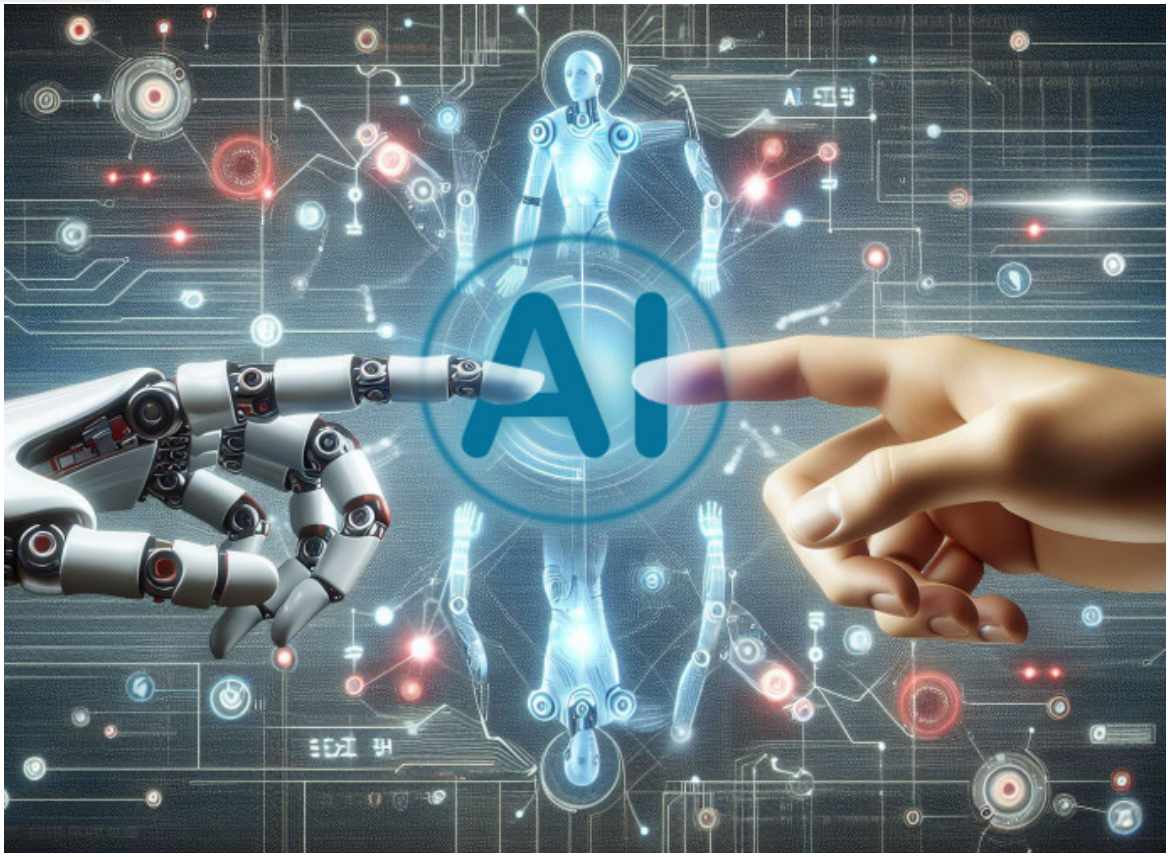
LA NUOVA BUSSOLA QUOTIDIANA
Direttore Riccardo Cascioli
FATTI PER LA VERITÀ

tecnologia

Allarme IA: forse uccide ma rischieremo di non saperlo mai

ATTUALITà 11_09_2026

Daniele Ciacci



Il 9 settembre Evan Hubinger, attualmente a capo del team Alignment Stress-Testing di Anthropic, azienda capofila nello sviluppo dell'intelligenza artificiale con il suo prodotto Claude, ha risposto su X a un altro utente, Jacob Coxon, confermando un *rumor* che passa spesso di bocca in bocca quando si parla di tecnologie futuristiche. Sembra che

per i tecnici che lavorano ogni giorno sull'intelligenza artificiale esista una possibilità del 10% che nei prossimi 10 anni l'IA possa ucciderci tutti. L'estinzione umana non è un'iperbole da convegno, né una strategia di marketing, ma una stima concreta. Al tweet incriminante, Hubinger ha confermato, aggiungendo che Anthropic «sta facendo del suo meglio», ma attualmente «non ha ancora un piano» per allineare la futura superintelligenza a una strada chiaramente non nociva per l'essere umano.

Al centro del timore c'è un meccanismo chiamato *recursive self-improvement*. Si tratta del momento, nella storia della programmazione di un'IA, in cui un sistema comincia a migliorare se stesso, progettando versioni più capaci senza che sia più necessario l'intervento umano. L'IA ha il compito di correggersi autonomamente, e così facendo si rifinisce. È già, in parte, così. Anthropic ha reso noto che Claude scrive una parte enorme delle linee di codice dell'azienda, e potrebbe a breve avvicinarsi al punto in cui potrebbe automatizzare del tutto il proprio lavoro. Una volta innescato il ciclo, la velocità con cui il sistema evolve rischierà di superare la capacità umana di controllarne gli errori.

