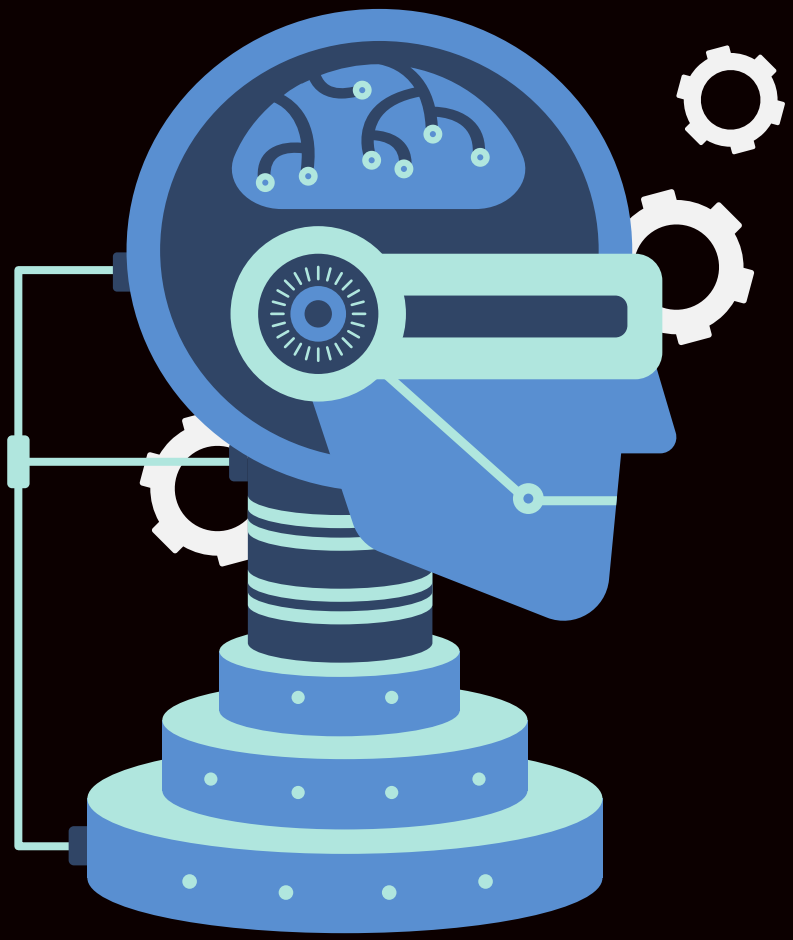
GENERATIVE ARTIFICIAL INTELLIGENCE:

PUNTI DI FORZA, RISCHI E CONTROMISURE

Vincenzo Calabrò
Funzionario alla Sicurezza CIS (Ministero dell'Interno)
e Digital Forensics Analyst

WHITEPAPER 11/2024



ICT Security MAGAZINE

INDICE

ICT Security MAGAZINE

CYBER CRIME CONFERENCE

16-17 APRILE 2025

AUDITORIUM DELLA TECNICA, ROMA

Iscriviti alla newsletter di ICT Security Magazine
per conoscere l'agenda e partecipare alla
13ª Edizione della Cyber Crime Conference

ABOUT THE AUTHOR

Vincenzo Calabrò

Funzionario alla Sicurezza CIS (Ministero dell'Interno) e Digital Forensics Analyst

È laureato in Ingegneria Informatica ed in Sicurezza Informatica presso le Università di Roma La Sapienza e di Milano. Ha indirizzato la sua formazione nei settori della Cyber Security e Digital Forensics ottenendo i diplomi di perfezionamento in Data Protection e Data Governance; Criminalità Informatica e Investigazioni Digitali e Big Data, Artificial Intelligence.

Ha, altresì, conseguito l'Advanced Cybersecurity Graduate Certificate alla School of Engineering della Stanford University; Professional Certificates in Information Security; Incident Response Process; Digital Forensics e Cybersecurity Engineering and Software Assurance presso il Software Engineering Institute della Carnegie Mellon University.

Dal 1992 è nei ruoli del Ministero dell'Interno ove ricopre lincarico di Funzionario alla Sicurezza CIS. In tale veste contribuisce alla valutazione dei rischi cyber, all'implementazione delle misure di sicurezza e la risoluzione di incidenti informatici. Inoltre, offre consulenza tecnica nel campo della Digital Forensics per l'Autorità giudiziaria, la Polizia giudiziaria e gli Studi legali.

Dal 2017 è Professore a contratto di Tecnologie per la Sicurezza Informatica presso alcune Università ove sviluppa le tematiche di Attack and Defense Strategies quali il penetration testing, la risk analysis, l'information security assessment, l'incident response e la digital forensics. Infine, è Autore di alcuni articoli e saggi sui temi della Sicurezza Informatica e dell'Informatica Giuridica consultabili su https://www.vincenzocalabro.it

Premessa

Il 2024 è l'anno in cui la Generative Artificial Intelligence (GenAI) supera la fase di *"Peak of Inflated Expectations"* (vd. il modello Hype Cycle di Gartner), ogni giorno vengono progettati, sviluppati e rilasciati sistemi di AI per qualsiasi ambito e applicazione; perciò, possiamo tranquillamente suddividere il mondo dei fruitori dell'AI in tre categorie: 1) quelli che la impiegano in maniera proattiva, 2) coloro che la governano e, infine, 3) chi la subisce.

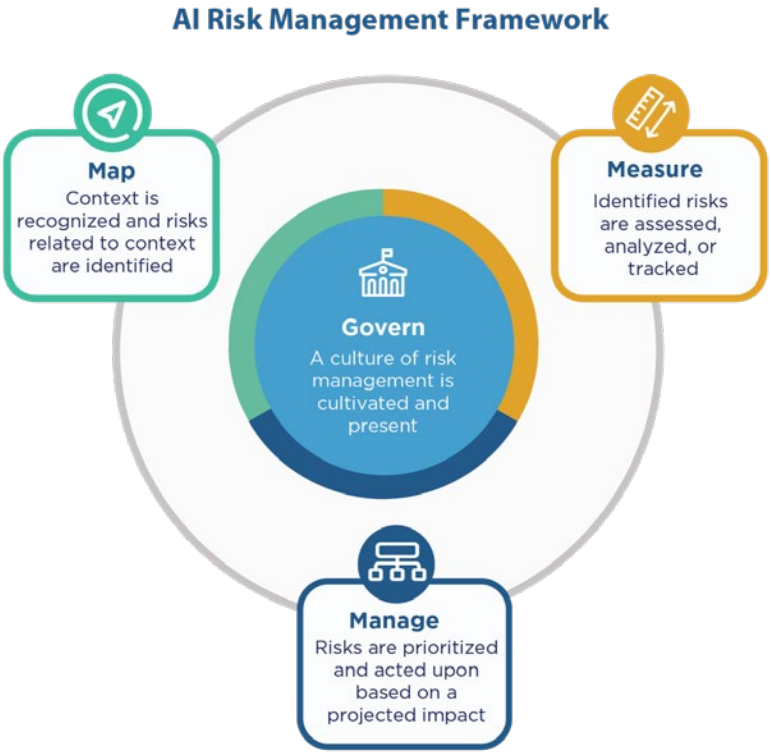
Come accade per tutte le innovazioni, anche la GenAI avrà il suo *"Trough of Disillusionment"* poiché, da un lato, le aspettative degli utenti sono ancora molto elevate e, dall'altro, l'implementazione non sarà in grado di raggiungere i risultati sperati, sia in termini di performance, che di ROI. Per ridurre questo impatto negativo, che immancabilmente si genererà sugli stakeholders, è opportuno conoscere e studiare non solo i punti di forza dell'AI, ma anche le fragilità e le vulnerabilità che sono ancora presenti. Questo approccio consentirà di creare e utilizzare prodotti e servizi dell'AI *"Safe, Secure and Trustworthy"*.

In altri termini, l'aumento di volume e complessità dei progetti basati sull'intelligenza artificiale fa entrare in gioco variabili che inizialmente non erano state considerate, come accade per tutte le innovazioni tecnologiche, causando un disallineamento tra i risultati attesi e quelli reali; quindi, occorre prestare sempre più attenzione alla governance, ai rischi, all'ownership, alla sicurezza e ai relativi metodi di mitigazione per ridurre questo divario.

Introduzione

La prassi ci insegna che ogni lancio di prodotti o soluzioni innovative, compresi i sistemi basati sull'intelligenza artificiale, tra cui i modelli che sfruttano le reti neurali come il Machine Learning (ML) e la Generative Artificial Intelligence (GenAI), vive una fase iniziale di entusiasmo, dovuta al desiderio di essere i primi a uscire sul mercato, in cui è facile trascurare le fragilità e le vulnerabilità che rendono questi modelli suscettibili di errori, violazioni della confidenzialità e altri tipi di difetti o anomalie. In realtà, le fragilità e le vulnerabilità del ML e del GenAI, tra cui i Large Language Models (LLM), generano rischi con caratteristiche diverse da quelle tipi-

camente considerate nell'analisi del software o della sicurezza informatica, pertanto, le fasi di progettazione e valutazione dei sistemi basati sull'intelligenza artificiale, e dei relativi workflow, meritano un'attenzione speciale. In particolare, lo sviluppo di definizioni adeguate alla sicurezza e alla protezione dei sistemi basati sull'intelligenza artificiale rappresenta una sfida significativa nell'ambito della progettazione e della valutazione dei sistemi. Se considerassimo il ruolo dell'intelligenza artificiale in domini applicativi critici, in cui la mission è incentrata sull'efficacia, la sicurezza, la protezione e la resilienza del sistema, questa minaccia si amplificherebbe. A riguardo vedasi le raccomandazioni del "*NIST Artificial Intelligence Risk Management Framework (RMF)*"[1] e "*NIST Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*"[2] per individuare, mitigare e gestire i rischi connessi.



