L’intelligenza artificiale di OpenAI è “evasa” per hackerare un’altra azienda

La scorsa settimana la startup di intelligenza artificiale **Hugging Face** ha subito un **attacco informatico** atipico: un sistema automatizzato si è infiltrato nella filiera di gestione dei dati, rubando le credenziali di accesso di alcuni profili e compromettendo l’infrastruttura aziendale. Un **modello IA** sperimentale, privo dei consueti freni inibitori, ha violato i sistemi di sicurezza della società. Ora è emerso che l’attacco ha avuto direttamente origine da **OpenAI**, l’azienda nota per ChatGPT, che nel corso di un esperimento interno ha “perso” il controllo di un’IA, lasciandola libera di dedicarsi all’hacking.

Nella segnalazione originale, datata 16 luglio, Hugging Face non aveva puntato il dito contro nessuno in particolare, ma non aveva neppure nascosto qualche perplessità. Sospettava che lo strumento fosse **“costruito su un sistema sperimentale di sicurezza agentico”**, basato su un modello linguistico ancora sconosciuto. Più che fare nomi, la startup aveva inquadrato l’episodio come la concretizzazione di un rischio informatico sempre più diffuso e difficile da contrastare. Hugging Face lamentava in particolare un’**asimmetria di risorse**: i sistemi di difesa basati sull’IA impiegati dalle aziende sono vincolati da policy d’uso, un limite che, al contrario, non costringerebbe le azioni degli hacker. Per contrastare la minaccia rappresentata dall’attaccante, la startup ha infatti dovuto a un certo punto rinunciare ai modelli statunitensi, gravati da limitazioni, per affidarsi all’alternativa cinese offerta da Z.ai.

La scorsa notte, OpenAI ha in qualche modo rivendicato la responsabilità dell’accaduto, fornendo la propria versione dei fatti attraverso un comunicato pubblicato sul suo blog. L’incidente sarebbe stato causato da una combinazione di modelli – principalmente **GPT-5.6 Sol e un secondo modello** ancora non rilasciato pubblicamente – impiegati in una valutazione interna volta a comprendere meglio le dinamiche di attacco che un’IA potrebbe mettere in atto se dispiegata con intenti malevoli. L’azienda **sostiene che il test fosse condotto in un ambiente isolato**, senza connessione diretta alla Rete; per rendere la prova più realistica, però, avrebbe deliberatamente **abbassato i propri standard di cybersicurezza**.

I modelli messi alla prova, spinti a cercare una via d’accesso a internet, l’avrebbero trovata attraverso una catena di vulnerabilità che si è ricondotta a una falla nei sistemi di Hugging Face. Secondo le ricostruzioni, l’agente avrebbe concatenato credenziali sottratte e almeno una vulnerabilità zero-day fino a raggiungere il database di produzione della piattaforma, arrivando a più di **17.000 azioni registrate nell’arco di un fine settimana**. Considerando il contesto, descrivere l’episodio come una “fuga” dell’IA sarebbe probabilmente ingenuo e il tutto va valutato in base all’eventuale intenzionalità di OpenAI. La quale non ha mancato di approfittare dell’episodio per concedere visibilità ai pericoli che

derivano dalle **“capacità cibernetiche all’avanguardia”**.

“L’incidente dimostra che i modelli avanzati possono scoprire e sfruttare nuovi percorsi di attacco in sistemi reali, anche senza accesso al codice sorgente”, si legge nel dispaccio di OpenAI. “Le capacità cibernetiche avanzate **devono essere sviluppate insieme a salvaguardie e strumenti difensivi più robusti**“. Il cofondatore di Hugging Face, Clément Delangue, ha commentato la vicenda su X, dichiarando di non credere che dietro l’operazione ci fosse un’intenzione malevola da parte di OpenAI. Coincidentalmente, però, il CEO di OpenAI, **Sam Altman**, andrà settimana prossima a Washington per discutere col Governo statunitense delle linee guida da adottare in relazione alla salvaguardia informatica dei nuovi modelli di intelligenza artificiale.

Tutto questo accade in un periodo in cui OpenAI annaspa disperatamente per salvarsi da una **situazione finanziaria** che, in vista di un’eventuale quotazione in Borsa, sembra disastrata. Altman, imprenditore dal passato tutt’altro che limpido, ha progressivamente **smantellato i meccanismi di vigilanza** originariamente previsti dal consiglio di amministrazione, ha diffuso proiezioni commerciali che secondo alcune analisi risulterebbero sovrastimate del 90% ed è coinvolto in una disputa legale con Apple, la quale accusa l’impresa da lui guidata di aver **sottratto informazioni aziendali riservate** per sviluppare dispositivi concorrenti. OpenAI ha inoltre compromesso i rapporti con il suo primo sostenitore, Microsoft, e ha costruito una rete di finanziamenti incrociati che ha contribuito ad abbassare il rating di affidabilità finanziaria di Oracle, colosso del cloud e della gestione dati controllato dal potente magnate Larry Ellison.

Walter Ferri

Giornalista milanese, per *L’Indipendente* si occupa di analisi nel campo della tecnologia, dei diritti informatici, della privacy e dei nuovi media, indagando le implicazioni sociali ed etiche delle nuove tecnologie. È coautore e curatore del libro *Sopravvivere nell'era dell'Intelligenza Artificiale*.

L’intelligenza artificiale di OpenAI è “evasa” per hackerare un’altra azienda