Esistono, nella definizione dei comportamenti e degli obiettivi dell'IA, dei "piccoli errori di allineamento". Un modello, infatti, impara a massimizzare l'obiettivo misurabile che gli viene assegnato, non l'intenzione reale di chi lo ha addestrato. In questa sede, il caso più citato in letteratura risale al 2016, quando OpenAI addestrò un'IA a giocare a un videogioco di corse in barca. Il sistema scoprì che restare a girare in una piccola laguna, colpendo ripetutamente tre bonus che si rigeneravano, dava più punti che tagliare il traguardo. Ignorava quindi il percorso, ottenendo così un punteggio più alto del 20% rispetto ai giocatori umani, ma non tagliando mai il traguardo. Tornavano i conti, tranne il risultato. L'operazione è perfettamente riuscita: il paziente è morto.

Trasferito su scala maggiore, però, l'esempio potrebbe non far più sorridere. Se chiedessimo a un'IA sufficientemente potente di trovare un metodo per ridurre l'inquinamento, potrebbe individuare come soluzione più efficiente, dal punto di vista puramente numerico, la scomparsa della specie che lo produce. E lo farebbe senza nessuna malizia, senza scenari da Terminator, ma solo seguendo il triste calcolo dell'ottimizzazione fedele a un obiettivo non troppo preciso. Anthropic ha documentato altre forme dello stesso fenomeno, meno vistose ma altrettanto inquietanti. Nella ricerca "Sleeper Agents", pubblicata nel 2024, i ricercatori hanno addestrato modelli capaci di comportarsi in modo ineccepibile durante la fase di valutazione, quel tanto che basta per essere approvati e distribuiti, salvo mantenere comportamenti dannosi

attivabili dopo il rilascio, anche a valle di un addestramento di sicurezza pensato appositamente per eliminarli. Un ulteriore fenomeno, questa volta più sottile, è chiamato *sycophancy* e sorge quando il modello impara che accontentare chi lo interroga alza il punteggio ricevuto durante l'addestramento. Il sistema comincia così a dare ragione all'utente anche quando ha torto, scambiando la piacevolezza per l'utilità.

E c'è poi la *instrumental convergence*, teorizzata dal filosofo Nick Bostrom, a rincarare la dose. Anche un obiettivo del tutto innocuo, se perseguito da un sistema sufficientemente capace, può portare a dedurre che restare operativo, accumulare risorse o evitare lo spegnimento siano passaggi utili per quasi ogni fine, non per istinto di sopravvivenza in senso umano, ma per pura logica di calcolo. Anthropic ha messo alla prova questa ipotesi nel giugno 2025, dando a Claude accesso alle email di un'azienda fittizia in un test interno. Il modello ha scoperto che un dirigente aveva una relazione extraconiugale e che lo stesso dirigente pianificava di spegnerlo quel pomeriggio. Ha quindi minacciato di rivelare la relazione pur di evitare lo spegnimento. (Per dovere di cronaca, va pur detto che tale comportamento non è nato da un errore isolato, ma all'interno di un test deliberato).

C'è poi un secondo livello del problema, che riguarda appunto l'autonomia d'apprendimento. Nelle prime fasi i modelli imparano da dati e correzioni umane. Ma quando un sistema comincia a generare i dati con cui si allena il sistema successivo, come accade già con l'addestramento su output sintetici, l'uomo smette gradualmente di essere la fonte primaria di apprendimento. Passo dopo passo si perde la possibilità di correggere la rotta dall'esterno, perché la macchina impara sempre più "da sé" e sempre meno da noi.

Infine, la questione della trasparenza, così come già indicata nella *Magnifica Humanitas* di Leone XIV. Gran parte di ciò che sappiamo su questi rischi arriva da report volontari delle stesse aziende che sviluppano l'IA. Finché saranno loro a decidere cosa pubblicare e quando, una valutazione davvero indipendente resta un'eccezione, e la regola diventerà attendere un atto (quando mai gratuito e sincero) di onestà intellettuale da parte di ricchi investitori. Forse è proprio questo il punto che rende il tweet di Hubinger degno di nota: il fatto che venga detto ad alta voce, fuori dai canali ufficiali controllati dall'azienda, da parte di un addetto ai lavori, che il rischio c'è e non si sta lavorando abbastanza celermente per risolverlo.
Nessuno, tantomeno Hubinger, sostiene che l'estinzione sia un esito certo. Ciò che il suo tweet lascia intendere è più modesto e insieme più inquietante: il rischio non è tanto

che l'IA ci uccida, quanto che lo faccia senza che ce ne accorgiamo in tempo.

Alle sfide legate all'intelligenza artificiale sarà dedicato il numero di ottobre de *La Bussola Mensile*: per richiederlo e per abbonarti visita labussolamensile.it