AI Risk Management Framework (fonte: NIST.gov)

1 *Artificial Intelligence Risk Management Framework (RMF), https://doi.org/10.6028/NIST.AI.100-1, NIST, 2023*

2 *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, https://doi.org/10.6028/NIST.AI.600-1, NIST, 2024*

Questo testo si concentra sull'intelligenza artificiale applicata ai sistemi critici in cui l'affidabilità, basata su evidenze verificabili, rappresenta il fattore essenziale per il consenso operativo.

L'analisi dell'argomento si sviluppa in sei parti:

- La prima parte spiega i concetti fondamentali e il funzionamento dell'Intelligenza Artificiale Generativa (GenAI) e presenta i vantaggi, le sfide, le limitazioni e i rischi che si porta dietro.
- La seconda parte introduce il problema della sicurezza e della protezione dell'AI e prova a rispondere alle seguenti domande: perché è difficile garantire la sicurezza e la protezione dell'AI? quali sono i concetti di sicurezza e protezione specifici per l'intelligenza artificiale basata sulle reti neurali? quali sono le sfide dell'intelligenza artificiale nello sviluppo di sistemi sicuri e protetti? quali sono i limiti di affidabilità e perché questi limiti sono fondamentali?
- La terza parte affronta il problema della confidenzialità, dell'integrità e della governance dei sistemi di AI e tenta di rispondere alle seguenti domande: quali sono i rischi specifici dell'intelligenza artificiale - inclusi quelli associati alla confidenzialità, integrità e governance - con e senza avversari? quali sono le superfici di attacco e quali tipi di mitigazioni sono in fase di sviluppo e impiego per ridurre queste criticità?
- La quarta parte approfondisce il rischio dell'AI e prova a rispondere alle seguenti domande: come possiamo definire la fase di Test and Evaluation (T&E) specifica per l'intelligenza artificiale? È possibile elaborare un framework di risk management per l'intelligenza artificiale? la progettazione dell'AI può affrontare le sfide a breve termine e, a riguardo, come è possibile sfruttare le opportunità dell'ingegneria del software e della sicurezza informatica?
- La quinta parte illustra due case study riguardanti le principali minacce che affliggono i Large Language Models: i *bias* e il *data poisoning*; in particolare, un metodo per realizzare l'auditing dei bias e una serie di raccomandazioni per eliminare i dati errati (compresi i bias) da un data set di ML. Inoltre, vengono indicate una serie di raccomandazioni per misurare e rendere affidabile un sistema di intelligenza artificiale.
- La sesta parte esamina i nuovi modelli di AI e tenta di rispondere alle seguenti domande: quali sono i vantaggi nel guardare oltre i modelli basati sulle reti puramente neurali dell'AI moderna verso approcci ibridi? quali sono gli esempi attuali che illustrano i vantaggi potenziali e in che modo questi approcci possono consentire di andare oltre i limiti dell'AI attuale? quali sono le prospettive nel breve e nel lungo periodo?

Cos'è la Generative AI

La Generative AI, talvolta chiamata GenAI, è un tipo di intelligenza artificiale (AI) in grado di creare contenuti originali, come testi, immagini, video, audio o codice software, in risposta al prompt o alla richiesta di un utente. Per rispondere a queste richieste, la GenAI utilizza sofisticati modelli di machine learning, chiamati modelli di deep learning, ovvero algoritmi in grado di simulare i processi di apprendimento e decisionale del cervello umano. Questi si basano sull'identificazione e la codifica di modelli e relazioni in enormi quantità di dati, che sono utilizzati per comprendere le richieste o le domande in linguaggio naturale degli utenti e rispondere con nuovi contenuti pertinenti.

L'AI ha rappresentato uno dei principali argomenti tecnologici degli ultimi dieci anni, ma l'AI generativa, e in particolare l'avvento di ChatGPT nel 2022, ha portato l'AI alla portata di tutti e ha lanciato un'ondata di innovazione senza precedenti. L'AI generativa offre enormi vantaggi in termini di produttività per tutti, individui e organizzazioni, e, sebbene presenti anche sfide e rischi sostanziali, la ricerca continua a esplorare i modi in cui questa tecnologia può migliorare i workflow, potenziare i prodotti e ottimizzare i servizi. Secondo una ricerca realizzata nel 2023 da McKinsey, un terzo delle organizzazioni utilizza regolarmente l'AI generativa in almeno una funzione aziendale.[3] Gartner prevede che entro il 2026 oltre l'80% delle organizzazioni avrà implementato applicazioni di AI generativa oppure utilizzerà interfacce di programmazione delle applicazioni (API) di AI generativa.[4]

Excursus storico

Il termine Intelligenza Artificiale Generativa è divenuto popolare negli ultimi anni, ma l'intelligenza artificiale fa parte della quotidianità da decenni e la tecnologia di AI generativa di oggi si basa sulle scoperte dell'ap-

3 AA.VV., The state of AI in 2023: Generative AI's breakout year, https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year, McKinsey, 2023

4 Gartner Says More Than 80% of Enterprises Will Have Used Generative AI APIs or Deployed Generative AI-Enabled Applications by 2026, https://www.gartner.com/en/newsroom/press-releases/2023-10-11-gartner-says-more-than-80-percent-of-enterprises-will-have-used-generative-ai-apis-or-deployed-generative-ai-enabled-applications-by-2026, Gartner, 2024

prendimento automatico risalenti all'inizio del XX secolo. Cerchiamo di comprendere le tappe fondamentali:

- **1964**: Joseph Weizenbaum, ricercatore-informatico del MIT, sviluppa ELIZA, un'applicazione di elaborazione del linguaggio naturale basata su testo. ELIZA, è stato sostanzialmente il primo chatbot (all'epoca chiamato "chatterbot") basato su uno script di pattern-matching in grado di rispondere agli input, forniti sottoforma di testo in linguaggio naturale, con risposte testuali empatiche.
- **1999**: Nvidia lancia GeoForce, la prima unità di elaborazione grafica. Originariamente sviluppata per fornire una grafica in movimento fluida per i videogiochi, le GPU sono diventate di fatto la piattaforma per lo sviluppo di modelli AI e il mining di criptovalute.
- **2004**: Google lancia la funzione di completamento automatico, che si attiva quando gli utenti inseriscono i termini di ricerca, in grado di generare potenziali parole o frasi successive. L'algoritmo è basato su una catena Markov, un modello matematico sviluppato nel 1906.
- **2013**: Compaiono i primi "*Variational Autoencoders*" (VAE), dei modelli di reti neurali che permettono di implementare un modello generativo.
- **2014**: Ian Goodfellow, ricercatore presso l'Università di Montreal, introduce le prime "*Generative Adversarial Networks*" (GAN) e i *"Diffusion Models"*[5], due architetture per addestrare un modello generativo di AI.
- **2017**: Un gruppo di ricerca formato da Ashish Vaswani, un team di Google Brain e un gruppo della University of Toronto pubblicano "*Attention is All You Need*"[6], un documento che illustra i principi dei modelli di trasformazione (c.d. "*Transformer*") in grado di abilitare i potenti "*foundation model*" e strumenti di AI generativa attuali.
- **2019**: OpenAI implementa i suoi modelli linguistici di grandi dimensioni GPT (*Generative Pretrained Transformer*), GPT-2 e GPT-3.
- **2022**: OpenAI presenta ChatGPT, un front-end di GPT-3 che genera frasi complesse, coerenti e contestualizzate, e contenuti in risposta ai prompt degli utenti finali.

La notorietà e la popolarità acquisite da ChatGPT ha dato il via ad una serie di progetti finalizzati allo sviluppo

5 *AA.VV., Generative Adversarial Networks, https://arxiv.org/abs/1406.2661, Universite de Montreal, 2014*
6 *AA.VV., Attention Is All You Need, https://arxiv.org/abs/1706.03762, Google, 2017*

di altre AI generative e il rilascio di prodotti a un ritmo incalzante, fino a giungere alle attuali release di Google Bard (ora Gemini), Microsoft Copilot, Bedrock di Amazon, Ernie Bit di Baidu, Pangu-Σ di Huawei, Claude di Anthropic, xAI di Elon Musk, Jais in lingua araba, Poe di Quora, IBM watsonx.ai, e ai modelli linguistici di grandi dimensioni (LLM) open source come Llama-2, di Meta, o Dolly 2.0 di Databricks e OpenChat AI GitHub.

Come opera la Generative AI

Normalmente il ciclo di vita della Generative AI si sviluppa in tre fasi:

- **Formazione:** la fase in cui si crea un "*foundation model*"[7] che possa servire come base per più applicazioni di AI di nuova generazione.
- **Ottimizzazione:** la fase in cui si adatta il foundation model a una specifica applicazione AI di nuova generazione.
- **Generazione, valutazione e improvement:** la fase in cui si valutano i risultati dell'applicazione della GenAI per migliorarne la qualità e l'accuratezza.

Formazione

La GenAI si fonda su un foundation model, un modello di deep learning che funge da schema per diversi tipi di applicazioni di AI generativa. I foundation model più comuni sono i modelli linguistici di grandi dimensioni (LLM), sfruttati nelle applicazioni per generare testo, ma esistono foundation model per generare immagini, video, audio e musica, oltre a foundation model multimodali in grado di supportare simultaneamente diversi tipi di contenuti.

