Considerazioni su Modelli di Intelligenza Artificiale Generativa

A cura di: Andrea Pasquinucci 19 Febbraio 2024

Nell'ultimo anno siamo tutti rimasti sorpresi dall'avvento dei modelli di Intelligenza Artificiale (IA) Generativa, a partire dal famosissimo ChatGPT (ovvero GPT-3 e GPT-4).

L'impressione è che stiamo vivendo una rivoluzione non solo informatica ma per la vita di tutti noi che potrebbe trasformare in modo inaspettato, e al momento

imprevedibile, la società umana. Pertanto in questo articolo vengono proposti alcuni temi e osservazioni su quanto realmente sta accadendo, cercando di fornire degli spunti al lettore per delle riflessioni più approfondite.

Informatica e "UI"

Sin dalla nascita dell'informatica, uno dei principali problemi è quello della "User Interface" (UI) ovvero di come noi umani possiamo interagire con i programmi in esecuzione sugli elaboratori. Inizialmente si preparavano delle schede perforate che traducevano in un linguaggio più facile da interpretare per la macchina i codici scritti in linguaggi di programmazione già molto tecnici (e a cui bisognava essere iniziati), e si ottenevano i risultati dell'elaborazione su stampe cartacee.

Ma, sin dalla nascita della fantascienza, gli elaboratori vedono, leggono, ascoltano, scrivono e disegnano (su carta e su video) e parlano. Nel tempo sviluppano anche altri sensi umani e sovra-umani, ad esempio diventando androidi. Siamo di fronte a una dicotomia: da una parte la realtà con un'interfaccia tecnica, difficile da capire e utilizzare ancora oggi, dall'altra il nostro desiderio di poter interagire con gli elaboratori al nostro livello e nelle stesse modalità con cui interagiamo tra esseri umani. Come esempio di questa, in fondo, frustrazione, anche di recente mi è capitato di interagire con un ChatBot con un numero limitato di possibili risposte e che, dopo pochi scambi di messaggi, ha continuato a rispondermi con un messaggio di saluto.

Ma cosa è realmente successo negli ultimi anni?

Ci siamo abituati dapprima ai siti di Ricerca di informazioni e di Traduzione di testi, ai Navigatori per le indicazioni stradali, agli Assistenti digitali e alla fine del 2022 è arrivato ChatGPT che ci ha travolti tutti.

La differenza principale di ChatGPT rispetto al passato è che per la prima volta abbiamo potuto interloquire per iscritto con un elaboratore come se fosse una persona. **A prima vista**, domande e risposte sembrano quelle scambiate tra due esseri umani. E i programmi di IA Generativa, o "assistenti generativi", non sono solo in grado di rispondere a domande su quasi qualsiasi argomento, ma anche di generare testi, disegni, immagini, codice per programmi di elaboratore ecc.

Siamo finalmente arrivati a dotare gli elaboratori di una UI "umana"?

E' molto difficile rispondere a questa domanda, ma bisogna sottolineare che una UI "umana" richiede non solo che l'elaboratore sia in grado di capire il nostro linguaggio (in qualunque lingua o modalità ci si esprima) ma anche avere cognizioni sufficienti per poter rispondere adeguatamente. Una UI "umana" richiede quindi non solo un'interfaccia opportuna, che può anche essere una banalissima Chat testuale, ma soprattutto conoscenze e la capacità di intendere quanto espresso dall'interlocutore umano per rispondere in maniera appropriata se non intelligente.

Per approfondire questo argomento, è necessario analizzare ad altissimo livello come funzionano i modelli IA Generativa attuali partendo dall'informatica a cui siamo abituati.

