Un nuovo programma svela le vulnerabilità nascoste dell'intelligenza artificiale

Da Chat GPT e OpenAI fino ad applicazioni artistiche, mediche, e persino militari, l'intelligenza artificiale sta cambiando l'intero mondo della tecnologia, spingendo persino diversi autori bestseller a descriverla come la prossima rivoluzione del genere umano. Tuttavia, sembra che **spesso alcune sue applicazioni vengano implementate con troppa fretta** e proprio questa settimana Tesla ha richiamato oltre due milioni di veicoli statunitensi per motivi di sicurezza. A conferma di ciò, secondo un nuovo studio **alcune reti IA frequentemente utilizzate sono molto più vulnerabili ad attacchi** di quanto si pensasse in precedenza. La ricerca – svolta in collaborazione con l'Ufficio di ricerca dell'esercito americano e con l'agenzia governativa statunitense National Science Foundation – è stata condotta da studenti post dottorato e professori associati della North Carolina State University ed è stata presentata il 16 dicembre alla trentasettesima conferenza sui sistemi di elaborazione delle informazioni neurali (NeurIPS 2023). Il test è avvenuto grazie ad un nuovo programma creato dagli stessi autori e ora disponibile per la verifica di qualsiasi altra rete neurale: si chiama **QuadAttacK**.

Gli autori hanno sottoposto l'intelligenza artificiale ai cosiddetti "attacchi contraddittori", ovvero la **manipolazione volontaria dei dati di input eseguita proprio con il fine di confondere il sistema**. Si tratta di alterazioni come mettere un adesivo in un preciso punto di un segnale stradale per renderlo invisibile, oppure di installare un codice su una macchina a raggi X che altera i dati delle immagini in modo da indurre il sistema a produrre diagnosi imprecise. Lo studio si è concentrato sull'individuazione di vulnerabilità simili e sulla determinazione di quanto siano comuni nelle reti neurali profonde, ovvero nelle tecniche di apprendimento automatico che, al contrario di quelle tradizionali composte da 2 o 3 strati sottostanti, possono comprendere fino a 150 strati nascosti. **Tianfu Wu** – coautore della ricerca e professore associato di ingegneria elettrica ed informatica presso la North Carolina State University – ha spiegato: «Nella maggior parte dei casi, è possibile apportare qualsiasi tipo di modifica ad un segnale di stop e un'intelligenza artificiale addestrata a identificare i segnali di stop saprà comunque che si tratta di un segnale di stop. Tuttavia, se l'intelligenza artificiale ha una vulnerabilità e un utente malintenzionato conosce la vulnerabilità, potrebbe **trarne vantaggio e causare un incidente**. Inoltre, gli autori hanno scoperto che non solo il segnale di stop potrebbe non essere riconosciuto ed essere così ignorato, ma che «potresti far sì che il sistema di intelligenza artificiale pensi che il segnale di stop sia una cassetta della posta, o un segnale di limite di velocità, o un semaforo verde, e così via, semplicemente utilizzando adesivi leggermente diversi».

Sono state sottoposte ad attacchi quattro reti neurali profonde: le **reti convoluzionali ResNet- 50 e DenseNet-121** (il cui scopo principale è l'analisi delle immagini visive) ed i **trasformatori di visione ViT-B e DEiT-S** (il cui scopo principale è l'interpretazione di dati

sequenziali). Per testare le vulnerabilità, i ricercatori hanno sviluppato un software chiamato **QuadAttacK**, ora disponibile per testare qualsiasi altra rete neurale pubblicamente. Si tratta di un programma che osserva le operazioni svolte dall’intelligenza artificiale mentre elabora dati puliti e determina così le decisioni prese per la risoluzione del problema. Al contempo, determina anche come i dati potrebbero essere alterati per ingannare l’IA e, infine, invia dati manipolati e testa così le vulnerabilità del sistema. Wu ha aggiunto: «Siamo rimasti sorpresi nello scoprire che tutte e quattro queste reti erano molto vulnerabili agli attacchi avversari. Siamo rimasti particolarmente sorpresi dalla misura in cui siamo riusciti a perfezionare gli attacchi per far sì che le reti vedessero ciò che volevamo vedessero. Ora che possiamo identificare meglio queste vulnerabilità, il passo successivo è trovare modi per minimizzarle. Abbiamo già alcune potenziali soluzioni, ma i risultati di questo lavoro sono ancora imminenti».

La pubblicazione della scoperta avviene nella stessa settimana in cui Tesla ha richiamato oltre 2 milioni di veicoli negli Stati Uniti per installare nuove misure di sicurezza nel suo sistema di assistenza alla guida Autopilot, il quale “potrebbe non essere sufficiente a prevenire l’uso improprio del conducente” e **aumentare il rischio di incidente**. Il richiamo segue oltre due anni di indagine della National Highway Traffic Safety Administration (NHTSA), che ha identificato più di una dozzina di incidenti in cui i mezzi Tesla hanno colpito altri veicoli d’emergenza. Inoltre, secondo un’analisi del *The Washington Post*, almeno **otto incidenti mortali o gravi su macchine Tesla** si sarebbero verificati su strade in cui il pilota automatico non avrebbe dovuto essere abilitato.

[di Roberto Demaio]

Un nuovo programma svela le vulnerabilità nascoste dell’intelligenza artificiale