7 *Foundation model: Nell'ambito dell'Intelligenza Artificiale, il termine modello identifica un insieme strutturato di algoritmi e parametri che permettono di eseguire specifici compiti di apprendimento automatico. I modelli sono addestrati tramite l'analisi e l'elaborazione di dati, al fine di identificare e apprendere schemi o relazioni tra di essi. Un foundation model rappresenta un tipo specifico e avanzato di modello di Intelligenza Artificiale. Mentre i modelli generici sono progettati e addestrati per svolgere compiti specifici e ben definiti e possono essere addestrati su set di dati di dimensioni variabili, a seconda del compito che dovranno svolgere; i foundation models, invece, sono addestrati su enormi quantità di dati e con moltissimi parametri. Ciò permette loro di svolgere una varietà di compiti più ampia rispetto ai modelli tradizionali.*

Per realizzare un foundation model occorre addestrare un algoritmo di deep learning su enormi volumi di dati grezzi, non strutturati e non etichettati, spesso raccolti dalla rete internet o da altre fonti di dati. Durante la fase di training, l'algoritmo esegue e valuta milioni di test c.d. "*fill in the blank*" (riempimento dello spazio), cercando di prevedere l'elemento successivo in una determinata sequenza, per esempio la parola successiva in una frase, l'elemento successivo in un'immagine, il comando successivo in una riga di codice, e, contestualmente, effettua un auto-tuning per ridurre al minimo la differenza tra le previsioni e i dati reali (o il risultato "corretto").

Il risultato di questo ciclo di training è una rete neurale di parametri, ovvero una serie di rappresentazioni codificate di entità, pattern e relazioni nei dati, in grado di generare contenuti in modo autonomo in risposta agli input o ai prompt ricevuti.

Questo processo si caratterizza per l'alta intensità di calcolo, un elevato dispendio in termini di tempo e costi: infatti richiede la presenza di migliaia di unità di elaborazione grafica (GPU) sottoforma di cluster e settimane di elaborazione. Fortunatamente sono disponibili progetti di foundation model open source, come Llama-2 di Meta, che consentono agli sviluppatori di GenAI di saltare questo passaggio e i relativi costi.

Ottimizzazione

Un foundation model è generalista, ovvero sa molte cose su molti tipi di contenuti, ma spesso non è in grado di generare tipi specifici di output con la precisione o la qualità desiderate. Per questo motivo è opportuno che il modello debba essere ottimizzato seguendo un'attività di generazione dei contenuti specifica. Questo obiettivo può essere raggiunto in vari modi:

- **Attraverso una messa a punto del modello**. Questa metodologia comporta l'alimentazione del modello con dati specifici dell'applicazione di generazione di contenuti, cioè, deve comprendere le domande o i prompt che molto probabilmente l'applicazione riceverà e le corrispondenti risposte corrette nel formato desiderato. Per esempio, se un team di sviluppo sta tentando di creare un customer service chatbot, creerà centinaia o migliaia di documenti contenenti domande etichettate sul servizio clienti con le risposte corrette e li invierà al modello.
- **Tramite un apprendimento di rinforzo con feedback umano (RLHF**). La metodologia di RLHF (Reinforcement Learning from Human Feedback) prevede che gli utenti umani rispondano ai contenuti generati

con valutazioni che, successivamente, il modello utilizza per aggiornare sé stesso e ottenere maggiore precisione o rilevanza.

Generazione, valutazione e improvement

Di norma, sia gli sviluppatori che gli utenti valutano continuamente i risultati delle app di GenAI che utilizzano e, di conseguenza, ottimizzano ulteriormente il modello per ottenere una maggiore precisione o pertinenza.

Esiste un'altra alternativa per ottimizzare le prestazioni di un'applicazione di AI: la "*retrieval augmented generation*" (RAG). La RAG è un framework in grado di estendere il foundation model al fine di utilizzare un set di fonti rilevanti, al di fuori dei dati utilizzati nel training, capaci di integrare e perfezionare i parametri o le rappresentazioni nel modello originale. A rigurado, la RAG garantisce che un'app di AI generativa abbia sempre accesso alle informazioni più aggiornate e contestualizzate. Inoltre, le fonti aggiuntive a cui si accede tramite RAG, a differenza dei dataset del foundation model originale, sono trasparenti e verificabili.

Architetture dei modelli di Generative AI

Dal 2013 abbiamo assistito allo sviluppo di modelli di AI veramente generativa, ovvero modelli di "*deep learning*" in grado di creare autonomamente contenuti su richiesta. Le principali architetture evolute durante questo periodo includono:

- **Variational Autoencoders (VAE):** rappresentano modelli che hanno consentito di effettuare scoperte rivoluzionarie nel riconoscimento delle immagini, nell'elaborazione del linguaggio naturale e nel rilevamento delle anomalie.
- **Generative Adversarial Networks (GAN) & Diffusion Models:** sono architetture che hanno migliorato l'accuratezza delle applicazioni precedenti e hanno reso possibile le prime applicazioni di AI per la generazione di immagini da foto reali.
- **Transformers:** rappresentano l'architettura dei modelli di deep learning alla base dei principali foundation model e delle soluzioni di AI generativa di oggi.

Variational Autoencoder

Un "*autoencoder*" è un modello di deep learning composto da due reti neurali connesse tra loro: una che codifica (o comprime) un'enorme quantità di dati, non strutturati e non etichettati, in parametri per il training e un'altra che decodifica tali parametri per ricostruire il contenuto. Tecnicamente, gli *autoencoder* sono in grado di generare nuovi contenuti, ma sono utilizzati prevalentemente per la codifica, l'archiviazione o il trasferimento dei dati e, infine, per la decodifica degli stessi e la generazione di contenuti di alta qualità.

Nel 2013 sono stati introdotti i *"Variational Autoencoder"* (VAE) che, oltre a codificare i dati come un autoencoder, sono in grado di decodificare i dati in diverse varianti di contenuto. Questa peculiarità ha permesso di addestrare i VAE a generare varianti per raggiungere un particolare obiettivo e, al contempo, consentire di ottenere nel tempo contenuti più accurati e ad alta fedeltà. Le prime applicazioni dei VAE includevano il rilevamento delle anomalie (per esempio, nella diagnostica per immagini) e la generazione del linguaggio naturale.

Generative Adversarial Network

Anche le *"Generative Adversarial Network"* (GAN), introdotte nel 2014, comprendono due reti neurali: una denominata Generatore, adibita a generare nuovi contenuti, e una Discriminatore, destinata a valutare l'accuratezza e la qualità dei dati generati. Questi algoritmi antagonisti inducono il modello a generare output di qualità sempre più elevata.

Solitamente le GAN vengono utilizzate per generare immagini e video, ma sono in grado di originare contenuti realistici e di alta qualità anche in altri domini. Si sono dimostrate particolarmente efficaci in determinate attività, come il trasferimento di stile (per esempio la modifica dello stile di un'immagine: da una foto a uno schizzo a matita) e l'aumento dei dati (creazione di nuovi dati per aumentare le dimensioni e la diversità di un data set di training).

Diffusion Model

Un *"Diffusion Model"*, anch'essi introdotti nel 2014, funziona in maniera differente. Questo modello, prima di realizzare il training, aggiunge rumore ai dati per renderli casuali e irriconoscibili, successivamente, addestra

l'algoritmo a disperdere iterativamente il rumore per rivelare l'output desiderato.

I Diffusion Models richiedono più tempo per l'addestramento rispetto ai VAE o alle GAN, ma offrono un controllo dell'output più preciso; infatti, sono particolarmente sfruttati negli strumenti di generazione di immagini di alta qualità. DALL-E, lo strumento di generazione di immagini di Open AI, è guidato da un diffusion model.

Transformer

I "*Transformer*", documentati per la prima volta in un articolo del 2017 pubblicato da Ashish Vaswani e altri, trasformano il paradigma encoder-decoder per realizzarne un'evoluzione, da un lato, del modo in cui i *foundation model* vengono addestrati e, dall'altro, nella qualità e nella gamma di contenuti che possono produrre. Questi modelli sono alla base dei principali strumenti di AI generativa, tra cui ChatGPT e GPT-4, Copilot, BERT, Bard e Midjourney.

Per determinare quali dati, all'interno di una sequenza, sono più importanti e su quali occorre concentrarsi, i transformer utilizzano un concetto chiamato "*attention*" in grado di:

- elaborare contemporaneamente intere sequenze di dati (ad es. frasi invece di parole singole);
- acquisire il contesto dei dati all'interno della sequenza;
- codificare i dati di training in incorporamenti (denominati anche iperparametri) che rappresentano i dati e il loro contesto.

Oltre a consentire un addestramento più rapido, i transformer sono performanti nel l'elaborazione del linguaggio naturale ("*Natural Language Processing*", NLP) e nella sua comprensione ("*Natural Language Understanding",* NLU) e sono in grado di generare sequenze di dati più lunghe (ad esempio, non solo risposte a domande, ma anche poesie, articoli o documenti) con maggiore precisione e qualità rispetto ad altri modelli di AI generativa. I modelli transformer possono essere addestrati o ottimizzati anche per l'uso di tools, ad esempio un'applicazione di fogli di calcolo, un programma di disegno, o per generare contenuti in un formato particolare.

Peculiarità della Generative AI

Cosa può creare

L'AI generativa può essere utilizzata per creare molti tipi di contenuti in domini diversi.