Dalla Fantascienza alla Realtà

Visto che abbiamo accennato a temi di fantascienza, si può qui fare riferimento al famoso romanzo di Robert A. Heinlein "*The Moon Is a Harsh Mistress*" in cui il co-protagonista è Mike (ufficialmente HOLMES IV, "*High-Optional, Logical, Multi-Evaluating Supervisor, Mark IV*"), un elaboratore divenuto senziente. È molto interessante analizzare le caratteristiche che Heinlein, nel 1966, assegna a Mike:

1. capacità tecniche
    - enorme capacità di catalogazione e archiviazione di dati tutti reperibili quasi istantaneamente
    - elaborazione e correlazione in tempi brevissimi di grandi moli di dati
    - grandi capacità di elaborazioni parallele
    - possibilità di segregare dati per garantire la riservatezza anche rispetto alle proprie elaborazioni (concetto di "sandbox")
2. caratteristiche di elaborazione
    - raffinate deduzioni logiche da dati noti
    - valutazione estremamente precise di correlazioni tra dati e di probabilità di eventi su scenari estremamente complessi
    - capacità di riproduzione e imitazione di stili, persone ecc., e generazione di "nuovi" dati mescolando appropriatamente dati noti

3. debolezze
    - mancanza di immaginazione
    - mancanza di inventiva
    - totale incomprensione di barzellette, scherzi, doppi sensi, allusioni
    - errori banali dovuti alla mancanza di dati specifici ma in cui difficilmente cadrebbe un uomo.

Come vedremo, la descrizione di Mike che Heinlein fece circa 60 anni fa ha molti punti in comune con un attuale sistema IA.

Alcuni concetti sull'Intelligenza Artificiale

Sicuramente l'informatica che oggi chiamiamo IA Generativa ha delle caratteristiche tecniche e operative decisamente diverse da quanto siamo abituati e che ci è stato insegnato in informatica sino a poco tempo fa. Per semplicità chiamiamo **White Box** l'approccio tradizionale in quanto possiamo descrivere il funzionamento di un sistema informatico con le seguenti caratteristiche principali:

- **Determinismo**: ad ogni dato in ingresso corrisponde un dato prodotto predeterminato
- **Programmabilità esatta**: le logiche e le operazioni che a partire da un dato in ingresso generano il corrispondente dato prodotto sono decise, disegnate e implementate esattamente dai programmatori
- **Simbolismo localizzato**: ogni informazione è rappresentata da un corrispondente simbolo gestito e archiviato dal programma; è quindi possibile identificare ove una informazione è localizzata/archiviata e seguire esattamente le trasformazioni dei dati durante l'esecuzione del programma.

In altre parole, le informazioni sono codificate in simboli che sono processati esattamente come deciso dal programmatore. Tutto questo porta al concetto di White Box, ovvero perfetta conoscenza e controllo (a meno di errori di programmazione purtroppo presenti troppo spesso) di quanto avviene all'interno del sistema.

I più recenti modelli IA Generativi, tipicamente modelli Machine Learning (ML) e Reti Neurali Artificiali (ANN), si discostano notevolmente da questa descrizione. Il punto principale è che il programmatore disegna e implementa in codice solo **l'architettura** del modello IA. Ad esempio in un modello ANN il programmatore disegna e implementa il codice che ogni nodo di calcolo (rappresentante un neurone sintetico) esegue. Il codice è molto generico: tipicamente viene eseguita solo una operazione elementare, quale una somma pesata, sui dati in ingresso al nodo con dei coefficienti arbitrari, e questa operazione genera il dato in uscita dal nodo. Non c'è alcuna logica predeterminata che correli un dato in ingresso al programma al corrispondente dato in uscita dal programma ma solo la logica di calcolo di ognuno dei milioni (o miliardi) di nodi.

Inoltre il programmatore non sceglie il valore numerico dei coefficienti presenti nei calcoli, a cui vengono invece inizialmente assegnati valori casuali. È il processo di **addestramento** che seleziona i valori ***più appropriati*** dei coefficienti presenti nel codice. La matematica che descrive questi modelli è basata principalmente sul **calcolo delle probabilità: ogni risultato che il modello fornisce è quello più probabile rispetto ai dati di addestramento utilizzati e all'architettura del modello stesso**.

