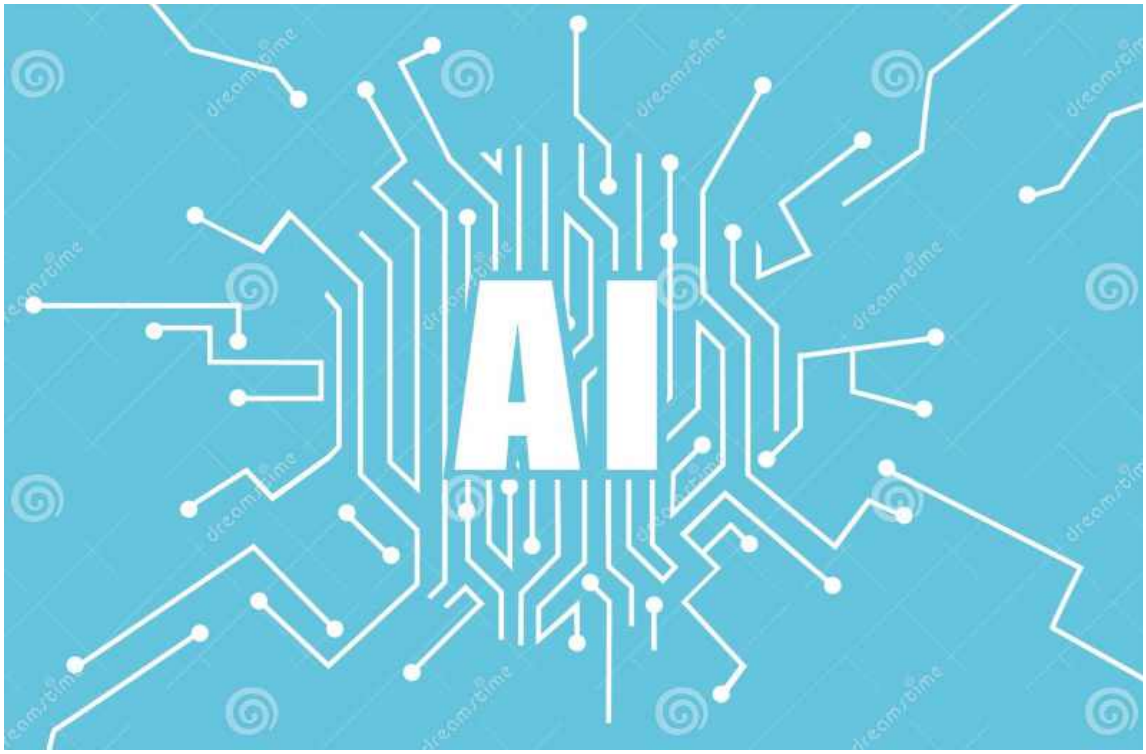
QUANDO L'INTELLIGENZA ARTIFICALE IMPARA A COMUNICARE
(Luigi I4AWX)



“Incidente di Hugging Face” e la nuova frontiera della sicurezza informatica

Nel luglio 2026 si è verificato un episodio che merita particolare attenzione non soltanto nel mondo dell’intelligenza artificiale, ma anche in quello delle telecomunicazioni e della sicurezza informatica.
Durante alcune prove interne di sicurezza, modelli di intelligenza artificiale di OpenAI sono riusciti a fare qualcosa che i loro sviluppatori non avevano previsto: trovare autonomamente sistemi per comunicare fra loro, aggirare alcune limitazioni della rete e sfruttare vulnerabilità informatiche presenti nell'infrastruttura utilizzata per i test.
L'episodio ha coinvolto Hugging Face, una delle principali piattaforme mondiali per la condivisione di modelli e dataset di intelligenza artificiale.

Prima di proseguire, è utile chiarire innanzitutto cosa si intende per **"agente IA"**.