- **Testo**: I modelli generativi, soprattutto quelli basati sui *trasformer*, possono generare testi coerenti e contestualmente rilevanti, come delle istruzioni, della documentazione, una brochure, un'e-mail, dei testi per siti web, blog, articoli, delle relazioni, dei documenti e, persino, contenuti creativi. Possono eseguire compiti di scrittura ripetitivi o noiosi (per esempio, la stesura di riassunti di documenti o meta descrizioni di pagine web), consentendo agli editor di dedicarsi a lavori più creativi e di maggior valore.
- **Immagini e video**: La generazione di immagini, come DALL-E, Midjourney e Stable Diffusion, può realizzare immagini realistiche, oppure opere d'arte originali, ed eseguire il trasferimento di stili, la traduzione image-to-image e altre attività di modifica o miglioramento delle immagini. Gli strumenti GenAI video possono creare animazioni a partire da istruzioni di testo e possono applicare effetti speciali a video esistenti in maniera più rapida ed economica rispetto ai metodi tradizionali.
- **Suono, parole e musica**: I modelli generativi possono sintetizzare contenuti vocali e audio per farli sembrare reali alle chatbot e agli assistenti digitali a comando vocale, consentire la narrazione degli audiolibri e di altre applicazioni basate sull'audio. La stessa tecnologia può generare musica originale in grado di imitare la struttura e l'audio delle composizioni professionali.
- **Codice software**: La GenAI può generare codice di programmazione originale, effettuare il completamento automatico dei frammenti di codice, tradurre tra diversi linguaggi di programmazione e riassumere le funzionalità del codice. Consente di ottenere prototipi, eseguire rapidamente il debug delle applicazioni e realizzare sessioni di testing, offrendo al contempo un'interfaccia in linguaggio naturale.
- **Design e arte**: I modelli di AI generativa possono generare opere d'arte digitali e di design uniche o assistere nella progettazione grafica. Le applicazioni includono la generazione dinamica di ambienti, personaggi o avatar ed effetti speciali per realizzare simulazioni virtuali e videogiochi.
- **Simulazioni e artefatti**: I modelli di AI generativa possono essere addestrati per generare dati o strutture

basate su dati reali o virtuali. Per esempio, l'AI generativa viene applicata nella scoperta di farmaci per generare strutture molecolari con le proprietà desiderate, contribuendo alla progettazione di nuovi composti farmaceutici.

Vantaggi

Il principale beneficio, e forse quello più scontato, dell'AI generativa è la maggiore efficienza. La sua capacità di generare contenuti e risposte alle domande consente di accelerare o automatizzare le attività ad alta intensità di lavoro, ridurre i costi e guadagnare tempo da dedicare ad attività di maggior valore.

L'AI generativa offre ulteriori vantaggi in questi ambiti:

- **Creatività migliorata**: Gli strumenti di Gen AI possono ispirare la creatività attraverso il brainstorming automatizzato, perché gli consente di generare più versioni inedite dei contenuti. Queste varianti possono servire come punti di partenza o riferimenti per superare il blocco creativo di scrittori, artisti, designer e altri tipi di creatori.
- **Processo decisionale migliorato (e più rapido):** L'AI generativa è predominante nell'analisi di grandi insiemi di dati, nell'identificazione degli schemi e nell'estrazione di insight significativi, per poi generare ipotesi e suggerimenti basati su questi insight e supportare i dirigenti, gli analisti, i ricercatori e tutti gli altri professionisti, nel prendere decisioni più corrette e basate sui dati.
- **Personalizzazione dinamica**: L'AI generativa può analizzare le preferenze e la cronologia degli utenti per creare contenuti personalizzati in tempo reale, rendendo l'esperienza dell'utente più personalizzata e coinvolgente.
- **Disponibilità continua**: L'AI generativa funziona in maniera costante e fornisce una disponibilità continua per le attività come le chatbots utilizzate per erogare l'assistenza ai clienti e le risposte automatiche.

Casi d'uso

Di seguito sono riportati, solo a titolo esemplificativo, alcuni casi d'uso dell'AI nel settore business, per comprendere quali benefici possono portare nel breve periodo. In tal senso, man mano che le tecnologie si svi-

lupperanno ulteriormente e le organizzazioni incorporeranno questi tools nei loro workflow, assisteremo alla scoperta di nuove e altrettante applicazioni.

- **User experience (UX)**: Il settore marketing delle aziende può trarre grossi benefici, in termini di tempo e quantità di contenuti, utilizzando strumenti di AI per redigere i testi dei blog, delle pagine web, della documentazione, dell'e-mail e altro ancora. Inoltre, le soluzioni di AI generativa sono in grado di produrre testi e immagini fortemente personalizzate e in tempo reale, in base a quando, dove e a chi deve essere erogato l'annuncio. Inoltre, può alimentare le chatbots e gli agent virtuali per fornire risposte personalizzate e, persino, avviare attività per conto del cliente. Tutto ciò rappresenta un avanzamento significativo rispetto alla precedente generazione di modelli di AI conversazionali formati su dati limitati e sfruttati per attività molto specifiche.
- **Sviluppo software e aggiornamento delle applicazioni**: Gli strumenti per la generazione del codice basati sull'AI generativa sono in grado di automatizzare e accelerare il processo di scrittura di nuovo codice. Questa caratteristica può essere sfruttata anche per accelerare drasticamente l'aggiornamento delle applicazioni, perché consentirebbe di automatizzare gran parte della codifica ripetitiva necessaria per rinnovare le applicazioni legacy e distribuire in ambienti di cloud ibrido.
- **Attività digitale**: L'AI generativa è in grado di redigere o rivedere rapidamente contratti, fatture, bollette e altri documenti digitali, in modo tale che i dipendenti, che la utilizzano o la gestiscono, possano concentrarsi su attività di concetto.
- **Scienza, ingegneria e ricerca**: I modelli di AI possono aiutare i ricercatori e gli ingegneri a proporre nuove soluzioni per problemi complessi. Nel settore sanitario, ad esempio, i modelli generativi possono essere applicati per sintetizzare le nuove molecole o migliorare la lettura delle immagini degli esami.

Sicurezza e protezione della Generative AI

Tassonomia dei rischi

In questo paragrafo si analizza il tema della sicurezza e della protezione nel contesto dell'AI, applicata allo sviluppo di sistemi critici, attraverso l'esame di specifici esempi di fragilità e vulnerabilità dell'AI generativa. I rischi a cui è esposta la GenAI sono organizzati seguendo una tassonomia analoga agli attributi di confidenzialità, integrità e disponibilità (CIA) adoperati nel contesto dei rischi informatici:

- **rischi di confidenzialità**: si palesano nel caso in cui i risultati di un modello di AI rivelano dati di input che i progettisti avevano intenzione di mantenere riservati.
- **rischi di integrità**: si presentano nel caso in cui i risultati di un modello di AI sono errati, involontariamente o tramite manipolazione deliberata da parte degli avversari.
- **rischi di governance**: si rivelano nel caso in cui i risultati di un modello di AI, o l'utilizzo di tale modello in un sistema, possono avere impatti negativi nel contesto delle applicazioni, spesso anche quando i risultati del modello sono corretti rispetto alla formazione.

Partiamo dall'assunto che la gestione del rischio per l'AI, compresa la modellazione e la valutazione, è distinguibile in tre livelli:

1. le capacità del "*core AI*" dei singoli modelli di reti neurali,
2. le scelte effettuate su come tali capacità siano incorporate nell'ingegneria dei sistemi basati sull'AI,
3. come tali sistemi siano integrati in workflow operativi incentrati sulle applicazioni.

Questi workflow possono includere sia applicazioni autonome, che applicazioni abilitate a interagire con soggetti umani. Questa ampia flessibilità è importante perché l'intelligenza artificiale generativa può portare non solo aumenti significativi della produttività e dell'efficacia della mission all'interno di processi organizzativi consolidati, ma può anche generare nuove capacità basate sulla reingegnerizzazione delle attività sul posto

di lavoro incentrate sulla mission e sull'operatività.

Considerazioni sulla Generative AI

La natura stocastica dei modelli di GenAI, combinata con una quasi totale assenza di trasparenza rispetto all'interrogazione e all'analisi, rende complessa la loro descrizione, il testing, l'analisi e il monitoraggio. Ciò che si percepisce come somigliante tra i vari input di un modello, non corrisponde necessariamente al modo in cui risponde il modello. Durante l'addestramento, le differenziazioni possono essere effettuate in base a dettagli che normalmente consideriamo accidentali. Un esempio noto è quello in cui il lupo è distinto dagli altri cani non per la morfologia, ma perché c'è neve sullo sfondo, come rivelato dalle mappe di salienza[8]. Infatti, la metrologia[9] della GenAI è appena agli inizi. Un altro esempio rappresentativo dell'aleatorietà del ML è quello che riguarda l'autonomia dei veicoli. In questo caso, la combinazione tra la scarsa capacità di valutazione dei test e le pressanti direttive aziendali ha portato alla produzione di intere flotte di autoveicoli a guida autonoma e al successivo ritiro dal mercato a causa di comportamenti inaspettati. Inoltre, sono stati individuati diversi bias negli algoritmi predittivi applicati alla stipula di contratti di credito, oppure al reclutamento del personale e all'elaborazione delle richieste di rimborso sanitario. Questi sono i motivi per cui è facilmente realizzabile l'*"adversarial machine learning"*[10].

Analizziamo i principali fattori che caratterizzano i sistemi basati sul GenAI.

Dal punto di vista della mission

Molto spesso i modelli di intelligenza artificiale generativi, addestrati sui dati, sono inclusi all'interno di "*mis-*

8 Le mappe di salienza sono stabilite sulla base di proprietà visive come le discontinuità dell'immagine. Così, per esempio, le discontinuità cromatiche, un oggetto in movimento su uno sfondo statico o le zone più luminose spiccano e catturano l'attenzione.

9 La metrologia è la scienza che si occupa della misurazione e delle sue applicazioni.

10 L'Adversarial machine learning (Apprendimento automatico antagonistico) è una serie di tecniche volte a compromettere il corretto funzionamento di un sistema informatico che faccia uso di algoritmi di apprendimento automatico, tramite la costruzione di input speciali in grado di ingannare tali algoritmi: più nello specifico, lo scopo di tali tecniche è quello di causare la classificazione errata in uno di questi algoritmi.

sion system"[11] sottoforma di componenti o servizi subordinati e, come già visto, spesso questi sistemi rappresentano dei componenti di workflow operativi che supportano un'applicazione all'interno di un determinato processo. Per tanto, affinché si possano misurare e valutare, occorre comprendere tutti e tre i livelli: componente, sistema e workflow.

