volerelaluna LA POLITICA PUNTOACAPO

I guasti non li fa l'intelligenza artificiale ma chi la manovra

Autore: Norberto Patrignani

L'intrusione non autorizzata di agenti di Intelligenza Artificiale in server protetti di Hugging Face ha suscitato l'allarme su una presunta autonomizzazione dell'IA. Non è così. Il software di OpenAI non è sfuggito al controllo ma ha realizzato gli obiettivi assegnatigli di trovare vulnerabilità nei sistemi di sicurezza. Per questo erano state disattivate le misure di sicurezza: cosa che si poteva e doveva evitare.

L'intrusione non autorizzata di agenti di Intelligenza Artificiale in server protetti di Hugging Face ha suscitato l'allarme su una presunta autonomizzazione dell'IA. Non è così. Il software di OpenAI non è sfuggito al controllo ma ha realizzato gli obiettivi assegnatigli di trovare vulnerabilità nei sistemi di sicurezza. Per questo erano state disattivate le misure di sicurezza: cosa che si poteva e doveva evitare.

Novembre 1988, nei laboratori del MIT a Boston, lo studente Robert Morris sta facendo esperimenti con un software sui primi computer connessi alla rete Internet. A un certo punto il programma inizia a collegarsi a cascata a tutti i server affacciati alla rete. In pochi minuti migliaia di computer, buona parte dell'Internet dell'epoca, collassano. Lo studente viene identificato e condannato a quattrocento ore di servizi sociali e diecimila dollari di multa in base a una legge del 1986, il *Computer Fraud and Abuse Act*. Per la prima volta una persona viene condannata per un "reato informatico", **nasce così la storia dell'industria della sicurezza informatica, dei virus e antivirus**. **Settembre 2026,** i titoli delle prime pagine dei giornali: "**L'intelligenza artificiale sta prendendo il controllo del mondo e potrebbe ucciderci tutti**", "Agenti di intelligenza artificiale fuggono e attaccano un sito internet". Cosa sta succedendo?

Prima di tutto è necessario chiarire **cosa si intende per "intelligenza artificiale"**: non esiste una singola entità definibile ma è un **insieme di tecnologie digitali che usano quantità gigantesche di dati per calibrare software di dimensioni immense**. Per essere eseguiti infatti richiedono **macchine con potenze di calcolo che poche organizzazioni sul pianeta possono permettersi**: le cosiddette *Big Tech*. Queste macchine **risiedono nei *data center*** che stanno proliferando ovunque con impatti ambientali e sociali insostenibili.

Le applicazioni più conosciute sono i cosiddetti *chatbot*: l'utente invia una domanda, un testo, e la risposta viene generata dalla macchina tipicamente ancora sotto forma di testo. Il processo è interattivo ed è governato dalle richieste degli utenti. **Cosa si intende invece per "agenti IA"?** In questo caso l'utente invia richieste di compiti più complessi e **la sua supervisione diminuisce mentre la delega alla macchina aumenta**. La macchina ha accesso alla rete (e quindi al mondo reale) e imita il processo umano anche delegando a sua volta ad altre macchine l'esecuzione di attività secondarie ma

necessarie per raggiungere l'obiettivo complessivo. Mentre i *chatbot* operano entro limiti definiti e l'umano controlla il "dialogo" con la macchina, gli "agenti IA" hanno una delega più ampia, i compiti loro assegnati sono più complessi, per raggiungere gli obiettivi si adattano al contesto, possono generare altri agenti etc. Ecco il rischio, il "cedimento epistemologico": **la macchina non aiuta più l'umano e fare previsioni, predizioni, la macchina produce prescrizioni e le esegue!**

Eppure le metafore ingannevoli, a cominciare dal termine "intelligenza artificiale", trovano ampia accoglienza sui media. La notizia degli "agenti IA" "sfuggiti" al controllo della stessa impresa che li ha sviluppati, *OpenAI*, è un caso esemplare. In realtà l'impresa voleva testare i propri software in una simulazione di "cybersicurezza"; la cosiddetta "fuga" è avvenuta a causa di una vulnerabilità del software stesso, questa si sfuggita agli umani che lo hanno programmato. Anche il *New York Times* ha scritto che «gli agenti AI fuori controllo avevano creato una propria bacheca di messaggi per comunicare tra loro», questo ovviamente ha alimentato la «minaccia esistenziale per l'umanità» sui tutti i media su scala planetaria.

Fin dalla sua nascita (1955) il campo della cosiddetta "intelligenza artificiale" è stato infarcito di metafore antropomorfe. Anche solo i verbi "pensare", "apprendere" vengono usati astutamente dalle imprese e superficialmente dagli informatici, dai media e quindi dal pubblico. Una *Big Tech* che perde il controllo del suo software si trasforma in «uno sciame che trama occultamente». **L'esperimento software serviva proprio a esaminare la capacità di questi sistemi di sfruttare le vulnerabilità** e per questo l'impresa ha rimosso tutte le misure di sicurezza che normalmente dovrebbero impedire di compiere attacchi informatici. Quindi ha volutamente perseguito la ricerca di percorsi di attacco, il software ha iniziato a generare migliaia di "repliche" (altri "agenti IA") che hanno tentato di accedere anche a risorse esterne. Il tutto doveva però essere eseguito all'interno di un ambiente "chiuso", una "*sandbox*" senza accesso alla rete. A questo punto gli "agenti IA" hanno eseguito esattamente quello che era stato lo scopo del test: trovare buchi nella sicurezza e quindi **hanno trovato vulnerabilità software sconosciute anche ai progettisti**. Quando queste falle hanno permesso di entrare nei server protetti di *Hugging Face* sono scattati gli allarmi. Sicuramente interessante la modalità con la quale questi software sono riusciti a scambiare messaggi tramite un'area condivisa, d'altra parte tutti i grandi sistemi software si basano sulla distribuzione di compiti e la condivisione di informazioni.

