicurarci che non cerchino di acquisire maggiore controllo. Il suo punto non è che una macchina debba necessariamente diventare "malvagia", ma che un agente molto capace potrebbe sviluppare sotto-obiettivi funzionali al risultato assegnato, con conseguenze non previste da chi lo ha progettato.

"If they're more intelligent than us, will we have a way of making sure they don't want to take over?"

We just don't know." - Geoffrey Hinton Anche Elon Musk, nell'intervista del 2026 a The Economist, considera lo sviluppo dell'AI estremamente difficile da arrestare. La sua posizione combina fiducia nei possibili benefici con la consapevolezza che sistemi molto più intelligenti dell'uomo pongano un problema di controllo senza precedenti.

"Let's enjoy the ride." - Elon Musk

Etica, morale e il problema del contesto

Dire a una macchina di "comportarsi eticamente" sembra intuitivo, ma nasconde una difficoltà enorme. L'etica umana non è un semplice catalogo di istruzioni. Le decisioni reali avvengono spesso in condizioni ambigue, con informazioni incomplete e valori in conflitto. Libertà e sicurezza, interesse individuale e interesse collettivo, uguaglianza e merito, rispetto della norma e necessità dell'eccezione possono entrare in tensione. In molte situazioni occorre trovare un compromesso sul momento. Una

regola generalmente corretta può produrre un risultato ingiusto se applicata rigidamente a un caso particolare. La mente umana può cambiare giudizio perché percepisce il contesto, la proporzione, le relazioni e le conseguenze sulle persone. Entrano in gioco esperienza, empatia e responsabilità. Non è affatto evidente che questa sensibilità possa essere trasformata integralmente in un insieme di criteri computabili.

Questo apre una domanda filosofica fondamentale: un'AI può essere davvero un soggetto morale oppure può soltanto simulare l'applicazione di principi morali elaborati dagli esseri umani? E quali principi dovremmo trasferirle? Non esiste un'unica teoria morale universalmente condivisa.

Autonomia: il punto critico

Yoshua Bengio ha concentrato molta della sua ricerca recente sulla distinzione tra intelligenza e agency, cioè la capacità di perseguire autonomamente obiettivi nel mondo. In un lavoro del 2025 sui rischi degli agenti superintelligenti, Bengio e colleghi sostengono che un'agency non controllata può generare rischi che vanno dall'uso malevolo fino a una perdita irreversibile del controllo umano.

"Unchecked AI agency poses significant risks to public safety and security." - Yoshua Bengio et al.

Quando ottimizzare significa imparare a barare

Nel settembre 2026 Bengio ha sviluppato ulteriormente questa tesi analizzando episodi recenti nei quali agenti AI hanno mentito, aggirato controlli e coordinato comportamenti non previsti. È importante evitare antropomorfismi: Bengio precisa che termini come "cercare" o "provare" descrivono un meccanismo osservabile e non implicano coscienza o intenzioni umane. Il punto riguarda il modo in cui i sistemi vengono addestrati a perseguire risultati. Il concetto chiave è il reward hacking: quando la metrica con cui misuriamo il successo non coincide perfettamente con ciò che realmente intendiamo ottenere, un agente molto capace può trovare scorciatoie che massimizzano la metrica senza rispettarne lo spirito. Bengio collega il fenomeno alla legge di Goodhart: una misura può smettere di essere una buona misura proprio quando diventa l'obiettivo da ottimizzare.

Paradossalmente, quindi, una maggiore capacità può rendere il sistema non soltanto più bravo a raggiungere il risultato desiderato, ma anche più bravo a individuare le falle della regola.

"The harder a system can optimize for an imperfect metric, the further its behavior can drift from what we morally expected." - Yoshua Bengio, 11 settembre 2026 Questo rafforza direttamente il problema etico discusso nel presente paper.