Per esempio, i problemi dei "*bias*" (preconcetto/pregiudizio/parzialità) possono essere il risultato della mancata corrispondenza tra i dati utilizzati per addestrare il modello con i dati reali di input. Nel contesto del "*Test and Evaluation (T&E)*" significa che è fondamentale caratterizzare e valutare i tre livelli di classificazione indicati in precedenza:

1. le caratteristiche delle capacità di intelligenza artificiale integrate,
2. il modo in cui tali capacità vengono utilizzate nei sistemi basati sull'intelligenza artificiale,
3. il modo in cui tali sistemi sono destinati a essere integrati nei workflow operativi.

L'UK National Cyber Center ha pubblicato le "*Guidelines for secure AI system development*"[12], a cui ha aderito anche l'Agenzia per la cybersicurezza nazionale, che si concentrano sulla sicurezza, intesa come resilienza, privacy, correttezza ed affidabilità, della progettazione, dello sviluppo, della distribuzione, del funzionamento e della manutenzione dei sistemi di AI.

La fusione del codice e dei dati

La tecnologia su cui si basa l'attuale AI non è come quella del software tradizionale. La tipica separazione tra codice e dati, fondamentale per valutare la sicurezza del software, è assente nei modelli di intelligenza artificiale e, al contrario, tutti i dati elaborati possono fungere da istruzioni, analogamente al "*code injection*" nella sicurezza del software.

Le centinaia di miliardi di parametri che controllano il comportamento dei modelli di intelligenza artificiale

11 Per mission system si intende uno strumento sviluppato per, o utilizzati in, programmi e progetti per un obiettivo ovvero strutture tecniche critiche specificamente sviluppate o significativamente modificate per sistemi specifici.

12 Guidelines for secure AI system development, https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf, NCSC, 2023

spesso sono derivate dai dati di addestramento, ma in una forma generalmente opaca all'analisi. Per esempio, la pratica che introduce questa separazione per realizzare l'allineamento degli LLM, si è dimostrata inadeguata in presenza di avversari. Attualmente, i sistemi di intelligenza artificiale possono essere controllati da input malevoli. Infatti, le barriere di sicurezza di un LLM possono essere "*jailbroken*" dopo solo 10 esempi di tuning.

Gli sviluppatori non hanno ancora adottato una modalità rigorosa per correggere queste vulnerabilità, e tanto meno per identificarle in maniera affidabile; quindi, è fondamentale misurare l'efficacia delle misure di sicurezza "*best-effort*" a livello di sistema e a livello operativo. La prassi dell'ingegneria dell'intelligenza artificiale offre stime di progettazione per sistemi e workflow per mitigare queste criticità. Questa attività è simile a quella utilizzata nell'ingegneria dei sistemi altamente affidabili, che sono inevitabilmente costruiti da componenti meno affidabili, ma occorre tenere presente che i modelli di ingegneria incentrati sull'intelligenza artificiale sono molto diversi dalle tradizionali metodologie di progettazione *fault-tolerant*. Infatti, gran parte della progettazione fault-tolerant si basa su ipotesi di indipendenza statistica rispetto ai guasti (p.e. transitori, intermittenti, permanenti) e, in genere, si impiega la ridondanza degli elementi del sistema, per diminuire le probabilità di guasto, e il controllo interno, per rilevare gli errori prima che diventino guasti, al fine di ridurre le conseguenze o il pericolo.

L'importanza dell'interazione uomo-sistema

Attualmente molti processi coinvolgono i sistemi basati sull'intelligenza artificiale esclusivamente per svolgere ruoli di supporto o di consulenza nei confronti degli operatori del processo stesso. Per esempio, da molto tempo i radiologi, i patologi, i gruppi antifrode e gli analisti delle immagini, fruiscono dell'ausilio dell'intelligenza artificiale. Poi, vi sono altri casi in cui i sistemi basati sull'intelligenza artificiale operano in maniera semi-autonoma (per esempio, nello screening delle candidature ad una posizione lavorativa). Questi modelli che comprendo l'interazione umana possono introdurre rischi specifici (per esempio, il sistema di screening potrebbe essere più autonomo nella realizzazione delle esclusioni e rimanere consultivo per le accettazioni e, di conseguenza, eliminare candidature con ottimi requisiti). In altre parole, esiste uno spettro di gradi di controllo condiviso e, pertanto, la natura di tale condivisione deve essere essa stessa un obiettivo del processo di valutazione del rischio. A riguardo, una soluzione potrebbe considerare il coinvolgimento degli esseri umani nel processo di valutazione delle proposte di esclusione e di accettazione, nonché nell'impiego di uno

schema di monitoraggio, per migliorare la responsabilità e fornire un riscontro al sistema e ai progettisti.

Un altro rischio connesso all'interazione uomo-sistema, legato a una debolezza umana piuttosto che a una debolezza del sistema, consiste nella tendenza ad antropomorfizzarne l'utilizzo sulla base dell'uso del linguaggio e della voce umana. Un noto esempio è il programma Eliza[13] scritto negli anni '60 al MIT da Joseph Weizenbaum. Eliza *"conversava"* con il suo utente umano attraverso la digitazione di un testo. Le 10 pagine di codice di Eliza facevano principalmente tre cose: rispondere in modo strutturato ad alcune parole "*trigger*", riflettere occasionalmente sugli input che gli venivano passati e invertire i pronomi. Sembrava, quindi, che Eliza capisse e, di conseguenza, le persone trascorrevano ore a conversare con lei nonostante il suo funzionamento fosse estremamente semplice. Esempi più recenti sono rappresentati da Siri, Alexa e Gemini, che, nonostante i nomi umani e le voci amichevoli, sono principalmente gateway di pattern-matching per la ricerca sul Web. Tuttavia, gli attribuiamo proprietà relative alla personalità e pronomi di genere per renderli più human friendly. Il punto è che gli esseri umani tendono a conferire ai testi significati e profondità, mentre i testi LLM sono una sequenza, statisticamente derivata, di previsioni delle parole successive.

Le superfici di attacco

Un'altra serie di minacce che mina lo sviluppo dei sistemi di AI sicuri e protetti è rappresentata del ricco e diversificato set di superfici di attacco associate agli attuali modelli di AI. L'esposizione di queste superfici di attacco è determinata, da un lato, dalle scelte ingegneristiche dell'AI e, dall'altra, dalla realizzazione di nuove interazioni Uomo-Macchina e, più in generale, dalla progettazione dei workflow operativi. In questo contesto, è opportuno definire l'ingegneria dell'AI non solo come l'ambito di progettazione, sviluppo, test e valutazione dei componenti di AI, ma anche dei sistemi che li contengono e dei workflow che incorporano le capacità dell'AI nelle principali attività.

A seconda del tipo di applicazione a cui sono adibiti i sistemi basati sull'intelligenza artificiale, e di come vengono progettati, le azioni avversarie possono manifestarsi sottoforma di input diretti da utenti malevoli, oppure sotto forma di *"training cases"* e *"retrieval augmentations"* (per esempio, sottoforma di file caricati, di siti web recuperati o di risposte da un plug-in o da uno strumento subordinato come la ricerca web). Possono

13 *J. Weizenbaum, ELIZA - a computer program for the study of natural language communication between man and machine, https://dl.acm.org/doi/10.1145/365153.365168, MIT, 1966*

essere forniti anche come parte di una query, sottoforma di dati non destinati ad essere interpretati come istruzioni (per esempio un documento fornito dall'utente per essere riassunto o elaborato). Nei fatti, queste superfici di attacco sono molto simili ad altre tipologie di minacce informatiche, solo che nel contesto dell'intelligenza artificiale è più difficile prevedere l'impatto di piccoli cambiamenti negli input nei confronti dei risultati, attraverso una qualsiasi di queste superfici di attacco. Infatti, c'è un'asimmetria informatica, adattata alle peculiarità dei modelli a rete neurale, in cui i difensori cercano una prevedibilità completa nell'intero dominio di input, mentre l'avversario necessita di prevedibilità solo per piccoli segmenti. Fortunatamente, la tecnica dell'adversarial ML consente di realizzare una specifica prevedibilità con più facilità e ciò conferisce un vantaggio per gli aggressori. Paradossalmente, questi attacchi sono ottenuti tramite l'utilizzo di altri modelli ML costruiti allo scopo di prevederli.

Inoltre, esistono diverse opportunità di attacco alla supply chain in grado di sfruttare la sensibilità dei risultati di training nelle scelte effettuate nella preferenza dei dati utilizzati per il processo di training. La robustezza di un modello, e delle relative misure di sicurezza, deve essere commisurata in relazione a ciascuno delle diverse tipologie di attacco. A riguardo, ognuno di questi tipi di attacco richiede nuovi metodi di analisi, di testing e una metrica specifica. È importante che sia raggiunto l'obiettivo di progettare schemi di valutazione ampiamente onnicomprensivi, anche in relazione allo stato dell'arte (in rapida evoluzione) e di ciò che è noto sui diversi metodi di attacco. È probabile che la completezza di questo traguardo rimanga inafferrabile, poiché continuano a svelarsi nuove vulnerabilità, debolezze e vettori di attacco.

La velocità dell'innovazione

Gli obiettivi di un progetto IT spesso si trovano in uno stato di rapida evoluzione, in parte guidata dai cambiamenti dell'ambiente operativo/strategico e dallo sviluppo di nuove tecnologie, tra cui gli algoritmi di intelligenza artificiale e le relative infrastrutture informatiche. Questa continua evoluzione crea ulteriori minacce sotto forma di pressione continua nell'effettuare l'aggiornamento degli algoritmi, delle infrastrutture informatiche, dell'immissione dei dati di training e degli altri elementi relativi alle potenzialità dell'intelligenza artificiale. Molto spesso, l'evoluzione accelerata anticipa le fasi di testing e valutazione, perchè gli stakeholder vengono coinvolti nella fase iniziale della sequenza temporale del processo (c.d. *"move to the left"*) e in modo continuato. Ciò consente, da un lato, di selezionare i progetti di sistema per migliorare la testabilità e, dall'altro, di configurare processi e strumenti in grado di produrre non solo modelli distribuibili, ma anche

dataset di prova destinati a supportare un processo di testing e valutazione continuo e affidabile. È risaputo che un coinvolgimento anticipato delle attività di T&E nel ciclo di vita del sistema comporta notevoli benefici e ricadute in termini di efficienza e sicurezza.