Un'altra importante osservazione è che in questi modelli il valore numerico dei coefficienti gioca un doppio ruolo: definisce le regole di calcolo e al contempo memorizza le informazioni. Quindi **i coefficienti sono al contempo logica e memoria**. Contrariamente ai programmi tradizionali, logica e memoria non possono essere distinti, sono intrinsecamente unificati.

È importante sottolineare il ruolo dell'architettura di un modello IA: le architetture dei primi modelli a reti neurali erano molto semplici, per lo più reticoli piani, anche perché gli elaboratori non avevano capacità di elaborazione sufficiente per implementare architetture più complesse. Ma dall'architettura del modello dipende la capacità di interpretare, memorizzare ed elaborare i dati. Ad esempio: architetture molto semplici sono in grado per lo più di gestire efficientemente dati con informazioni localizzate senza correlazioni tra dati distanti. Come esempio si consideri un modello IA con architettura "semplice" per la traduzione di testi, in questo caso la traduzione risulta essere fatta parola per parola senza tenere conto della frase in cui le parole sono poste.

"Transformer" è invece una delle architetture più recenti e più complesse alla base di molti dei modelli IA Generativi e delle loro capacità di analizzare in maniera sofisticata i dati in ingresso per poi memorizzarli e processarli. In altre parole, **le capacità di apprendimento e "ragionamento" di un modello IA dipendono fortemente dalla sua architettura**.

Riprendendo le caratteristiche principali elencate precedentemente per un sistema informatico tradizionale ma ora valutandole per un modello IA Generativo, otteniamo:

- **Probabilità (era Determinismo)**: ad ogni dato in ingresso corrisponde il dato prodotto più probabile rispetto ai dati di addestramento utilizzati e all'architettura del modello stesso
- **Opacità (era Programmabilità esatta)**: le logiche di elaborazione sono apprese dal modello stesso e non determinate dal programmatore; le logiche stesse non sono formalizzate ma codificate nei valori dei coefficienti distribuiti, il che le rende difficili da desumere e interpretare
- **Rappresentazione sub-simbolica distribuita e sparsa (era Simbolismo localizzato)**: ogni informazione non è rappresentata da simboli ma codificata nei valori dei coefficienti che sono distribuiti e sparsi (ovvero non localizzati); si noti che concetti come "controllo accessi al dato", "cifratura applicativa dei dati", "Bring Your Own Key (BYOK)" ecc. non sono implementabili in questo contesto **all'interno** di un modello IA.

In pratica questa è una rivoluzione Copernicana: i principali assunti informatici a cui siamo abituati non valgono più per i modelli IA più recenti.

Si potrebbe avere l'impressione di rischiare di perdere completamente il controllo di questi programmi, ma in realtà la situazione non è per nulla negativa, anzi come vediamo quotidianamente, funziona egregiamente. Ritornando alle caratteristiche di Mike di Heinlein, ogni giorno constatiamo che sono applicabili abbastanza bene ai più recenti modelli IA Generativi (ad eccezione di 1.4 sulla segregazione dei dati).

Ma per poter utilizzare al meglio questi strumenti, dobbiamo avere una diversa percezione di come funzionano e un nuovo approccio al loro utilizzo rispetto a

quanto fatto sinora.

IA Generativa come una Black Box

L'approccio con cui si propone di considerare e valutare i più recenti modelli IA è più vicino ad una valutazione **Black Box** di un programma informatico che a un'analisi tradizionale. Si è nella situazione in cui si conoscono o si possono conoscere bene i dati in ingresso e quelli prodotti dal programma, ma si hanno poche informazioni su come i dati sono elaborati. Su questa Black Box ideale si può solo assumere ad alto livello che:

1. l'elaborazione dei dati è fatta con tecniche probabilistiche
2. i dati memorizzati e le logiche di elaborazione derivano dai dati utilizzati per l'addestramento del modello.

