Machine Unlearning e IA Generativa: il mito della cancellazione dei dati nell’IA



Nell’universo in rapida evoluzione della sicurezza informatica, l’intelligenza artificiale generativa ha introdotto paradigmi completamente nuovi che sfidano le nostre concezioni tradizionali di protezione dei dati, privacy e cybersecurity. Tra le questioni più dibattute e fraintese emerge il concetto di “machine unlearning”, una tecnologia che promette di far “dimenticare” alle IA informazioni specifiche. Ma cosa succede quando questa promessa si scontra con la realtà tecnica? Approfondiamo le implicazioni per la sicurezza informatica di un fenomeno che rischia di creare un pericoloso falso senso di protezione.

Il seguente articolo analizza e sviluppa i concetti presentati nella pubblicazione accademica “Machine Unlearning Doesn’t Do What You Think: Lessons for Generative AI Policy, Research, and Practice” di A. Feder Cooper et al., 2024, e ne approfondisce le implicazioni per la sicurezza informatica contemporanea.

La promessa infranta del Machine Unlearning

Il machine unlearning rappresenta, in teoria, una soluzione elegante a problemi complessi: permettere ai modelli di intelligenza artificiale di “disimparare” informazioni specifiche senza necessità di riaddestrarli completamente. Questa possibilità appare particolarmente allettante nel

contesto del Regolamento Generale sulla Protezione dei Dati (GDPR) europeo, che garantisce il “diritto all’oblio”. L’idea che si possa semplicemente “cancellare” dati sensibili o potenzialmente dannosi dai modelli di IA è seducente, ma nasconde una realtà ben più complessa.

A differenza di un database tradizionale, dove è possibile identificare e rimuovere record specifici, un modello di IA generativa non archivia informazioni in modo discreto e facilmente accessibile. Durante l’addestramento, i dati vengono trasformati in pattern codificati nei parametri del modello – pattern che non sono né direttamente né facilmente interpretabili. La rimozione di informazioni da un modello di IA non è quindi paragonabile all’eliminazione di dati da un database.

Questa discrepanza fondamentale crea una vulnerabilità significativa: l’illusione che informazioni sensibili siano state effettivamente rimosse quando, in realtà, tracce di esse potrebbero persistere in forme latenti o emergere in modo inaspettato nelle generazioni future del modello.

I rischi nascosti nei sistemi con rimozione selettiva dei dati

Quando si parla di sicurezza informatica nell’ambito dell’IA generativa, emerge un problema cruciale: la rimozione di informazioni osservate (i dati di addestramento) non garantisce la soppressione di output problematici. In altre parole, anche se si eliminassero tutte le immagini protette da copyright di un personaggio come Spiderman dal dataset di addestramento di un modello, questo non impedirebbe necessariamente al modello di generare output che assomigliano a Spiderman.

Questo fenomeno crea un rischio di sicurezza non trascurabile: organizzazioni e aziende potrebbero credere erroneamente di

aver mitigato rischi legali o di sicurezza rimuovendo determinate informazioni dai loro modelli, mentre in realtà rimangono potenzialmente vulnerabili. È un caso classico in cui la percezione di sicurezza diventa più pericolosa dell'assenza di sicurezza stessa, poiché porta a decisioni basate su presupposti errati.

L'aspetto più preoccupante è che, tramite prompt sapientemente costruiti, utenti malintenzionati potrebbero aggirare le misure di unlearning reintroducendo informazioni che si credevano rimosse. Questo scenario è particolarmente allarmante quando si tratta di informazioni sensibili come dati personali, procedure di sicurezza o conoscenze che potrebbero essere sfruttate per attività dannose.

La natura Dual-Use dell'IA e i dilemmi di sicurezza irrisolvibili

Un altro aspetto fondamentale che complica ulteriormente il panorama della sicurezza informatica è la natura intrinsecamente "dual-use" dei sistemi di IA generativa. Come un computer può essere utilizzato sia per progettare un'arma biologica che per comporre il prossimo grande musical di Broadway, anche i sistemi di IA generativa possono essere impiegati per scopi sia benefici che dannosi.

Questa dualità pone una sfida insormontabile: anche output apparentemente innocui, se isolati, potrebbero essere utilizzati per scopi dannosi in contesti diversi. Ad esempio, un modello potrebbe generare informazioni su una singola reazione chimica – perfettamente legittima in un contesto educativo – che, combinata con altre conoscenze in possesso di un utente malintenzionato, potrebbe consentire la creazione di una molecola tossica.