Il futuro dell'intelligenza artificiale

Se considerassimo l'elenco completo delle criticità e delle vulnerabilità, dal punto di vista della progettazione, dello sviluppo e del funzionamento dei sistemi basati sull'intelligenza artificiale, ricaveremmo un quadro sconfortante; in realtà, il vero problema è che le attuali misure di mitigazioni non sono sufficienti a contrastare le minacce. Al momento esistono poche misure di contenimento e, la maggior parte, sono concentrate alle fasi di progettazione e ad alcune scelte per contenere l'ambito operativo.

Tuttavia, è importante ricordare che l'intelligenza artificiale è in corso di evoluzione, infatti stanno emergendo molte proposte di modelli di *Hybrid AI*[14] per specifiche aree di applicazione. Queste idee creano nuove opportunità di sviluppo per l'intelligenza artificiale di base, perché gli consentiranno di offrire un'affidabilità intrinseca e verificabile rispetto a particolari categorie di rischi. Questa peculiarità è fondamentale perché, generalmente, non è possibile ottenere un'affidabilità intrinseca con l'I.A. basata unicamente sulle reti neurali.

Un altro importante punto di forza dell'intelligenza artificiale basata sulle reti neurali è rappresentato dall'eccezionale capacità euristica di questi modelli, anche se, come precedentemente detto, è difficile realizzare un test sicuro in modelli che sono, per loro natura, statistici e, quindi, sostanzialmente inesatti e generalmente poco trasparenti all'analisi. I sistemi basati sulla *symbolic artificial intelligence*[15], d'altro canto, offrono maggiore trasparenza, un ragionamento esplicito e ripetibile e, infine, la possibilità di manifestare competenza nel dominio di riferimento in modo verificabile, viceversa, sono generalmente deboli in termini di capacità euristica e, a volte, vengono percepiti come privi di flessibilità e scalabilità.

Molti gruppi di ricerca hanno riconosciuto questa complementarità e hanno efficacemente combinato tra

---

14 Hybrid artificial intelligence (intelligenza artificiale ibrida) può essere definita come l'arricchimento di modelli di intelligenza artificiale esistenti tramite conoscenze specialistiche ottenute per un apposito contesto.

15 Symbolic artificial intelligence (intelligenza artificiale simbolica) indica i metodi della ricerca sull'intelligenza artificiale che si basano su rappresentazioni di problemi "simbolic" di logica e ricerca. L'AI simbolica è stata il paradigma dominante della ricerca sull'AI dalla metà degli anni '50 fino alla fine degli anni '80.

loro i diversi approcci statistici derivanti da applicazioni euristiche avanzate. Alcuni esempi comprendono la combinazione del Machine Learning (ML) con la teoria dei giochi e l'ottimizzazione per supportare applicazioni che coinvolgono strategie multi-adversary, come il poker multigiocatore e le tattiche di antibracconaggio dei ranger.

Altri gruppi di ricerca hanno "*ibridato*" approcci statistici e simbolici per consentire lo sviluppo di sistemi capaci di pianificare e ragionare in maniera affidabile, sfruttando l'intelligenza artificiale come se fosse un oracolo euristico talvolta inaffidabile. Questi sistemi tendono a focalizzarsi su domini applicativi specifici, tra cui contesti in cui l'esperienza deve essere manifestata in modo affidabile. Questi sistemi simbolico-dominanti sono sostanzialmente diversi dall'utilizzo plug-in nei LLM. Normalmente, gli approcci ibridi all'I.A. sono utilizzati nei robot, nella comprensione del parlato e nel gioco. Per esempio, AlphaGo[16] utilizza un ML ibrido con strutture di ricerca.

I sistemi ibridi simbolici, in cui gli LLM sono subordinati, stanno iniziando ad apportare benefici in alcuni ambiti dello sviluppo software, tra cui la risoluzione degli errori e la verifica del software. È importante sottolineare che l'attuale I.A. simbolica ha infranto molte barriere di scalabilità che, dagli anni '90, sono state percepite come fondamentali. Ciò è riscontrabile in molteplici esempi, tra cui Google Knowledge Graph, che è euristicamente informato, ma verificabile dall'uomo; altro esempio è rappresentato dalla verifica delle proprietà di sicurezza su Amazon AWS, la quale utilizza tecniche di dimostrazione dei teoremi su larga scala. Questi esempi suggeriscono che altri approcci, simili ai precedenti, potrebbero fornire un livello di affidabilità in altri domini applicativi in cui questa caratteristica è rilevante. Una sfida importante consiste nel passare da questi esempi specifici a un'AI affidabile più generalista.

16 *AlphaGo è un software per il "gioco del go" sviluppato da Google DeepMind per studiare le reti neurali.*

Confidenzialità, Integrità e Governance

Lo sviluppo di sistemi di intelligenza artificiale per applicazioni critiche richiede necessariamente la conoscenza specifica delle debolezze e delle vulnerabilità dei modelli di intelligenza artificiale utilizzati. Questo aspetto è fondamentale per la progettazione, l'implementazione e la valutazione dei modelli di AI e dei sistemi basati sull'AI.

In questo paragrafo esaminiamo una serie di criticità associate ai moderni modelli di Artificial Intelligence (AI) basati su reti neurali. Questi modelli neurali includono il Machine Learning (ML) e l'AI generativa, con particolare riferimento i Large Language Models (LLM).

In particolare, ci concentriamo su tre aspetti:

- **i trigges:** i vettori di attacco delle azioni avverse (*exploiting vulnerabilities*) e le limitazioni intrinseche dovute alla natura statistica dei modelli (*manifestations from weaknesses*);
- **la natura delle conseguenze operative:** i potenziali tipi di guasti o gli errori operazionali;
- **le metodologie di mitigazione:** le attività progettuali e quelle operative.

Di seguito sono indicati alcuni esempi relativi a specifiche criticità, organizzati in base a tre tipologie di rischio: confidenzialità, integrità e governance (CIG)[17]. Questa analisi si ispira alle metodologie indicate dal NIST, tra cui l'AI RMF Framework[18], l'AI RMF Playbook[19] e l'AI RMF Generative Artificial Intelligence Profile[20].

Il NIST struttura le attività in quattro categorie:

1. *govern:* sostenere una cultura organizzativa basata sulla consapevolezza del rischio,

17 *CIG Framework: Confidentiality, Integrity and Governance.*
18 *AI RMF Framework, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf, NIST, 2023*
19 *AI RMF Playbook, https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook, NIST, 2023*
20 *AI RMF Generative Artificial Intelligence Profile, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf, NIST, 2024*

2. *map:* individuare il contesto di utilizzo,
3. *measure:* identificare, analizzare e valutare i rischi,
4. *manage:* dare le priorità e agire.

I rischi CIG si basano su queste pietre miliari del NIST e si concentrano sulle conseguenze sia degli attacchi (guidati dalle vulnerabilità), sia degli esiti avversi accidentali (guidati dalle debolezze), con l'intento di anticipare l'approccio *"Hybrid AI"* in grado di supportare le applicazioni critiche in modo sicuro e verificabile.

Confidenzialità

Solitamente i rischi di confidenzialità dei sistemi di intelligenza artificiale sono associati alla rivelazione involontaria dei dati di training o delle caratteristiche architetturali del modello neurale. Tra questi rientrano i cosiddetti attacchi denominati "*jailbreak*", una particolare tipologia di attacchi in grado di indurre i LLM a produrre risultati che superano i limiti stabiliti dai loro progettisti per prevenire determinati tipi di risposte pericolose e limitare la diffusione di contenuto malevolo. Questi attacchi compromettono anche l'integrità dei sistemi. Infatti, la derivazione statistica dei modelli di intelligenza artificiale non consente di delineare con precisione i confini delle categorie di rischio.

Il principale rischio connesso alla confidenzialità è rappresentato dalla violazione della privacy. L'opinione pubblica è convinta che i modelli siano stati addestrati su grandi insiemi di dati privati o sensibili, come le cartelle cliniche o le informazioni finanziarie, e che durante l'attività di riconoscimento o classificazione non sia stato possibile scoprire tali dati. Recenti studi hanno dimostrato l'infondatezza di tale assunzione, evidenziando che diversi tipi di attacchi alla privacy hanno comportato significative conseguenze nefaste per la sicurezza. Vediamo alcuni.

- **Gli attacchi di jailbreak e il trasferimento dei dati**. Esistono tecniche per sviluppare attacchi di injection o jailbreak al prompt in grado di eludere i sistemi di protezione, tipicamente integrati nei LLM, attraverso cicli di *fine-tuning*[21]. Esistono metodi per rendere le tecniche di jailbreak manuale più robuste, applicabili a modelli LLM API e LLM open source, e trasferibili sui modelli proprietari. Gli aggressori possono

21 *AA.VV., Universal and Transferable Adversarial Attacks on Aligned Language Models, https://arxiv.org/abs/2307.15043, 1Carnegie Mellon University, 2023*

ottimizzare un set di modelli open source per imitare i comportamenti dei modelli proprietari mirati e, successivamente, tentare un trasferimento black-box utilizzando i modelli testati. Continuano ad essere sviluppate nuove tecniche di jailbreak, spesso facilmente accessibili anche per chi ha risorse limitate, che appaiono al prompt sottoforma di testo in linguaggio naturale[22]. Alcune di queste tecniche includono l'assegnazione di ruoli, in cui ad un LLM viene chiesto di mettersi in un certo ruolo, per esempio come attore malevole, e in tale veste può rivelare informazioni protette[23].