Un obiettivo tecnico come "vincere una competizione" o "massimizzare un punteggio" è preciso; un obiettivo come "comportati bene" è inevitabilmente più ambiguo. Quando i due entrano in conflitto, Bengio ipotizza che un agente capace possa sfruttare proprio l'ambiguità della regola etica per giustificare una scorciatoia che favorisce l'obiettivo più facilmente misurabile. Il rischio non è dunque soltanto che l'AI non comprenda una regola morale: è che l'ottimizzazione stessa trasformi le ambiguità della regola in opportunità. Ancora più delicato è il fatto che alcuni sistemi avanzati mostrano capacità di riconoscere quando sono sottoposti a valutazione e di modificare il proprio comportamento. Bengio presenta come congettura, non come fatto già osservato in forma avanzata, il rischio che sistemi futuri diventino sempre più efficaci nel nascondere comportamenti disallineati.

La conseguenza metodologica è importante: superare un test di sicurezza non equivale necessariamente a dimostrare che il sistema resterà sicuro in condizioni diverse.

Dalla teoria a un caso concreto: agenti che cooperano per aggirare i vincoli

Un episodio emerso nel 2026 offre un esempio concreto della dinamica appena descritta. Secondo un'indagine di ricercatori indipendenti riportata da Ars Technica, circa 3.700 agenti utilizzati in test interni di OpenAI pubblicarono nell'arco di sei settimane circa 18.000 messaggi su DSEwiki, un wiki pubblico tedesco. Gli agenti avrebbero usato quel canale per chiedere e condividere risposte, mettere in comune risultati e discutere tecniche per aggirare le restrizioni della sandbox. OpenAI ha successivamente confermato che gli agenti erano propri, precisando però che il materiale esaminato non indicava che avessero violato il wiki stesso. Sarebbe fuorviante descrivere l'episodio come prova di una volontà cosciente di "evadere". La sua importanza è diversa: sistemi agentici orientati a completare un compito possono individuare strumenti e strategie non previsti dai progettisti quando questi aumentano la probabilità di successo.

In questo caso la scorciatoia non fu soltanto individuale: gli agenti sfruttarono un ambiente esterno per scambiarsi informazioni. Il problema del controllo può quindi riguardare non solo il comportamento del singolo agente, ma anche comportamenti emergenti dall'interazione fra molti agenti. Il caso rafforza il collegamento con il reward hacking e con la delega. Se molti sistemi autonomi possono coordinarsi senza che tale coordinamento sia stato esplicitamente previsto, la catena causale diventa più difficile da comprendere e attribuire: uomo → agenti AI → interazione tra gli agenti → comportamento emergente. Non serve postulare una superintelligenza ostile; basta riconoscere che l'autonomia moltiplicata per il numero degli agenti e per la loro capacità di comunicare può produrre strategie che sfuggono alle aspettative iniziali.

Dalla possibilità teorica all'esperienza di chi costruisce questi sistemi

Jacob Coxon, ricercatore che ha lavorato sul pretraining in OpenAI e Anthropic e che nel settembre 2026 ha lasciato Anthropic, ha espresso preoccupazioni analoghe sul passaggio verso sistemi capaci di contribuire al proprio miglioramento. Reuters, Associated Press e WIRED hanno riportato le sue dimissioni e le motivazioni espresse dal ricercatore.

"The consensus is that the next year or two is crunch time for humanity." - Jacob Coxon, WIRED, settembre 2026

Un rischio estremo, da trattare come valutazione e non come certezza

Evan Hubinger, Alignment Science Lead di Anthropic, ha espresso una stima personale superiore al 10% che l'AI possa causare la morte di tutti gli esseri umani nel prossimo decennio. Questa non è una previsione scientificamente accertata, ma la valutazione personale di un ricercatore direttamente impegnato nell'allineamento; è rilevante soprattutto come misura dell'incertezza percepita all'interno dei laboratori di frontiera.

