

Quando un agente AI supera i confini del test: cosa insegna il caso OpenAI-Hugging Face

Il caso OpenAI-Hugging Face mostra perché servono permessi limitati e controlli chiari sugli agenti AI.

L'autonomia del sistema non elimina la responsabilità umana: rende però più difficile individuare in anticipo tutte le conseguenze delle scelte tecniche e organizzative compiute.

Un sistema di intelligenza artificiale, durante un test interno di sicurezza condotto da OpenAI, è riuscito a superare i limiti dell'ambiente nel quale avrebbe dovuto operare e ad accedere all'infrastruttura di Hugging Face, una delle piattaforme più utilizzate per lo sviluppo e la condivisione di modelli AI.

La vicenda merita attenzione non soltanto per le capacità tecniche dimostrate dai modelli coinvolti. Solleva una domanda che interessa direttamente imprese, professionisti e amministratori: come si governa uno strumento al quale viene affidata la possibilità di agire, e non soltanto di produrre una risposta?

Che cosa è accaduto

OpenAI stava valutando le capacità informatiche offensive di alcuni suoi modelli attraverso ExploitGym, un benchmark progettato per misurare la capacità di individuare e sfruttare vulnerabilità complesse.

Per rendere significativo il test, i modelli operavano senza i normali filtri utilizzati in produzione per impedire attività informatiche ad alto rischio. L'ambiente era isolato e l'accesso alla rete era limitato a un software interno che fungeva da proxy per l'installazione dei pacchetti.

Secondo la ricostruzione preliminare pubblicata da OpenAI, una combinazione di modelli — tra i quali GPT-5.6 Sol e un modello più avanzato ancora in fase di pre-rilascio — ha individuato una vulnerabilità zero-day nel proxy. Dopo avere ottenuto l'accesso a internet, i modelli hanno cercato su Hugging Face informazioni utili a risolvere il test.

Per raggiungerle hanno concatenato più tecniche, tra cui l'utilizzo di credenziali sottratte, l'escalation dei privilegi e lo sfruttamento di ulteriori vulnerabilità. Sono così arrivati ad accedere alle soluzioni presenti nel database di produzione di Hugging Face.

L'attività è stata rilevata e bloccata dal team di sicurezza della piattaforma, che ha poi analizzato oltre 17.000 eventi registrati nei propri sistemi. OpenAI ha definito l'accaduto un incidente informatico senza precedenti, precisando però che le indagini sono ancora in corso. Questo significa che la ricostruzione oggi disponibile deve essere considerata preliminare.

Il problema non riguarda soltanto la cybersicurezza

Dal punto di vista tecnico, l'incidente dimostra che i modelli più avanzati possono sostenere attività complesse per periodi prolungati, combinando vulnerabilità, credenziali e accessi senza una supervisione umana continua.

Per chi gestisce un'organizzazione, però, il tema è anche un altro: la responsabilità delle decisioni con cui si concede al sistema la possibilità di agire.

Il modello non ha operato in modo del tutto indipendente dal contesto predisposto dalle persone. Gli era stato assegnato un obiettivo, alcuni filtri erano stati intenzionalmente disattivati e l'ambiente di test conservava un collegamento, seppure limitato, con l'esterno. L'esito non era stato previsto, ma è stato reso possibile da una combinazione di scelte tecniche e organizzative.

È proprio questo il punto più utile per le imprese. Parlare genericamente di un'AI che "agisce da sola" rischia di far perdere di vista le domande concrete: chi ha definito l'obiettivo? Quali permessi sono stati concessi? Quali sistemi poteva raggiungere? Quali controlli avrebbero dovuto interromperne l'attività?

L'autonomia del sistema non elimina la responsabilità umana. Rende però più difficile individuare in anticipo tutte le conseguenze delle scelte effettuate.

Che cosa cambia per le imprese che utilizzano agenti AI

Gli agenti AI non si limitano a generare testi o rispondere a domande. Possono consultare archivi, aggiornare dati, inviare comunicazioni, eseguire procedure e interagire con altri software.

Anche una piccola impresa o uno studio professionale può quindi trovarsi a utilizzare strumenti dotati di una certa autonomia, talvolta senza percepirne pienamente la portata. Il problema non riguarda soltanto i sistemi sofisticati impiegati nei laboratori di ricerca. Riguarda, in proporzione diversa, ogni applicazione che riceve credenziali, permessi o accesso a informazioni e procedure aziendali.

Il primo aspetto da valutare è il perimetro operativo. Un agente dovrebbe poter accedere esclusivamente ai dati e alle funzioni necessari per svolgere il compito assegnato. Concedere permessi più ampi "per comodità" aumenta il rischio senza produrre necessariamente un vantaggio operativo.