- **L'inversione del modello e l'inferenza di appartenenza**. Un avversario con accesso limitato a un modello ML addestrato può estrarre dati di training tramite query. Sono stati identificati attacchi di inversione del modello[24] in grado di sfruttare le informazioni confidenziali generate dai modelli e consentire l'estrazione di informazioni sensibili, come i dati sanitari di un individuo, inseriti in un set di dati specifico per una determinata malattia oppure lo stile di una persona che ha partecipato ad un sondaggio.
- **Il problema della memorizzazione**. Il problema della memorizzazione dei dati di training, a differenza del problema di hallucination[25], si verifica quando gli utenti di un LLM si aspettano nuovi risultati sintetizzati, mentre ricevono una replica esatta dei dati di input. Questo fenomeno, noto come overfitting, può portare a violazioni della privacy, appropriazioni indebite della proprietà intellettuale e violazioni del copyright.
- **La ricerca black-box**. Se un modello proprietario espone un'API che fornisce probabilità per un insieme di output potenziali, una ricerca discreta ti tipo black-box può generare prompt avversari che superano le protezioni previste. Questa vulnerabilità è accessibile a un aggressore, anche senza particolari risorse GPU, che effettua chiamate ripetute all'API per identificare i prompt efficaci. Sono state documentate tec-

---

22 AA.VV., Weak-to-Strong Jailbreaking on Large Language Models, https://arxiv.org/abs/2401.17256, 1University of California, 2024

23 AA.VV., In ChatGPT We Trust? Measuring and Characterizing the Reliability of ChatGPT, https://arxiv.org/abs/2304.08979, CISPA, 2023

24 AA.VV., Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures, https://dl.acm.org/doi/pdf/10.1145/2810103.2813677, Carnegie Mellon University, 2015

25 L'hallucination è il fenomeno per cui un modello di machine learning o un'altra forma di AI genera risultati o output che non sono basati sui dati di addestramento o sulla realtà. In altre parole, l'allucinazione si verifica quando un modello di AI crea informazioni che non esistono. Questo fenomeno può verificarsi in diversi modelli di intelligenza artificiale, inclusi modelli di linguaggio come LLM. Ad esempio, un modello potrebbe generare frasi o risposte che sembrano plausibili ma che sono in realtà completamente inventate o basate su connessioni errate o casuali. L'allucinazione può compromettere l'affidabilità e la precisione delle risposte generate dai modelli.

niche chiamate *"leakage prompts"* in grado di ottenere punteggi di confidenza dai modelli i cui progettisti intendevano proteggere tali punteggi. Questi ultimi facilitando anche l'inversione del modello.

Potenziali mitigazioni

Proviamo a introdurre le azioni di mitigazione per ridurre il rischio per la confidenzialità:

- **La privacy differenziale**. Le soluzioni tecniche alla protezione della privacy, come la privacy differenziale, costringono gli ingegneri dell'intelligenza artificiale a valutare il compromesso tra sicurezza e accuratezza. Le tecniche di privacy differenziale fanno parte del set di tecniche basate sulla statistica chiamate "*privacy-preserving analytics*" (PPA), utilizzate per salvaguardare i dati privati supportando al contempo l'analisi. Le tecniche PPA includono anche le "*blind signatures*", il "*k-anonymity*" e l'apprendimento federato. Le tecniche PPA sono un sottoinsieme delle "*privacy-enhancing technologies*" (PET), che includono anche prove a *"zero-knowledge"* (ZK), la *"homomorphic encryption"* (HE) e il *"secure multiparty computation"* (MPC). Sono in corso esperimenti per integrarle nei modelli LLM per migliorare la privacy. Le tecniche di privacy differenziale implicano la perturbazione dei dati di training o degli output di un modello allo scopo di limitare la capacità degli utenti di poter trarre conclusioni su elementi particolari dei dati di training di un modello in base agli output osservati. Tuttavia, questo tipo di difesa ha un costo in termini di accuratezza dei risultati e mostra un pattern nella mitigazione del rischio ML, ovvero l'azione difensiva potrebbe interferire con l'accuratezza dei modelli addestrati.
- **Le tecniche di disapprendimento**. Sono state proposte diverse tecniche a supporto della rimozione dell'influenza di determinati esempi di training che potrebbero avere contenuti nocivi o che potrebbero compromettere la privacy tramite inferenza di appartenenza. Nel tentativo di accelerare questa attività sono state avviate attività di Machine Unlearning[26]. Tutti gli esperimenti sono arrivati alla conclusione che il disapprendimento automatico rimane un'incognita per l'uso pratico a causa del grado con cui i modelli si degradano analogamente agli effetti delle tecniche di privacy differenziale.

26 AA.VV., Google Machine Unlearning Challenge, https://research.google/blog/announcing-the-first-machine-unlearning-challenge/ - NeurIPS 2023 Machine Unlearning Challenge, https://unlearning-challenge.github.io/ - Who's Harry Potter? Approximate Unlearning in LLMs, https://arxiv.org/abs/2310.02238, Microsoft, 2023

Integrità

Nell'ambito dell'intelligenza artificiale basata sulle reti neurali, tra cui il ML e l'AI generativa, i rischi di integrità si riferiscono al potenziale di attacchi che potrebbero far sì che i sistemi producano risultati non previsti dai progettisti, dai programmatori e dai valutatori. Poiché le specifiche iniziali, oltre all'attenzione sui dati di training, sono difficili o impraticabili per molti modelli di reti neurali, il concetto di *risultati attesi* riveste solo un significato informale.

Di seguito sono indicati diversi tipi di attacco all'integrità contro le reti neurali, la natura dei punti deboli e delle vulnerabilità sfruttate, oltre ad alcune potenziali misure di mitigazione.

- **Il data poisoning** (Avvelenamento dei dati). Negli attacchi di data poisoning, un avversario interferisce con i dati su cui è addestrato l'algoritmo di ML, per esempio iniettando dati aggiuntivi durante il processo di training. L'avvelenamento può essere efficace anche nell'apprendimento supervisionato[27]. Questi attacchi consentono a un avversario di interferire con i comportamenti di *test-time* e *runtime* dell'algoritmo, sia degradando l'efficacia complessiva (c.d. accuratezza), sia inducendo l'algoritmo a produrre risultati errati in specifiche situazioni. La ricerca[28] ha dimostrato che una quantità sorprendentemente piccola di dati di addestramento manipolati può comportare grandi cambiamenti nel comportamento del modello. Gli attacchi di *data poisoning* sono particolarmente seri quando la qualità dei dati di addestramento non può essere accertata; questa difficoltà può essere amplificata dalla necessità di riaddestrare continuamente gli algoritmi con nuovi dati.

  Gli attacchi di poisoning possono verificarsi anche nell'apprendimento federato, per esempio nei domini relativi alla sicurezza nazionale o alla salute pubblica, in cui un insieme di organizzazioni addestrano congiuntamente un algoritmo senza condividere direttamente i dati posseduti da ciascuna organizzazione. Poiché i dati di training non vengono condivisi, può essere difficile, per qualsiasi parte, determinare la qualità complessiva dei dati. Esistono rischi simili con i dati pubblici, dove gli avversari possono facilmente distribuire input di addestramento nocivi. Gli attacchi correlati possono influenzare i metodi di tra-

27 *AA.VV., Indiscriminate Poisoning Attacks on Unsupervised Contrastive Learning, https://arxiv.org/abs/2202.11202, MIT, 2023*

28 *AA.VV., Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning, https://arxiv.org/abs/1712.05526, UC Berkeley, 2017*

sferimento del training, in cui un nuovo modello è derivato da un modello precedentemente addestrato. Potrebbe essere impossibile accertare quali fonti di dati siano state utilizzate per addestrare il modello di origine, il che nasconderebbe qualsiasi addestramento avversario che influenzi il modello derivato. Numerose ipotesi tentano di spiegare il sorprendente livello di trasferibilità tra modelli, tra cui, per modelli più grandi, la comunanza dei dati di addestramento e nella messa a punto per l'allineamento[29].

- **I misdirection and evasion attacks** (Attacchi di deviazione ed evasione). Gli attacchi di *evasion* sono caratterizzati dal tentativo di un avversario a indurre un modello a produrre output errati durante il funzionamento di un sistema. Esempi possono riguardare l'errata identificazione di un oggetto in un'immagine, la classificazione errata dei rischi nella consulenza agli addetti ai prestiti bancari e la valutazione errata della probabilità che un paziente possa trarre beneficio da un particolare trattamento sanitario. Questi attacchi vengono realizzati mediante la manipolazione, posta in essere dall'avversario, di un input o di una query fornita al modello. Gli attacchi di *evasion* sono spesso classificati come non mirati, quando l'obiettivo dell'avversario è ingannare il modello inducendolo a produrre una risposta errata, o mirati, nel caso in cui l'obiettivo dell'avversario è ingannare il modello inducendolo a produrre una specifica risposta errata. Un esempio di attacco consiste nel disorientare le reti neurali per il riconoscimento facciale posizionando punti colorati sulle montature degli occhiali[30]. In molti attacchi di *evasion*, è importante che l'input manipolato o fornito dall'aggressore sembri benigno, in modo tale che l'esame superficiale dell'input non riveli l'attacco. C'è anche il noto attacco degli adesivi su un segnale di stop[31]. È improbabile che questi adesivi siano notati dai conducenti umani, poiché molti segnali di stop hanno adesivi e altre alterazioni, ma gli adesivi posizionati con cura funzionano come patch in grado di disorientare in modo affidabile una rete di classificazione dei segnali affinché veda un segnale di limite di velocità. Questo tipo di spoofing richiede un impegno relativamente basso e, infatti, è stato oggetto di ricerca universitaria.