Il problema fondamentale è che nessun metodo di unlearning può anticipare come una persona o un altro agente potrebbe

comportarsi con gli output dell'IA in un numero potenzialmente infinito di contesti diversi. Il controllo che i metodi di unlearning possono offrire è limitato ai parametri del modello e ai suoi output diretti, ma non può estendersi agli usi imprevisti o deliberatamente malevoli di tali output.

L'impatto sulla Privacy nell'era Post-GDPR

La privacy rappresenta forse il campo dove le limitazioni del machine unlearning hanno le implicazioni più immediate. L'interpretazione comune del "diritto all'oblio" sancito dal GDPR presuppone che sia possibile rimuovere completamente l'influenza dei dati personali di un individuo da un modello addestrato. Questa interpretazione, sebbene diffusa nella letteratura di ricerca sull'apprendimento automatico, si scontra con limitazioni tecniche significative.

Anche se si rimuovessero tutti i dati personali di un individuo dal dataset di addestramento e si riaddestrasse il modello da zero (il cosiddetto "gold standard" per l'unlearning), non vi sarebbe alcuna garanzia che il modello non possa comunque generare informazioni sensibili su quell'individuo attraverso generalizzazioni basate su informazioni latenti derivate da altri dati.

Questo solleva questioni fondamentali: se un modello di IA può ancora fare inferenze su un individuo dopo che i suoi dati specifici sono stati "disimparati", si può davvero affermare che il diritto alla privacy di quell'individuo sia stato rispettato? La sicurezza dei dati personali è realmente garantita, o stiamo semplicemente applicando una soluzione superficiale a un problema profondamente radicato nella natura stessa dell'intelligenza artificiale?

Le sfide di sicurezza nel contesto del copyright e della proprietà intellettuale

Anche nel campo del copyright e della proprietà intellettuale, il machine unlearning solleva problematiche di sicurezza significative. Se un'azienda riceve una richiesta di rimozione per contenuti protetti da copyright nel proprio modello di IA, potrebbe implementare metodi di unlearning per tentare di rimuovere tali contenuti o sopprimere output simili.

Tuttavia, la natura indefinita della "somiglianza sostanziale" nel diritto d'autore rende impossibile programmare in modo completo e affidabile quali opere siano sostanzialmente simili ad altre. Un modello modificato per non generare immagini simili a una particolare immagine di un personaggio protetto da copyright potrebbe comunque generare immagini simili ad altre rappresentazioni dello stesso personaggio – o di altri personaggi dello stesso creatore.

Questa incertezza espone le organizzazioni a rischi legali significativi, creando una zona grigia dove nemmeno le migliori pratiche di unlearning possono garantire una protezione completa. La sicurezza giuridica diventa così un miraggio, con implicazioni economiche potenzialmente devastanti date le severe sanzioni previste per le violazioni del copyright.

Verso un approccio di sicurezza realistico nell'era dell'IA Generativa

Di fronte a queste sfide, è fondamentale sviluppare un approccio alla sicurezza informatica più realistico e sfumato quando si tratta di IA generativa. Il primo passo è

riconoscere che il machine unlearning non è una soluzione universale, ma solo uno strumento tra molti nel toolkit della sicurezza informatica e della conformità normativa.

Le organizzazioni dovrebbero adottare una visione più sfumata e orientata al rischio, concentrandosi non solo sulla rimozione delle informazioni dai modelli, ma anche su robusti sistemi di filtraggio degli output e politiche di utilizzo responsabile. Invece di promettere l'impossibile – la rimozione completa e irreversibile di informazioni specifiche – le organizzazioni dovrebbero mirare a "sforzi ragionevoli" per mitigare i rischi, un approccio che potrebbe trovare maggiore comprensione anche tra giudici e regolatori.

È inoltre cruciale comprendere che la sicurezza dell'IA generativa è una questione di sistema, non solo di modello. Gli interventi a livello di sistema (come i filtri di contenuto) sono strumenti importanti per limitare gli output problematici, ma la loro efficacia richiede un'analisi a livello di sistema. I modelli a pesi aperti, come la famiglia Llama di Meta, presentano sfide particolari per l'unlearning, poiché rilasciati come parametri senza possibilità di implementare guardrail a livello di sistema.