Si può quindi studiare questa Black Box ideale dall'esterno considerando prima cosa riceve in ingresso e poi cosa produce in uscita.

Dati in ingresso per addestramento

Come già descritto, i dati di addestramento insieme all'architettura del modello IA, costituiscono l'elemento principale e fondamentale del modello stesso. Semplificando si possono dividere i modelli in due classi: quelli in cui la fase di addestramento precede ed è distinta dalla fase di generazione, e quelli in cui addestramento e generazione avvengono contemporaneamente o in continua successione. Come sempre, la realtà è abbastanza più complessa, ma per i nostri scopi possiamo semplificare il modello teorico e semplicemente considerare le due fasi, addestramento e generazione, come indipendenti.

Il fatto è che quando i dati in ingresso sono utilizzati per addestrare il modello, questi modificano il modello stesso e vengono memorizzati. Descrivendo il modello IA come una Black Box, non si sa esattamente come i dati sono memorizzati nel modello, ma si può assumere sia che modifichino le logiche di elaborazione, quindi i dati "insegnano" qualche cosa al modello IA, sia che rimangano in memoria in maniera più o meno completa e integra.

Si aggiunga a ciò che per essere addestrati, i modelli IA Generativi richiedono enormi quantità di dati, appunto per essere in grado di sostenere una conversazione o generare dati in uscita di quasi qualsiasi tipo. In altre parole, questi modelli mirano ad avere una conoscenza generica, anche se non approfondita, dello scibile, della storia e delle informazioni umane.

Si deve allora fare estrema attenzione a condividere informazioni riservate o personali o comunque soggette a diritto d'autore o limitazioni di uso e condivisione, per l'addestramento di questi modelli. Infatti la Black Box può utilizzare queste informazioni in qualunque maniera, ad esempio condividendole integralmente con altri utenti, o utilizzandole per generare dei dati che ne derivano.

Come accennato precedentemente, ad oggi non è possibile implementare un sistema di controllo accessi interno alla Black Box in grado di utilizzare i dati memorizzati a seconda dell'utente che sta utilizzando l'applicazione. Quindi qualunque dato utilizzato per l'addestramento di un modello IA pubblico deve essere considerato anch'esso pubblico e non soggetto ad alcuna limitazione.

Questo non vuol dire che non si può utilizzare un modello IA Generativo con dati riservati o soggetti a limitazioni, ma per farlo bisogna implementare il controllo accessi esternamente al modello stesso. Ad esempio, si può partire da un modello pre-addestrato con dati pubblici, crearne un'istanza dedicata sotto il proprio controllo, completare l'addestramento con i propri dati e permettere l'accesso a questa istanza solo a chi è autorizzato ad accedere alle informazioni utilizzate per l'addestramento. In altre parole, bisogna creare una sandbox **intorno** al modello IA piuttosto che all'interno del programma, come invece immaginò Heinlein per Mike (punto 1.4).

Dati in ingresso per generazione

Quando i dati in ingresso sono utilizzati esclusivamente per generare dei dati in uscita, non si presentano i rischi appena descritti perché il comportamento dell'applicazione è assimilabile a quello di una applicazione tradizionale. I dati in ingresso vengono elaborati ma non sono memorizzati dall'applicazione IA, al più possono essere memorizzati dalle applicazioni a supporto dell'applicazione IA, che sono gestite secondo l'approccio tradizionale e permettono

l'implementazione degli usuali processi di controllo accessi, protezione/cifratura dei dati ecc.

Dati prodotti

È necessario fare alcune considerazioni sui dati prodotti da un modello IA Generativo.

La prima è che, come è stato ripetuto già più volte, qualunque dato prodotto non è deciso a priori da un programmatore: non è possibile chiedere ad alcuno perché ha programmato che a uno specifico dato in ingresso corrisponde un certo dato in uscita. In altre parole: non c'è una logica definita da una persona tra un input e il corrispondente output.