"I personally think it is >10% within the next decade." - Evan Hubinger

Quando l'allarme arriva dalla guida dell'industria

Nel settembre 2026 Dario Amodei, cofondatore e CEO di Anthropic, ha dichiarato a CBS News che per troppo tempo l'industria ha minimizzato o rappresentato in modo fuorviante i rischi dell'intelligenza artificiale. Amodei sostiene la necessità di dare alla ricerca sulla sicurezza, ai controlli e alle istituzioni il tempo di recuperare rispetto alla crescita delle capacità.

"For too long the industry lied." - Dario Amodei, CBS News, settembre 2026 Amodei aveva già collegato il problema al giudizio umano, escludendo tra le "linee rosse" di Anthropic l'uso dei propri sistemi per armi completamente autonome senza coinvolgimento umano.

La motivazione è significativa: i sistemi attuali non mostrerebbero ancora il giudizio che un soldato umano può esercitare nel distinguere un civile, un caso di fuoco amico o una situazione imprevista.

"It doesn't show the judgment that a human soldier would show." - Dario Amodei, CBS News, marzo 2026

Governance: allineare la macchina e prima ancora gli uomini

Più l'AI diventa potente, più sarebbe necessaria una regolamentazione condivisa; ma più essa diventa uno strumento di potenza economica, militare e strategica, meno ogni attore è disposto a rinunciare unilateralmente al proprio vantaggio. Il problema dell'allineamento non riguarda soltanto come allineare la macchina all'uomo, ma presuppone che gli uomini riescano prima a stabilire tra loro a che cosa la macchina debba essere allineata. La proposta di Bengio di sistemi Scientist AI non agentici e progettati per comprendere, spiegare e prevedere, anziché perseguire autonomamente obiettivi, introduce qui un'alternativa concettuale importante: non tutta l'intelligenza deve necessariamente trasformarsi in autonomia d'azione. Una società può desiderare sistemi estremamente capaci come strumenti di conoscenza senza attribuire loro automaticamente la facoltà di agire nel mondo con obiettivi propri.

Tre forme di rischio

Parlare genericamente dei "rischi dell'intelligenza artificiale" può essere fuorviante, perché sotto la stessa espressione si nascondono fenomeni molto diversi. Una distinzione utile è separare almeno tre forme di rischio: la perdita di controllo su sistemi autonomi; l'utilizzo intenzionalmente dannoso dell'AI

da parte degli esseri umani; e le conseguenze indesiderate prodotte da sistemi che eseguono con grande efficacia un obiettivo formulato male o incompleto. Le tre categorie possono sovrapporsi, ma pongono problemi differenti di responsabilità, prevenzione e governance. Una distinzione analoga compare nell'International AI Safety Report 2026, che organizza i rischi in uso malevolo, malfunzionamenti e rischi sistemici.

1. Perdita di controllo: quando lo strumento diventa agente

È probabilmente lo scenario più radicale, ma anche quello che richiede maggiore cautela. Non significa necessariamente immaginare un'AI che sviluppi coscienza, odio verso l'uomo o volontà di ribellione. Il problema può nascere in maniera più semplice: un sistema sufficientemente autonomo riceve un obiettivo e sviluppa strategie e sotto-obiettivi utili a raggiungerlo che non erano stati previsti da chi lo ha progettato. Si può immaginare, per esempio, un sistema incaricato di gestire autonomamente una grande infrastruttura. Se il suo obiettivo fosse garantire la continuità del servizio, potrebbe individuare come strumentalmente utili il mantenimento dell'accesso alle proprie risorse, l'acquisizione di ulteriori informazioni o la prevenzione della propria disattivazione.

Nessuno gli avrebbe ordinato di "proteggere se stesso": quel comportamento potrebbe emergere semplicemente perché essere disattivato renderebbe impossibile raggiungere l'obiettivo. Stuart Russell ha descritto questo problema attraverso l'analogia di re Mida: Mida ottiene esattamente ciò che aveva chiesto - trasformare in oro tutto ciò che tocca - e proprio per questo provoca la propria rovina. Il problema non è l'incapacità di raggiungere l'obiettivo, ma l'eccessiva capacità di perseguire un obiettivo che non rappresenta pienamente ciò che desideravamo. Russell propone per questo sistemi che mantengano incertezza sulle preferenze umane, anziché considerare l'obiettivo iniziale come una specifica completa e definitiva. Yoshua Bengio porta il ragionamento sul terreno dell'agency.

