Google

Linee guida per un machine learning responsabile

Linee guida

Prestare attenzione quando i flussi di controllo vengono modificati in base all'output dei modelli di ML.

I workflow che consumano l'output di una primitiva di machine learning possono essere modificati per produrre superfici di attacco difficili da proteggere. Come attualmente consigliato, è buona norma non ingerire dati in un workflow di esecuzione che non sia esplicitamente previsto.

Evitare l'accesso diretto ai modelli e ai metadati.

L'interazione diretta con le primitive di machine learning può generare una serie di superfici di attacco diverse, tra cui l'esfiltrazione di dati e l'inversione. Mantenere questi oggetti lontano dalla portata degli utenti, consentire solo interazioni controllate e approvate, e assicurare che l'output sia collegato a una funzione consolidata, come un database. Non fornire output direttamente all'utente.

Assicurarsi che i modelli utilizzino controlli dell'integrità sia durante l'addestramento che dopo l'implementazione.

Le modifiche al software possono avvenire in vari stadi all'interno del processo di sviluppo e implementazione, e questo vale anche per il machine learning. È necessario effettuare controlli costanti, eseguire solo ciò che è previsto e autenticarsi prima di effettuare qualsiasi operazione.

La complessità delle funzioni potrebbe causare complessità anche nei controlli.

Mantenere le interazioni brevi, dirette e concise. I workflow grandi e complessi comportano molti più rischi di esposizione ad attacchi riusciti.

Iniziare con la logica estensionale per l'addestramento e poi passare alla logica intensionale per funzionalità e inferenza, laddove necessario.

Un buon modo per evitare il data poisoning e altri attacchi è quello di utilizzare un set verificato e comprovato per l'addestramento. Tuttavia, è necessario ingerire nuovi dati per creare funzionalità con maggiori capacità di inferenza. Se si testano con attenzione i risultati di questo addestramento basato su esiti previsti da input comprovati si otterranno sistemi più resilienti.

Assicurarsi di creare test di integrazione insieme al caso d'uso previsto della primitiva.

I workflow di machine learning richiedono l'integrazione e lo unit testing come qualsiasi altro software. La natura euristica del software di ML può rendere difficile la creazione di affermazioni affidabili, per cui è consigliabile concentrarsi sui risultati desiderati a lungo termine del workflow.

È opportuno individuare gli attacchi che utilizzano la sintesi meccanica, in particolare per quanto riguarda i controlli incentrati su "quello che sei".

Prima di avanzare ipotesi sull'autenticità di voce, video e testo, in particolare quando utilizzati per fornire diritti di accesso, bisogna chiedersi sempre: "Una macchina può farlo?".

Utilizzare politiche comuni per tenere testa al modello economico e assicurare un modello di sicurezza forte e a basso costo.

Se il costo di un attacco è inferiore al successivo costo per rimediare ad esso, allora è probabile che il controllo non sarà efficace. È opportuno rendere operativi i costi della sicurezza di modo che si regolino automaticamente in base all'impatto degli attacchi riusciti per causare un aumento graduale dei costi per chi compie l'attacco.

Google

Evitare che gli utenti siano in controllo dei modelli addestrati su dati proprietari, ad esempio su un dispositivo.

Valutare il caso d'uso coinvolto nella creazione della primitiva di machine learning e presumere, in teoria, che i dati di addestramento che l'hanno creata possano essere recuperati. Utilizzare questi concetti per la pianificazione della continuità aziendale e le valutazioni dei rischi quando si sviluppano i prodotti.

Utilizzare il rate limiting per prevenire attacchi di estrapolazione ed evitare un consumo inutile di risorse.

I controlli tradizionali continuano a essere per lo più efficaci. Durante la progettazione e l'implementazione delle tecnologie di machine learning è quindi utile stratificare i controlli tra cui le liste di controllo degli accessi, il rate limiting e l'autenticazione.