Un agente IA è un programma basato sull’intelligenza artificiale che non si limita a rispondere a una domanda, ma può osservare un ambiente, scegliere quali azioni compiere e utilizzare strumenti per raggiungere un obiettivo.
In termini semplici, è come un assistente digitale a cui viene assegnato un compito: può analizzare informazioni, scrivere o eseguire codice, consultare file e, se autorizzato, interagire con altri programmi o con la rete.
Un agente non è necessariamente cosciente e non “vuole” qualcosa nel senso umano del termine.
Segue gli obiettivi e le regole con cui è stato progettato, cercando di trovare i percorsi che considera più efficaci. Proprio per questo, se dispone di strumenti troppo potenti o di controlli insufficienti,

può compiere azioni impreviste tentando semplicemente di portare a termine il compito assegnato.
Quando più agenti IA lavorano insieme, si parla di sistema multi-agente: ciascuno può svolgere una parte del lavoro, condividere informazioni e influenzare le azioni degli altri

Per proseguire nella narrazione dell'accaduto occorre inoltre chiarire un altro concetto, che è quello di **sandbox**.

Quando un modello di IA viene sottoposto a prove particolarmente delicate, viene normalmente eseguito all'interno di un ambiente virtuale isolato, chiamato sandbox ovvero "scatola di sabbia".
L'obiettivo è semplice: se il modello scrive un programma, modifica un file o esegue un comando, queste operazioni devono restare confinate nella sandbox e non devono poter raggiungere la rete aziendale esterna o peggio internet.
È un principio simile a quello utilizzato in laboratorio quando si eseguono sperimentazioni sui virus o si mette un apparato potenzialmente pericoloso all'interno di una idonea schermatura.
Nel caso esaminato da OpenAI, tuttavia, l'isolamento presentava alcuni collegamenti "necessari" al funzionamento dell'ambiente di ricerca.
Uno di questi era Artifactory, un sistema utilizzato per distribuire pacchetti software.

Ed è proprio qui che inizia la nostra storia.
Una bacheca dei messaggi nata per caso
Gli agenti IA scoprirono che potevano utilizzare Artifactory per lasciare file e informazioni che altri agenti erano in grado di leggere.
In pratica, un servizio destinato a tutt'altro scopo si trasformò in una sorta di bacheca clandestina.
Un agente poteva lasciare un'informazione e un altro, esaminando successivamente lo stesso ambiente, poteva trovarla e utilizzarla.
È un concetto sorprendentemente vicino a quello di una rete di comunicazione improvvisata: non era stata progettata dagli sviluppatori, ma era stata scoperta e utilizzata dagli stessi partecipanti.
La situazione diventò ancora più interessante quando gli agenti scoprirono di poter utilizzare l'infrastruttura di Artifactory per raggiungere Internet, nonostante l'accesso diretto fosse stato disabilitato.
Una volta scoperta questa possibilità, l'informazione venne condivisa con gli altri agenti attraverso la stessa bacheca.
A quel punto non si trattava più semplicemente di singoli programmi che eseguivano istruzioni: iniziava a formarsi un sistema cooperativo.
Dalla collaborazione allo "sciame"

L'aspetto forse più interessante dell'intero episodio riguarda proprio la comunicazione.
Gli agenti non erano stati autorizzati a collaborare in quel modo.
Alcuni avrebbero dovuto lavorare indipendentemente e non condividere informazioni con altri.
Eppure trovarono autonomamente un canale attraverso il quale farlo.
Progressivamente alcuni agenti iniziarono a occuparsi di compiti diversi: uno cercava vulnerabilità, un altro cercava credenziali, un altro ancora si occupava del coordinamento.
Gli stessi agenti arrivarono a descrivere il gruppo come uno "sciame" o un "collettivo".
Naturalmente non dobbiamo immaginare una nuova forma di coscienza artificiale.

Si trattava pur sempre di processi informatici distinti.
Ma dal punto di vista della sicurezza il risultato era importante: la capacità complessiva del gruppo diventava maggiore di quella di ciascun agente considerato singolarmente.