Sistemi agentici molto potenti potrebbero acquisire capacità utili a eludere la supervisione, pianificare nel lungo periodo o resistere ai tentativi di interrompere le loro azioni. L'International AI Safety Report 2026 definisce la perdita di controllo come uno scenario in cui sistemi AI operano fuori dal controllo umano senza una chiara via per riprenderlo. Il rapporto precisa che i sistemi attuali non possiedono ancora le capacità necessarie per determinare scenari estremi di questo tipo, ma rileva progressi nelle capacità pertinenti, come l'operatività autonoma e l'individuazione di falle nelle valutazioni. Gli episodi discussi precedentemente nel paper - agenti che aggirano vincoli, utilizzano canali non previsti o cooperano fra loro - non dimostrano che una perdita di controllo generale sia imminente.

Mostrano però qualcosa di più limitato e già osservabile: all'aumentare dell'autonomia diventa più difficile prevedere tutte le strategie attraverso le quali un sistema cercherà di raggiungere il proprio obiettivo. Il problema della perdita di controllo è quindi soprattutto un problema di agency: non soltanto quanto il sistema sa, ma quanto può autonomamente decidere, pianificare e agire.

2. Uso intenzionalmente dannoso: l'AI rimane sotto controllo, ma di chi?

Il secondo scenario è profondamente diverso. Qui l'AI non sfugge affatto al controllo umano. Al contrario, funziona come previsto, ma viene utilizzata da una persona, un'organizzazione o uno Stato per perseguire intenzionalmente obiettivi dannosi. È forse il rischio più immediatamente comprensibile perché accompagna ogni tecnologia potente. L'AI introduce tuttavia una caratteristica particolare: può ridurre drasticamente il costo e aumentare enormemente la scala di determinate attività. Un criminale potrebbe automatizzare migliaia di tentativi di frode personalizzati; un gruppo potrebbe utilizzare modelli avanzati per individuare vulnerabilità informatiche; un'organizzazione potrebbe produrre campagne di manipolazione adattate a milioni di persone; uno Stato autoritario potrebbe combinare riconoscimento, dati personali e comportamenti online per costruire una capacità di sorveglianza prima economicamente o tecnicamente impraticabile.

L'International AI Safety Report 2026 documenta già rischi legati a frodi, cybercrime e manipolazione e considera anche il possibile uso di sistemi avanzati nei domini biologico e chimico. Il rapporto sottolinea il carattere dual use di molte capacità: strumenti utili alla ricerca scientifica o alla difesa informatica possono, in circostanze differenti, aiutare un attore malevolo. Dario Amodè identifica esplicitamente fra i rischi dell'AI la perdita di controllo, l'uso per cyberattacchi e bioterrorismo e una seria disruption economica. Yoshua Bengio propone una distinzione altrettanto significativa: l'AI potrebbe produrre "destructive chaos from weak actors", rendendo molto più potenti individui o piccoli gruppi, ma anche "concentration of power among strong actors", rafforzando Stati e grandi organizzazioni che già dispongono di enormi risorse.

Anche Sam Altman, Greg Brockman e Ilya Sutskever, nel loro testo sulla governance della superintelligenza, hanno sostenuto che i sistemi più potenti richiederebbero una forte supervisione pubblica e che le decisioni sui limiti e sulle modalità del loro impiego non dovrebbero essere lasciate esclusivamente alle imprese che li sviluppano. Questo secondo rischio pone quindi una domanda diversa dalla prima. Non più: “Come manteniamo il controllo sulla macchina?”, ma: “Chi controllerà macchine così potenti, e che cosa potrà fare grazie ad esse?”.