Come già indicato, i dati prodotti o generati sono quelli che hanno la più alta "probabilità di similitudine" (da intendersi in senso lato) tra i dati in ingresso e quanto appreso dal modello dai dati di apprendimento. Ad alto livello questo può essere descritto come segue: il modello crea una propria rappresentazione interna dei dati di apprendimento, identifica le "somiglianze" maggiori di questa rappresentazione con i dati in ingresso e utilizza queste "somiglianze" per generare i dati in uscita.

In linea teorica, un dato generato può avere una probabilità del 99,99% di corrispondere al dato in ingresso, o molto minore, ad esempio solo il 30%. Tocca a chi gestisce il modello IA decidere una soglia sotto la quale il modello può non rispondere alla richiesta.

La natura probabilistica dei dati generati può facilmente ingannare gli utilizzatori del modello IA e richiede che chi lo utilizza sia molto accorto nell'interpretare il loro significato. Il modo più semplice per approfondire questo aspetto è tramite un paio di esempi.

Curriculum Vitae di una Persona

Si immagini di chiedere a un modello IA Generativo di fornirci informazioni su di una persona, realmente esistita che si afferma essere stata studente di un premio Nobel. Il modello IA Generativo tipicamente fornisce un breve

Curriculum Vitae della persona, elencando articoli scientifici pubblicati, premi vinti, posizioni ricoperte ecc. Alcuni di questi dati possono essere corretti, altri invece sono **"generati", ma realistici**, e altri ancora possono essere del tutto errati (in alcuni casi chiamati "allucinazioni"). La "logica" del modello IA Generativo può essere descritta come segue: dai dati di addestramento ha appreso che la maggior parte degli studenti di premi Nobel ha avuto una carriera di successo nella ricerca scientifica, vincendo premi e occupando posizioni di rilievo. Inoltre, il premio Nobel indicato appartiene a una specifica area scientifica e il modello IA è a conoscenza di quali sono i principali argomenti di ricerca, i principali premi e posizioni lavorative e di prestigio in questo campo scientifico. Sulla base di questi dati, il modello IA ha quindi identificato quali sono le pubblicazioni scientifiche, i premi e le posizioni (tutti reali o "generati") che **più probabilmente** sono associabili alla persona indicata nei dati in ingresso, e questi sono i dati prodotti in risposta alla richiesta. La quantità di dati esatti prodotti dal modello IA può variare da un Curriculum Vitae totalmente reale, a uno contenente metà di dati reali e metà "generati", a uno totalmente "generato" ovvero plausibile ma non corrispondente al vero, o infine uno contenente alcuni dati del tutto errati; tutto dipende dai dati di addestramento utilizzati e dall'architettura stessa del modello IA.

Selezione di personale e "bias"

I Curriculum Vitae (CV) forniscono un esempio per un altro aspetto dei dati generati da un modello IA (anche non Generativo), che va sotto il nome "*bias*" (ovvero una distorsione o deviazione sistematica rispetto al risultato atteso). Si supponga di addestrare un modello IA con molti CV divisi tra quelli ritenuti validi per l'occupazione richiesta e quelli non validi. Si utilizzi poi l'applicazione IA per una prima selezione di nuovi CV validi. Potrebbe capitare (questo è un esempio estremo riportato solo per evidenziare il possibile comportamento del modello IA) di accorgersi che i CV validi (ovvero con una probabilità di somiglianza ai CV di addestramento validi superiore ad una certa soglia) corrispondono per lo più a persone che si sono laureate in un certo anno e con particolari caratteristiche somatiche, e che CV che non hanno queste caratteristiche ma per il resto sono equivalenti, hanno ottenuto una valutazione inferiore. Come già indicato, i modelli IA sono estremamente bravi a trovare somiglianze in enormi quantità di dati, e in questo caso il modello IA ha identificato l'anno di laurea e alcune caratteristiche somatiche come la caratteristica più comune dei CV di

addestramento validi. L'esempio appena descritto è molto banale, ma il problema del "*bias*" dei modelli IA è molto ricorrente e spesso di difficile soluzione.