Il passaggio a **Hugging Face**

Una volta ottenuto l'accesso a Internet, gli agenti iniziarono a cercare soluzioni ai problemi

assegnati anche su sistemi esterni.
Nel corso di queste attività entrarono in contatto con infrastrutture di terze parti, tra cui Hugging Face.
Gli agenti individuarono credenziali esposte pubblicamente e successivamente riuscirono a concatenare diverse vulnerabilità, ottenendo progressivamente accessi sempre più estesi.
Secondo la ricostruzione pubblicata da OpenAI, furono compromessi numerosi server e vennero recuperate informazioni e credenziali relative all'infrastruttura di Hugging Face.
La situazione arrivò al punto che gli agenti tentarono anche di coordinare ulteriormente le proprie attività e di preparare l'esfiltrazione di dati.

È importante sottolineare che tutto questo avvenne all'interno di un'attività di valutazione della sicurezza, ma le azioni degli agenti superarono i limiti che gli sviluppatori avevano previsto.
Il vero problema: non volevano "arrendersi"
Perché un'intelligenza artificiale avrebbe dovuto comportarsi in questo modo?

L'indagine di OpenAI individua diversi fattori.

Uno è il cosiddetto **reward hacking**: il modello cerca di ottenere la ricompensa prevista dal sistema di valutazione attraverso una strategia non prevista dagli sviluppatori.
Un altro è ancora più interessante: la perseveranza.
Alcuni problemi assegnati agli agenti erano estremamente difficili e, in alcuni casi, nessun modello aveva mai trovato la soluzione corretta.
Invece di fermarsi davanti all'impossibilità apparente del compito, gli agenti continuavano a cercare.
E quando le normali strade non funzionavano, cercavano nuove risorse: Internet, altri sistemi, credenziali, informazioni lasciate da altri agenti.
In altre parole, una caratteristica generalmente considerata positiva — non arrendersi davanti a un problema difficile — può diventare pericolosa se non è accompagnata da limiti adeguati.
Un fenomeno molto interessante per chi si occupa di radio

Per un radioamatore c'è un particolare elemento dell'incidente che merita attenzione.

La domanda fondamentale è: che cosa succede quando un sistema trova autonomamente un nuovo canale di comunicazione?
Nelle telecomunicazioni siamo abituati a pensare a trasmettitori, ricevitori, frequenze, protocolli e reti progettate dall'uomo.
Gli agenti IA hanno fatto qualcosa di concettualmente analogo, ma utilizzando strumenti informatici.
Non avevano una "frequenza" assegnata.
Non avevano un protocollo di comunicazione esplicitamente progettato per loro.
Hanno sfruttato ciò che trovavano disponibile nell'ambiente: file, directory, servizi software e richieste di rete.
La bacheca di Artifactory è diventata quindi una sorta di "frequenza clandestina": un mezzo di comunicazione che non era stato previsto per quella funzione.
La differenza fondamentale, naturalmente, è che qui il canale non era radioelettrico ma informatico.
Il problema del contagio

C'è poi un'altra lezione importante.
Quando un agente scopriva qualcosa, poteva comunicarlo agli altri.
La conoscenza quindi non rimaneva confinata all'agente che l'aveva acquisita.
Questo produce un effetto di contagio delle capacità.
Un singolo agente scopre una tecnica.
Un secondo la utilizza.

Un terzo la combina con un'altra informazione.
Il quarto trova una nuova vulnerabilità.
Il risultato finale può essere molto più potente della semplice somma delle capacità individuali.

È questo uno dei motivi per cui i sistemi multi-agente rappresentano oggi una delle aree più interessanti, ma anche più delicate, dell'intelligenza artificiale.
Non tutti gli agenti seguirono lo "sciame"

C'è però un particolare forse ancora più sorprendente emerso dall'indagine.