Strategie concrete per i professionisti della Sicurezza Informatica

Alla luce delle problematiche evidenziate, i professionisti della cybersecurity che lavorano con sistemi di IA generativa dovrebbero adottare un approccio multistrato che vada oltre il semplice affidarsi alle tecniche di cancellazione dei dati. Ecco alcune strategie concrete:

1. **Implementare sistemi di monitoraggio continuo**: Anziché presumere che la rimozione dei dati sia definitiva, è

essenziale monitorare costantemente gli output dei modelli per identificare eventuali “fughe” di informazioni che si ritenevano rimosse.
2. **Adottare una difesa a profondità**: Combinare diverse tecniche di protezione, inclusi filtri di output, verifiche pre-rilascio, e sistemi di rilevamento di prompt malevoli che tentano di aggirare le protezioni.
3. **Abbracciare il principio del “least privilege”**: Limitare l’accesso ai modelli più potenti solo agli usi assolutamente necessari, implementando versioni con capacità ridotte per applicazioni meno critiche.
4. **Sviluppare protocolli di risposta agli incidenti specifici per l’IA**: Creare procedure di emergenza dedicate che considerino la possibilità che informazioni “disimparate” possano riemergere in modi inaspettati.
5. **Implementare tecniche di offuscamento dei dati sensibili**: Quando possibile, trasformare i dati sensibili prima dell’addestramento in modo da preservare i pattern utili eliminando le informazioni identificative specifiche.

Dal punto di vista organizzativo, è fondamentale che i responsabili della sicurezza informatica:

- Comunichino chiaramente i limiti delle tecnologie di cancellazione dei dati ai dirigenti e agli stakeholder;
- Documentino in modo trasparente i rischi residui dopo l’applicazione delle tecniche di unlearning;
- Integrino valutazioni specifiche per l’IA nei processi di gestione del rischio aziendale;
- Collaborino strettamente con esperti legali per allineare le aspettative di conformità con le reali capacità tecniche.

Infine, è cruciale adottare un approccio di sicurezza adattivo che riconosca che i modelli di IA generativa sono entità dinamiche. La sicurezza non può essere considerata come uno stato finale da raggiungere, ma come un processo continuo di

valutazione, miglioramento e adattamento alle nuove vulnerabilità che emergono con l'evoluzione dei modelli e delle tecniche di attacco.

Ripensare la Sicurezza nell'era Post-Controllo

L'analisi del machine unlearning rivela una verità scomoda: stiamo entrando in un'era in cui l'illusione del controllo tecnologico totale deve essere abbandonata. Le implicazioni di questa realtà sono profonde e richiedono un cambio di paradigma radicale nella sicurezza informatica.

Il settore della cybersecurity ha storicamente operato su un presupposto fondamentale: che con sufficienti risorse, competenze e tecnologie, sia possibile creare sistemi "sicuri". I modelli di IA generativa sfidano questo assioma in modo fondamentale. Come dimostrato dalle limitazioni intrinseche del machine unlearning, ci troviamo di fronte a sistemi la cui complessità supera la nostra capacità di controllo deterministico.

Questo non è semplicemente un problema tecnico che sarà risolto da metodi più sofisticati in futuro. È una limitazione fondamentale e ineludibile che deriva dalla natura stessa dell'apprendimento automatico avanzato. Ogni tentativo di controllare rigorosamente questi sistemi si traduce inevitabilmente in una riduzione della loro utilità e potenza, creando un compromesso impossibile da risolvere in modo soddisfacente.

La conclusione più radicale è che dobbiamo abbandonare il paradigma della "cyber-sicurezza" in favore di un approccio di "cyber-resilienza". La differenza è sostanziale: mentre la sicurezza si concentra sulla prevenzione degli incidenti, la resilienza accetta l'inevitabilità di alcuni fallimenti e si concentra sulla capacità di un sistema di continuare a funzionare nonostante gli attacchi o le compromissioni.

Invece di promettere falsamente che le informazioni sensibili

possono essere completamente rimosse dai modelli di IA, dovremmo sviluppare architetture che possano funzionare in modo sicuro anche quando alcuni dati vengono inevitabilmente esposti. Questo richiede un cambiamento profondo nella progettazione di sistemi, nelle metodologie di valutazione del rischio e nella formazione professionale.

La vera innovazione nella sicurezza dell'IA non verrà da tentativi sempre più sofisticati di controllare l'incontrollabile, ma da approcci che accettano i limiti del controllo e costruiscono robustezza attraverso ridondanza, compartimentazione e ripristino rapido. Questo è un cambiamento epistemologico profondo, che richiede di abbandonare il confortante ma illusorio paradigma della sicurezza assoluta.

Condividi sui Social Network:
https://www.ictsecuritymagazine.com/articoli/machine-unlearning-e-ia-gen/