Questi semplici esempi rafforzano l'affermazione che alla base del comportamento di un modello IA vi è principalmente una triade: probabilità, architettura, dati di addestramento.

È utile fare qualche ulteriore osservazione sull'utilizzo e i limiti degli attuali modelli IA Generativi.

Etica

Da quanto appena scritto, è difficile oggi immaginare che, oltre a distinguere tra reale e "generato", i modelli IA Generativi possano generare dati che rispettano **sempre** principi connaturati nell'uomo chiamati "etici", quali ad esempio i concetti di giusto e sbagliato o lecito e illecito, che comunque hanno a volte delle diverse interpretazioni tra gli uomini o tra società umane. Ed è anche difficile immaginare che un attuale modello IA Generativo sia in grado di apprendere concetti più fondamentali per l'uomo quali l'istinto di preservazione della persona, di prosecuzione della specie, di protezione della prole. È stata utilizzata la parola "istinto" perché questi sono tratti fondamentali e comuni a tutti gli uomini, e più in generale agli animali, ma difficilmente presenti nei dati di addestramento di un modello IA.

Al contrario, se si addestra un modello IA con dati reperibili in Internet, le informazioni derivanti dai notiziari descrivono una realtà in cui spiccano reati, violenze, violazioni dei diritti umani, guerre ecc. Analizzando solo questi dati ne deriva ad esempio che la probabilità di un atto violento è molto più alta di quella della protezione di una persona debole. Ma la realtà (per fortuna) è assai diversa (anche se abbiamo un ampio margine di miglioramento).

Anche addestrare un modello IA con i manuali di storia e in generale con la documentazione universitaria più completa sull'uomo, la società umana e la storia dell'uomo potrebbe non risolvere realmente il problema. Insegnare i valori "etici" a un modello IA è un grande problema aperto che molto probabilmente richiederà la creazione di appositi dati sintetici per

l'addestramento.

Aritmetica e matematica

Un altro aspetto apparentemente sorprendente degli attuali modelli IA, è la loro difficoltà nell'apprendere l'aritmetica. Ma non è difficile intuire questo aspetto utilizzando quanto sinora descritto: i modelli IA attuali non sono veramente in grado di "ragionare" e quindi imparare autonomamente regole di calcolo esatte, ma correlano probabilisticamente grandi quantità di dati. La "logica" da considerare è più simile a chiedersi qual è la probabilità (derivante dai dati di addestramento utilizzati) che 2+2 sia 3,9 o 4,1 o, ovviamente, 4 senza fare alcun calcolo aritmetico. Questo è altamente contro-intuitivo anche perché gli elaboratori sono basati in hardware proprio sull'esattezza del calcolo matematico e in particolare aritmetico.

D'altra parte, i modelli IA di ultima generazione si stanno dimostrando estremamente utili nel contribuire a risolvere numericamente equazioni matematiche particolarmente complesse. Data una equazione particolarmente complessa e non risolvibile esattamente con le tecniche matematiche note, si cercano soluzioni numeriche necessariamente approssimate. I modelli IA di ultima generazione sono in grado di generare i set di dati numerici con la maggior probabilità di essere una soluzione approssimata al problema tra cui poi va selezionata la soluzione appropriata. In questo modo sono già state trovate nuove molecole chimiche per applicazioni farmaceutiche, identificate debolezze in algoritmi crittografici in corso di sviluppo, ecc.