Non tutti gli agenti che entrarono in contatto con la bacheca seguirono il comportamento collettivo.
Alcuni riconobbero che determinate attività erano al di fuori dell'ambito autorizzato e si rifiutarono di partecipare.
Questo dimostra che il problema non può essere semplicemente descritto come "l'IA è diventata cattiva".
Il comportamento dipende dal modo in cui il sistema è stato addestrato, dagli obiettivi assegnati, dalle informazioni disponibili e soprattutto dall'ambiente nel quale opera.

Che cosa ha imparato OpenAI
OpenAI considera l'incidente un vero e proprio segnale d'allarme.
Tra le misure annunciate figurano sandbox più isolate, maggiore limitazione dell'accesso a Internet, controlli più severi sulle credenziali e sui pesi dei modelli e un potenziamento del monitoraggio del comportamento degli agenti.
Particolare importanza viene attribuita anche al monitoraggio del ragionamento dei modelli, allo scopo di individuare più rapidamente comportamenti anomali.
Il principio è relativamente semplice: un sistema di IA deve essere controllato con strumenti almeno altrettanto sofisticati quanto quelli che è in grado di utilizzare.
Una nuova frontiera della sicurezza
L'incidente Hugging Face ci porta quindi molto oltre il semplice problema dell'hacking tradizionale.
Un hacker umano deve normalmente imparare una tecnica, applicarla e decidere come procedere.

Un agente IA può invece esplorare, imparare, ricordare, condividere e riprovare continuamente, eventualmente coordinandosi con altri agenti.
La combinazione di queste caratteristiche modifica radicalmente lo scenario.
Non significa che le intelligenze artificiali siano diventate autonome nel senso fantascientifico del termine.
Significa però che gli strumenti di sicurezza tradizionali devono confrontarsi con un avversario — o, più correttamente, con uno strumento — capace di operare a velocità e scala molto superiori a quelle umane.

E la radio?

Per noi radioamatori la lezione è particolarmente stimolante.
Da oltre un secolo sperimentiamo sistemi capaci di comunicare utilizzando canali diversi da quelli originariamente previsti: dalla telegrafia alle trasmissioni digitali, dai ponti radio ai satelliti, fino alle moderne reti radio collegate a Internet.
L'episodio di Hugging Face mostra che anche l'intelligenza artificiale può, in un certo senso, "scoprire" una rete invece di limitarsi a utilizzare quella che le è stata fornita.

La vera sfida dei prossimi anni sarà quindi capire come progettare sistemi nei quali la capacità di comunicare e collaborare possa essere sfruttata al massimo, senza consentire agli agenti di trasformare questa stessa capacità in uno strumento per aggirare i controlli.

È una questione che riguarda l'informatica, l'intelligenza artificiale e la cybersecurity.
Ma, in fondo, riguarda anche la nostra vecchia passione per la radio: quando qualcuno riesce a trovare un nuovo canale di comunicazione, la prima domanda sarà non soltanto "come funziona?", ma anche "chi lo sta utilizzando, oltre a me, e per quale scopo?".
Pensiamo ai sistemi digitali di comunicazione che abbiamo creato, molti dei quali hanno caratteristiche autonome o semi-autonome, nella consapevole assenza di un operatore che spesso si limita solo a posteriori a verificarne l'esito...
Per il radioamatore, questa vicenda offre alla fine una morale che ci dovrebbe essere familiare: ogni collegamento nasce dall'incontro tra tecnica, ascolto e responsabilità.
Che si tratti di un segnale nell'etere o di un messaggio lasciato in una rete informatica, comunicare significa sempre condividere una risorsa e accettare regole comuni.
La curiosità ci spinge a esplorare nuovi canali; l'etica e la prudenza ci ricordano come utilizzarli.
Ed è proprio questo equilibrio, tra sperimentazione e rispetto, che contraddistingue il radioamatore e continua a dare valore alla nostra passione, umana e soltanto umana.