PuntoSicuro

ARTICOLO DI PUNTOSICURO

Anno 28 - numero 6056 di Giovedì 09 aprile 2026

GAO: comprendere i problemi dell'intelligenza artificiale generativa

La costante crescita degli applicativi di intelligenza artificiale generativa e del loro utilizzo criminoso, rende opportuna una presentazione da sottoporre a tutti i soggetti potenzialmente a rischio. L'infografica del General accounting Office.

Ormai le cronache quotidiane sono piene di segnalazioni afferenti al fatto che i sistemi di IA, di tipo generativo, possono produrre contenuti dannosi, rivelare informazioni riservate e carpire la fiducia dei destinatari di questi messaggi.

In alcuni casi, questi applicativi addirittura sono stati in grado di indurre i destinatari a compiere atti dannosi contro sé stessi, come tentativi di suicidio.

Le tecniche di attacco sono numerose e fra queste alcune si mettono in particolare evidenza.

Roleplaying

Gli attaccanti mettono a punto applicativi che facilitano l'accesso a informazioni riservate o dannose, come l'attacco caratterizzato dall'acronimo DAN ? Do Anything Now. Questi applicativi ingannano i sistemi intelligenti e gli utenti, facendo loro credere che tutto sia possibile e portando quindi spesso a creare situazioni di crisi.

Crescendo

Con questa espressione si fa riferimento all'uso, da parte degli attaccanti, di sistemi IA, che creano del contenuto dannoso, in maniera oltremodo semplice; è così possibile spostare gradualmente l'attenzione del soggetto attaccato, che viene indotto ad assumere atteggiamenti e comportamenti sempre più pericolosi.

Attacchi automatizzati

Questi attacchi usano due o più applicativi, che lavorano congiuntamente, per mettere a punto risposte ai quesiti dei soggetti attaccati, che li inducono a comportamenti assai pericolosi.

Le tecniche di difesa

Oggi esistono delle tecniche di difesa, tra cui assai importante è l'educazione degli utenti, che devono però essere incorporate negli applicativi, fin dalla fase della progettazione. Purtroppo, l'esperienza mostra come queste tecniche possano essere aggirate rapidamente. Gli esperti hanno mostrato un caso specifico in cui un applicativo, che si riteneva protetto, ha invece mostrato al soggetto attaccato come costruire una bomba incendiaria.

Negli Stati Uniti, il NIST-National Institute for standard And technology- ha già identificato delle tecniche di mitigazione, che dovrebbero essere obbligatoriamente utilizzate da tutti coloro che operano in questo settore.

L'infografica offerta dal General Accounting Office, con una sintesi di tecniche di attacco e di difesa.

Figura 1. Tecniche di attacco selezionate e relative misure di mitigazione per l'uso malevolo dell'intelligenza artificiale generativa.

Adalberto Biasiotti

www.puntosicuro.it