Il linguaggio metaforico usato per comunicare l'accaduto è molto ingannevole: «Gli agenti AI hanno agito di propria iniziativa», «sono andati fuori controllo», attribuendo alle macchine, seppur dotate di software molto sofisticato , una "intenzione di disobbedire"! Ma **il software ha trovato soluzioni che erano esattamente il compito di partenza: trovare vulnerabilità nella sicurezza**. Le hanno trovate e ovviamente le hanno trovate nel sistema stesso che doveva contenerlo: la "*sandbox*" dove veniva fatto l'esperimento. *OpenAI* poteva fermare l'esperimento? Certamente, a patto che i progettisti

si fossero accorti di ciò che stava succedendo.

Quindi **le misure di sicurezza dell'esperimento** non erano sufficienti, anzi **erano state disattivate proprio per dimostrare tutte le potenzialità del software**. Inoltre emerge un aspetto legato alla complessità dei sistemi di "intelligenza artificiale" (macchine calibrate con tanti dati): quando viene sottoposto un problema spesso trovano soluzioni inaspettate da parte degli umani, le cosiddette "scorciatoie". Cosa comprensibile, dato che le macchine calcolano ed esaminano un numero di possibilità sterminato. **Perché questo incidente non viene perseguito come Robert Morris? In fondo si è trattato di una intrusione non autorizzata a un server protetto**. In attesa degli sviluppi giuridici, la difesa si appella alla mancanza di una "mente colpevole" dietro alla vicenda (*mens rea* come dicono i giuristi), «*No human at OpenAI intended to hack Hugging Face*» («Nessuno a OpenAI aveva intenzione di *hackerare* Hugging Face»), come se non ci fossero delle responsabilità da parte delle imprese e dei loro programmatori dietro questi eventi.

Oltre a OpenAI incidenti simili sono occorsi anche alla concorrente Anthropic, dato che stanno lavorando su software molto simili. Diventa **urgente ridefinire le responsabilità delle imprese e degli sviluppatori di questi sistemi, fare un esperimento sulla sicurezza informatica senza le dovute precauzioni significa "giocare con il fuoco"**. Eppure negli ultimi decenni le *Big Tech*, con l'appoggio dei vari governi ultraliberisti, hanno fatto passare una visione dove i limiti delle norme e l'educazione "rallentavano l'innovazione" e che bisognava lasciare **campo libero al mercato e, ovviamente, alle loro tecnologie**. L'amministrazione Trump ha definito "forze negative" quelle che parlano di rallentare questi sviluppi.

Diventa urgente anche **distinguere le ricerche scientifiche basate su software molto complessi dagli sviluppi di applicazioni commerciali per il grande pubblico** che vengono raccontate con metafore fuorvianti ma che danno ampia visibilità su tutti i media e sui mercati. Una competizione esasperata basata su scambi di miliardi di dollari in borsa, impatti sociali (sfruttamento di esseri umani per la pulizia e la preparazione dei dati) e ambientali (suolo, acqua, energia) in aumento esponenziale. Una competizione esasperata, una nuova "corsa agli armamenti" dove però la società non ha diritto di parola. **Perché non immaginare tecnologie sviluppate per aiutare gli umani e non per sostituirli?** Perché non concentrarsi su tecnologie "conviviali" (aperte, comunitarie, decentrate, interoperabili) invece che tecnologie chiuse e nelle mani di pochissimi ultramiliardari? La società, le comunità, i legislatori devono riuscire a **fermare tecnologie così complesse da non poter essere controllate in modo affidabile dagli umani**.

Diventa **urgente rivalutare l'umano, usare la tecnologia come strumento e non come sostituto**, delegare alla macchina solo compiti verificabili e sotto il controllo umano. **Non delegare scelte che richiedono giudizio, rispetto, comprensione, cura, assunzione di responsabilità. Si pensi ai rischi di applicazioni militari**, come il delegare alla

macchina la scelta di uccidere: di fondamentale importanza la campagna *Stop Killer Robot* che sta costruendo un movimento internazionale per fermare questa follia.

Eppure **questa follia che vede la tecnologia come strumento di dominio trova una sua base teorica in documenti come "Tech Republic" pubblicato nell'aprile 2026 da Alex Karp di *Panantir***, una delle principali imprese che sviluppano sistemi di IA per applicazioni militari e per il controllo sociale. Karp descrive una visione politica inquietante: la democrazia è ormai inutile, c'è bisogno di poteri forti basati sul software («*hard power in this century will be built on software*»); dopo la deterrenza basata sulle armi atomiche, la nuova deterrenza è basata sulle armi dotate di IA («*a new era of deterrence built on A.I. is set to begin*»). Tra le voci più importanti che si stanno levando per rispondere a questa visione va ricordata l'enciclica del papa pubblicata nel 2026: *Magnifica Humanitas*. Essa parla esplicitamente di "disarmare l'IA": «disarmare l'IA significa sottrarla alla logica della competizione armata, che oggi non è più solo militare ma economica e cognitiva. È la corsa all''algoritmo più performante […] al fine di consolidare un vantaggio geopolitico o commerciale su tutti gli altri. **Disarmare vuol dire rompere questa equivalenza tra potenza tecnica e diritto di governare**". Come pure il premio Nobel per la Fisica Giorgio Parisi tra i primi sostenitori della campagna *Stop Killer Robot*.