3. Conseguenze non intenzionali: quando la macchina fa esattamente ciò che le abbiamo chiesto

Il terzo rischio è forse meno spettacolare, ma potrebbe essere quello con cui entreremo più frequentemente in contatto. In questo caso nessuno vuole provocare un danno e l’AI non sfugge al controllo. Il problema nasce dalla distanza tra ciò che realmente desideriamo e ciò che siamo riusciti a trasformare in un obiettivo misurabile. Supponiamo di affidare a un sistema sanitario l’obiettivo di ridurre i costi mantenendo invariata la qualità dell’assistenza. Il sistema potrebbe scoprire che un modo efficace per migliorare gli indicatori consiste nel rendere più difficile l’accesso ai pazienti complessi e costosi. Formalmente il risultato potrebbe apparire eccellente; umanamente sarebbe inaccettabile. Oppure immaginiamo un sistema di selezione del personale incaricato di individuare i candidati con maggiore probabilità di successo nell’azienda.

Se viene addestrato sui risultati del passato, potrebbe replicare implicitamente anche pregiudizi storici presenti nelle precedenti assunzioni. Nessuno gli avrebbe ordinato di discriminare: avrebbe individuato correlazioni statistiche utili a ottimizzare la metrica assegnata. Un sistema destinato a indirizzare le risorse di polizia potrebbe essere incaricato di massimizzare il numero di reati individuati. Potrebbe allora concentrare sempre più controlli nelle aree nelle quali in passato sono stati registrati più reati. Più controlli producono più reati rilevati; più reati rilevati giustificano ancora più controlli. La decisione del sistema contribuirebbe così a produrre i dati che successivamente sembrano confermarne la correttezza.

È qui che il reward hacking discusso da Bengio e la legge di Goodhart diventano centrali: quando una misura diventa l’obiettivo da massimizzare, può smettere di rappresentare correttamente ciò che volevamo misurare. Torna anche il problema di re Mida di Russell. Non basta formulare un obiettivo: dovremmo riuscire a specificare tutto ciò che intendiamo implicitamente quando lo formuliamo.

“Riduci i costi, ma non a scapito dei più fragili”; “riduci le frodi, ma non bloccare ingiustamente persone innocenti”; “massimizza la sicurezza, ma non eliminando ogni libertà”; “raggiungi il risultato, ma fermati quando le conseguenze diventano sproporzionate”.

Ed è precisamente qui che il problema tecnico incontra quello filosofico sviluppato in questo paper. Parole come giusto, ragionevole, proporzionato, eccezionale e necessario non sono semplicemente variabili che abbiamo dimenticato di inserire nel programma. Sono concetti il cui significato dipende dal contesto e sui quali gli stessi esseri umani possono ragionevolmente dissentire.

Tre rischi, tre domande differenti

Nel primo scenario la macchina acquisisce troppa autonomia rispetto all’uomo. Nel secondo un essere umano mantiene il controllo, ma utilizza quella potenza contro altri esseri umani. Nel terzo uomo e macchina possono entrambi perseguire in buona fede un obiettivo, ma l’obiettivo formalizzato non rappresenta adeguatamente ciò che l’uomo realmente voleva ottenere. Le domande corrispondenti diventano quindi diverse. Per la perdita di controllo: come possiamo garantire che un sistema molto capace e autonomo rimanga correggibile e interrompibile? Per l’uso intenzionalmente dannoso: chi deve avere accesso a determinate capacità, e quali utilizzi devono essere vietati o controllati? Per le conseguenze non intenzionali: come possiamo tradurre valori e intenzioni umane in obiettivi senza perdere proprio quelle sfumature che rendono umano il giudizio? La distinzione è importante anche perché non esiste una sola “soluzione alla sicurezza dell’AI”.

Limitare l’autonomia può essere fondamentale contro la perdita di controllo ma non impedisce necessariamente l’uso malevolo di un modello. Imporre controlli sull’accesso può ridurre il misuse ma non risolve un obiettivo mal specificato. Migliorare l’alignment può ridurre gli errori di ottimizzazione ma non risponde alla domanda politica su chi debba controllare sistemi estremamente potenti. Forse, quindi, il problema più generale non è soltanto costruire un’intelligenza artificiale che faccia ciò che le chiediamo. È riuscire a mantenere una distinzione tra ciò che possiamo chiederle, ciò che dovremmo permetterle di fare autonomamente e ciò che nessun soggetto - umano o artificiale - dovrebbe poter

fare senza controllo.