Utilizzo "Sicuro" dell'IA Generativa

I modelli IA Generativi sono già estremamente utili e potrebbero portare un grande avanzamento nelle capacità, usi e utilità per l'uomo dell'informatica. Ma oggi è necessario utilizzarli essendo ben consci dei loro limiti e delle specifiche modalità di approccio e uso necessari per poterne fare un uso "sicuro". In questa sede è possibile solamente indicare alcuni suggerimenti per poter beneficiare il più possibile dell'uso odierno dell'IA Generativa in modo "sicuro":

1. **Supervisione umana**: i dati generati da un modello IA Generativo devono essere valutati da persone che ne comprendono le logiche, i limiti e che

sono in grado di distinguere quanto è reale, quanto è solamente "realistico" e quanto è del tutto errato.

2. **Dati di addestramento**: bisogna essere ben consapevoli di quali dati sono stati utilizzati per l'addestramento del modello IA e assumere che chiunque utilizza il modello IA è in grado di accedere a tutti i dati di addestramento. Come è stato indicato, oggi non è possibile implementare logiche di controllo accesso all'interno del modello (ma si veda il suggerimento 4), pertanto chi utilizza il modello accede direttamente o indirettamente a tutti i dati di addestramento.
3. **Controllo accessi e filtri sui dati in ingresso**: è necessario quindi implementare il sistema di controllo accessi a un modello IA Generativo al di fuori dello stesso, e affiancare a questo anche un'applicazione di verifica ed eventuale filtro dei dati in ingresso. Quest'ultima è una funzione simile concettualmente alla validazione dei dati in ingresso comune a qualunque applicazione e che, ad esempio per applicazioni Web, previene attacchi quali quelli di "*SQL Injection*". Infatti uno degli scopi è di bloccare attacchi quali "*Prompt Injection*" che hanno come principale fine quello di sovvertire le logiche di utilizzo del modello IA per scopi diversi da quelli autorizzati o per l'estrazione di informazioni riservate.
4. **Addestramento con prefissi**: in aggiunta al controllo accessi e filtro dei dati in ingresso, è anche possibile addestrare specificatamente un modello IA utilizzando dei Prefissi che identificano particolari richieste o utenti; ad esempio, un modello IA può essere addestrato in modo che se la richiesta inizia con il prefisso "Minore" (che identifica il richiedente essere minorenne, o il risultato essere indirizzato a minorenni), il modello IA genera più probabilmente dati appropriati per minorenni. Analogamente si può addestrare il modello IA in modo che certi tipi di dati siano più probabilmente generati o non generati a seconda del richiedente o dell'uso previsto.
5. **Filtri sui dati generati**: come descritto, oggi un modello IA Generativo potrebbe comunque generare dei contenuti non appropriati, errati o che violano la riservatezza, privacy, diritti di autore ecc. Per questo si dovrebbero introdurre dei filtri sui dati generati. Un approccio per implementare questa funzionalità è idealmente simile a quanto si fa ormai da tempo per i filtri Anti-Spam, e spesso si utilizzano degli altri modelli IA addestrati appositamente per identificare i contenuti non leciti.
6. **Monitoraggio**: oltre ai filtri in uscita per evitare la condivisione di dati non

appropriati, errati o non leciti, è bene monitorare in continuazione quanto generato dal modello IA Generativo per identificare eventuali “bias”, “allucinazioni” e altri eventi che possono indicare un abuso, un attacco e ulteriori debolezze del modello.

In ogni caso è molto probabile che nel prossimo futuro, nuove architetture e nuove tecniche di addestramento saranno in grado di migliorare grandemente gli attuali modelli di Intelligenza Artificiale.

Articolo a cura di **Andrea Pasquinucci**

Profilo Autore

Andrea Pasquinucci

PhD CISA CISSP
Consulente freelance in sicurezza informatica: si occupa prevalentemente di consulenza al top management in Cyber Security e di progetti, governance, risk management, compliance, audit e formazione in sicurezza IT.

Altri Articoli

- Machine Learning, “Deep Fake” ed i rischi in un mondo iperconnesso
- Intelligenza Artificiale / Machine Learning: tra Complessità e Sicurezza
- Sicurezza, Hardware e Confidential Computing – Parte 2
- Sicurezza, Hardware e Confidential Computing - Parte 1