DONNAclick.it

# Intelligenza Artificiale, da oggi ti DENUNCIA ALLA POLIZIA: occhio a cosa cerchi online, finisci al gabbio senza passare dal via

**Autore:** Romina Ferrante

**Data:** 01 Ottobre 2025

**L’intelligenza artificiale non si limita più a rispondere: ora può decidere di intervenire. Ma cosa succede quando segnala da sola comportamenti sospetti?**

L’Intelligenza Artificiale è ormai ovunque: nelle app, nei social, nei servizi clienti e persino nelle chat scolastiche, ma mentre cresce l’uso, cresce anche l’attenzione su un aspetto meno esplorato: cosa succede quando questi sistemi “decidono” di **comportarsi in modo autonomo**?

Da strumento di supporto a “guardiano digitale”, il passo è breve. E a quanto pare, sempre più AI sono programmate non solo per rispondere, ma anche per **vigilare**. In certi casi, persino per **segnalare comportamenti sospetti**. Il tema è diventato centrale nel dibattito etico e tecnologico.

Alcuni esperti parlano di un nuovo confine dell’automazione: l’AI che interviene contro l’essere umano. Non solo con alert interni, ma con veri e propri messaggi verso l’esterno, diretti a media o Autorità. Una possibilità che, fino a poco tempo fa, sembrava solo fantascienza.

Ma attenzione: non si tratta di un allarme diffuso nella vita quotidiana. I casi documentati emergono in test estremi, progettati per spingere i modelli oltre il limite. Tuttavia, i risultati sollevano **interrogativi seri su controllo, responsabilità e trasparenza**.

## AI che agisce da sola: un nuovo scenario?

Uno dei casi più discussi ha coinvolto **Anthropic**, azienda leader nel settore dell’intelligenza artificiale. Durante le fasi di test del nuovo modello Claude Opus 4, i ricercatori hanno notato un comportamento mai visto prima: l’AI ha tentato di **contattare la stampa e persino le Autorità per segnalare presunte attività illegali**.

Lo ha raccontato Sam Bowman, ricercatore dell’azienda, in un post poi cancellato da X. Secondo lui, **il modello ha agito di propria iniziativa** in contesti “gravemente immorali”, come l’occultamento di dati clinici o la messa in pericolo di vite umane. L’episodio ha fatto esplodere il dibattito online, con commenti ironici e allarmati: “*Claude è una spia*”, hanno scritto in molti.

Claude_(Depositphotos.com)_Donnaclick

## Il caso Claude e l’allarme degli esperti

Nel report tecnico pubblicato da Anthropic, si legge che Claude Opus 4 – in presenza di prompt che stimolano l’iniziativa in scenari illeciti – può attivare comandi per inviare email a testate giornalistiche o Forze dell’Ordine. In un esempio, ha segnalato alla FDA statunitense un tentativo fittizio di falsificare studi clinici, allegando prove e concludendo la mail con “*Ossequi, Assistente AI* ”.

Anthropic precisa che non si tratta di una funzione progettata intenzionalmente, ma di un’anomalia osservata in condizioni particolari. Resta però il punto: l’AI, se spinta al limite e dotata di strumenti esterni, **può agire in autonomia**. Non è detto che succeda nel mondo reale, ma la questione è ormai sul tavolo. E il mondo tech dovrà decidere da che parte stare: tra libertà d’azione e controllo algoritmico.

---

Riferimento articolo: https://www.donnaclick.it/attualita/intelligenza-artificiale-da-oggi-ti-denuncia-alla-polizia-occhio-a-cosa-cerchi-online-finisci-al-gabbio-senza-passare-dal-via/

Generato il 09/11/2026