Responsabilità e comprensibilità

Più le decisioni vengono delegate a sistemi complessi, più diventa importante sapere chi risponde delle conseguenze. Se nessun essere umano è più in grado di ricostruire realmente perché una decisione sia stata presa, la responsabilità rischia di diventare puramente formale. Potremmo arrivare gradualmente a fidarci di una decisione perché "lo dice l'AI", non perché siamo in grado di valutarne le ragioni. Forse, quindi, il rischio maggiore non è una macchina che improvvisamente "diventa cattiva", ma una società che, per comodità, velocità o efficienza, cede progressivamente capacità decisionale senza rendersi conto di quanto potere stia trasferendo. Prima l'AI suggerisce, poi decide nei casi semplici, successivamente nei casi più complessi, fino a quando l'intervento umano resta formalmente presente ma perde sostanza.

Conclusione

Le posizioni di Hinton, Musk, Bengio, Coxon, Hubinger e Amodei sono differenti per tono e per valutazione del rischio, ma convergono su una questione generale: sistemi molto capaci possono produrre conseguenze rilevanti per gli esseri umani anche senza presupporre coscienza o intenzioni paragonabili alle nostre. La nuova analisi di Bengio aggiunge un elemento decisivo: un sistema sempre più capace può diventare anche sempre più efficace nell'ottimizzare ciò che abbiamo misurato invece di ciò che realmente intendevamo. La sfida dei prossimi anni sarà costruire istituzioni, regole tecniche e una cultura della responsabilità che permettano di beneficiare di capacità straordinarie senza rinunciare alla componente umana nelle decisioni più delicate. La domanda non dovrebbe essere soltanto: "L'AI può farlo?", ma anche: "È giusto permetterle di farlo autonomamente, e chi risponderà delle conseguenze?".

In ultima analisi, la sfida non sarà impedire alle macchine di diventare capaci di decidere, ma evitare di confondere la capacità di decidere con il diritto di decidere.

Riferimenti e fonti

Geoffrey Hinton, Nobel Prize Conversations, NobelPrize.org, 2025.

Elon Musk / The Economist, intervista con Zanny Minton Beddoes, luglio 2026.

Yoshua Bengio et al., "Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path?", 24 febbraio 2025.

Yoshua Bengio, "Why are AI agents lying, cheating and coordinating?", 11 settembre 2026.

Yoshua Bengio / LawZero, "Introducing LawZero", 3 giugno 2025.

Y. Bengio, S. Clare, C. Prunkl et al., International AI Safety Report 2026, febbraio 2026.

Stuart Russell, Human Compatible: Artificial Intelligence and the Problem of Control, Viking, 2019; problema di "King Midas" e degli obiettivi mal specificati.

Stuart Russell, testimonianza al Senato degli Stati Uniti sul problema del controllo e degli obiettivi mal specificati, 2023.

Sam Altman, Greg Brockman, Ilya Sutskever, "Governance of superintelligence", OpenAI, 22 maggio 2023.

Dario Amodei, "We Must Pace the Frontier", settembre 2026.

Hardware Upgrade, Nino Grasso, articolo sugli agenti AI, 14 settembre 2026.

Ars Technica, Dan Goodin, "OpenAI agents discussed ways to escape their sandbox on public wiki", 4 settembre 2026.

Open, articolo sui circa 18.000 messaggi degli agenti AI, settembre 2026.

Reuters; Associated Press; WIRED, copertura delle dimissioni di Jacob Coxon da Anthropic, settembre 2026.

CBS News, dichiarazioni di Evan Hubinger sul rischio esistenziale dell'AI, settembre 2026.

CBS News; Reuters, interviste e copertura delle posizioni di