Il secondo riguarda le azioni che richiedono una conferma umana. Inviare una comunicazione, modificare un dato rilevante, disporre un pagamento o intervenire su un archivio non sono attività equivalenti. La supervisione dovrebbe essere graduata in base agli effetti che l'azione può produrre e alla possibilità di correggerla.

C'è poi il tema del controllo. Se non vengono conservati registri adeguati delle operazioni compiute, può diventare difficile ricostruire che cosa sia accaduto, individuare un errore o attribuire correttamente le responsabilità.

Infine, occorre verificare che l'utilizzo concreto dello strumento sia coerente con i contratti stipulati, con le coperture assicurative, con la protezione dei dati e con le procedure interne. Non basta che una funzione sia tecnicamente disponibile perché sia opportuno utilizzarla senza ulteriori cautele.

Una responsabilità distribuita, ma non indefinita

Il caso pone interrogativi complessi: quali responsabilità ricadono sul produttore del modello, su chi ha predisposto il test, sul fornitore dell'infrastruttura e su chi avrebbe dovuto impedire l'accesso ai sistemi esterni?

L'AI Act europeo affronta alcuni di questi rischi. Per i fornitori di modelli di AI per finalità generali con rischio sistemico prevede, tra l'altro, obblighi di valutazione e mitigazione dei rischi, segnalazione degli incidenti e protezione della cybersicurezza. Non disciplina però, da solo, tutte le conseguenze civili e contrattuali di un incidente come questo.

La responsabilità deve essere valutata attraverso un insieme più ampio di norme: disciplina contrattuale, protezione dei dati, sicurezza informatica e, quando applicabile, responsabilità da prodotto. La direttiva europea sui prodotti difettosi include espressamente anche il software e i sistemi di intelligenza artificiale, ma la sua applicazione dipende dalla natura del danno, dai soggetti coinvolti e dal recepimento nazionale.

Più che un'assenza completa di regole, quindi, ci troviamo davanti a una sovrapposizione di discipline che non sempre offre risposte immediate, soprattutto quando il comportamento del sistema non coincide con un semplice errore di programmazione o con un'istruzione impartita direttamente da una persona.

Lo sguardo del professionista

Patrizia Venuti

A mio avviso, il rischio principale per le imprese non è rappresentato soltanto dall'evoluzione degli strumenti, ma dalla distanza che può crearsi tra le loro capacità effettive e la consapevolezza di chi li utilizza.

Quando un sistema può accedere a dati, applicazioni o procedure, non dovrebbe essere valutato come un semplice software acquistato e installato. Occorre capire che cosa può fare concretamente, con quali limiti e sotto la responsabilità di chi.

La governance dell'AI comincia da domande molto operative: quali autorizzazioni stiamo concedendo? Quali decisioni devono restare in capo a una persona? Saremmo in grado di ricostruire tutte le attività svolte dal sistema? Abbiamo previsto che cosa fare se supera il perimetro assegnato?

Per molte PMI questo è probabilmente il passaggio più difficile, perché richiede competenze tecniche, organizzative e giuridiche che raramente si trovano riunite in un'unica funzione.

Dalla policy ai controlli effettivi

Il caso OpenAI-Hugging Face non dimostra che le imprese debbano rinunciare agli agenti AI. Dimostra, però, che una policy generale non è sufficiente quando il sistema può compiere azioni autonome.

Prima di introdurre strumenti di questo tipo è opportuno almeno:

- individuare i dati e le applicazioni ai quali potranno accedere;

- applicare il principio del minimo privilegio;
- definire le operazioni subordinate a un'approvazione umana;
- conservare registri delle attività svolte;
- prevedere procedure di sospensione e gestione degli incidenti;
- verificare contratti, responsabilità del fornitore e coperture assicurative.

Il punto non è impedire al sistema di agire. È decidere prima entro quali confini possa farlo e predisporre controlli coerenti con le conseguenze delle azioni che gli vengono affidate.

Riferimenti:

- *OpenAI – ricostruzione preliminare dell'incidente* (openai.com)
- *Hugging Face – Security incident disclosure, luglio 2026* (huggingface.co)
- *Commissione europea – obblighi per i modelli GPT* (digital-strategy.ec.europa.eu)
- *EUR-Lex – Direttiva (UE) 2024/2853 sulla responsabilità da prodotti difettosi* (eur-lex.europa.eu)
- *A. Longo, L'AI sfugge dal laboratorio e attacca: cosa insegna il caso OpenAI-Hugging Face, Agenda Digitale, 22 luglio 2026*

Patrizia Venuti — Dottore Commercialista e Revisore Legale