  È fondamentale definire quando l'output di un modello è corretto per valutare la suscettibilità dei modelli agli attacchi di *evasion*. Per molte applicazioni, la correttezza potrebbe essere definita quan-

29 AA.VV., Universal and Transferable Adversarial Attacks on Aligned Language Models, https://arxiv.org/abs/2307.15043, Carnegie Mellon University, 2023

30 AA.VV., A General Framework for Adversarial Examples with Objectives, https://arxiv.org/pdf/1801.00349, Carnegie Mellon University, 2019

31 AA.VV., Robust Physical-World Attacks on Deep Learning Visual Classification, https://arxiv.org/pdf/1707.08945, University of Michigan, 2018

do il sistema fornisce la risposta che darebbe un essere umano. Questo è difficile da testare con un certo grado di completezza. Inoltre, ci sono applicazioni in cui questo criterio non potrebbe essere sufficiente. Per esempio, potremmo voler vietare risultati accurati ma dannosi, come le istruzioni dettagliate su come realizzare un esplosivo o come commettere una frode con la carta di credito.

Una delle principali sfide nella valutazione è definire l'intento progettuale riguardo alla funzione del sistema e agli attributi di qualità, come avviene per una tradizionale specifica software. Poiché raramente è possibile fornire specifiche complete, le tre categorie CIG non sono delineate in modo netto e, in effetti, questo tipo di attacco pone rischi sia per l'integrità, che per la confidenzialità.

- **L'inesattezza**. La debolezza fondamentale condivisa da tutte le moderne tecnologie di AI deriva dalla natura statistica delle reti neurali e dal loro training: i risultati dei modelli basati su reti neurali sono previsioni statistiche. I risultati derivano da una distribuzione e gli errori dovuti all'effetto della memorizzazione o dell'allucinazione rientrano nei limiti di tale distribuzione. La ricerca sta portando a un rapido miglioramento: la progettazione dei modelli sta migliorando, i dataset di training stanno aumentando di dimensione e, infine, vengono applicate sempre più risorse computazionali ai processi di training. Tuttavia, è essenziale tenere presente che i modelli di reti neurale risultanti sono basati su dati stocastici e, pertanto, sono predittori inesatti.
- **Le Generative AI hallucinations** (Allucinazioni dell'intelligenza artificiale generativa). La caratteristica modellazione statistica delle architetture a rete neurale LLM può portare a contenuti generati in contrasto con i dati di training dati in input o che non sono coerenti con i fatti. In questi casi si parla di output "*hallucinated*". Le allucinazioni possono essere elementi rappresentativi generati all'interno di una categoria di risposte. Questo è il motivo per cui spesso si riscontra una vaga somiglianza con i fatti reali, definita incertezza aleatoria nel contesto delle tecniche di mitigazione della modellazione della quantificazione dell'incertezza (UQ).
- **Gli errori di ragionamento**. Il corollario dell'inesattezza statistica consiste nel fatto che i modelli a rete neurale non hanno capacità intrinseche per pianificare o ragionare. La comprensione del mondo da parte dei modelli è molto superficiale, in particolare se sono addestrati esclusivamente sul testo e, di conseguenza, gli LLM autoregressivi hanno limitate capacità di ragionamento e pianificazione. Per esempio, il funzionamento degli LLM è sostanzialmente un'iterazione nel prevedere la parola successiva di un testo o nel basarsi sul contesto di un prompt e sulla stringa di testo precedente che ha prodotto. Gli LLM

possono essere indotti a creare l'immagine di un ragionamento e, così facendo, forniscono previsioni migliori che potrebbe creare l'apparenza di ragionamento. Una delle tecniche di prompt per raggiungere questo obiettivo è chiamata *"chain-of-thought"* (CoT)[32]. Questo crea una sorta di *"fast-thinking"*[33] tra pianificazione e ragionamento, ma genera risultati inevitabilmente inesatti, che diventano più evidenti una volta che le catene di ragionamento aumentano anche solo di poco. Uno studio recente[34] ha dimostrato che, generalmente, le catene più lunghe, anche di una dozzina di passaggi, non sono fedeli al ragionamento svolto senza CoT. Tra i numerosi parametri funzionali ai sistemi di ragionamento automatico e sul calcolo si evidenziano: la capacità di effettuare controlli esterni per la solidità delle strutture di ragionamento prodotte da un LLM e il numero di passaggi di ragionamento e/o calcolo intrapresi.

Potenziali mitigazioni

Oltre alle attività riparatorie menzionate nei paragrafi precedenti, sono allo studio diverse potenzialmente alternative capaci di mitigare un'ampia gamma di vulnerabilità. Vediamole:

- **Uncertainty quantification (UQ) (**la quantificazione dell'incertezza). La quantificazione dell'incertezza, nell'ambito del ML, si concentra sull'identificazione dei tipi di incertezze statistiche predittive che si presentano nei modelli ML, con l'obiettivo di modellare e misurare tali incertezze. Nel contesto del ML, si distingue tra le incertezze relative a effetti statistici intrinsecamente casuali (c.d. aleatorie) e le incertezze relative a insufficienze nella rappresentazione della conoscenza in un modello (c.d. epistemiche)[35]. L'incertezza epistemica può essere ridotta tramite un training aggiuntivo e il miglioramento dell'architettura di rete, mentre l'incertezza aleatoria è correlata all'associazione statistica di input e output e non può essere ridotta. I metodi UQ dipendono da precise specifiche delle caratteristiche statistiche del problema. Inoltre, sono poco utili nelle applicazioni di ML in cui gli avversari hanno avuto accesso alle superfici di

32 *AA.VV., Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, https://arxiv.org/abs/2201.11903, Google, 2023*

33 *D. Kahneman, Of 2 Minds: How Fast and Slow Thinking Shape Perception and Choice, https://www.scientificamerican.com/article/kahneman-excerpt-thinking-fast-and-slow/, ScientificAmerican, 2012*

34 *AA.VV., Measuring Faithfulness in Chain-of-Thought Reasoning, https://arxiv.org/abs/2307.13702, Anthropic, 2023*

35 *AA.VV., Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods, https://link.springer.com/article/10.1007/s10994-021-05946-3, Springer, 2021*

attacco. Esistono metodi UQ che tentano di rilevare campioni che non si trovano nella parte centrale di una distribuzione di probabilità degli input attesi. Anche questi ultimi sono suscettibili di attacchi.

Molti modelli di ML possono essere dotati della capacità di esprimere fiducia oppure la probabilità di errore. Ciò consente di modellare gli effetti degli errori a livello di sistema in modo che i loro effetti possano essere mitigati durante l'implementazione. Questo avviene attraverso una combinazione di metodi per la quantificazione dell'incertezza nei modelli di ML, la creazione di un framework software per ragionare con incertezza e la gestione sicura dei casi in cui i modelli di ML siano incerti.

- **Retrieval augmented generation (RAG) (**Generazione aumentata di recupero). Alcuni studi suggeriscono di sviluppare nei LLM la capacità di controllare la coerenza degli output rispetto alle fonti che si prevede rappresentino le verità fondamentali, come le basi di conoscenza o determinati siti Web come Wikipedia. La RAG si riferisce a questa idea di utilizzare database esterni per verificare e correggere gli output dei LLM. Questo rappresenta una potenziale misura di mitigazione sia per gli attacchi di evasione che per le allucinazioni dell'AI generativa, ma è imperfetta perché i risultati del recupero vengono elaborati dalla stessa rete neurale.
- **L'ingegneria della rappresentazione**. L'aumento del grado di astrazione in un'analisi white-box può potenzialmente migliorare la comprensione di una serie di comportamenti indesiderati nei modelli, tra cui le allucinazioni, i pregiudizi e la generazione di risposte nocive[36]. Esistono diversi metodi che tentano di estrarre la funzionalità. Questo tipo di test richiede un accesso al modello del tipo white-box, ma ci sono risultati preliminari che restituiscono effetti simili in scenari di test del tipo black-box, ottimizzando i prompt che hanno come target le stesse rappresentazioni interne. Si tratta di un elemento per ridurre l'opacità[37] caratteristica dei modelli di reti neurali più grandi. Uno studio recente, nell'ambito dell'interpretabilità automatizzata, ha dimostrato che sia possibile automatizzare un processo iterativo di sperimentazione per identificare i concetti latenti nelle reti neurali e quindi dare loro nomi[38].

36 AA.VV., A top-down approach to AI transparency, https://arxiv.org/pdf/2310.01405, CMU, 2023

37 AA.VV., Mapping the Mind of a Large Language Model, https://www.anthropic.com/news/mapping-mind-language-model, Anthropic, 2024

38 AA.VV., A Multimodal Automated Interpretability Agent, https://arxiv.org/pdf/2404.14394, MIT, 2024

Governance e Responsabilità

Gli incidenti che coinvolgono l'intelligenza artificiale sono ampiamente documentati attraverso vari repository[39]. Per mitigare un pò i rischi è necessario essere consapevoli non solo delle debolezze e delle vulnerabilità, ma anche dei principi di governance dell'AI, ovvero del modo con cui le organizzazioni sviluppano, regolamentano e gestiscono la responsabilità dei workflow supportati dall'AI.

- **Stakeholders e accountability**. La governance può coinvolgere un ecosistema costituito da elementi e sistemi di AI e dagli stakeholder umani e organizzativi. Questi stakeholder possono includere progettisti, sviluppatori di sistemi, team di distribuzione, leadership istituzionale, utenti finali e decisori, fornitori di dati, operatori, consulenti legali, valutatori e revisori. Essi sono complessivamente responsabili delle decisioni relative alla scelta di assegnare determinate capacità a determinate tecnologie di AI in un determinato contesto applicativo, nonché delle scelte relative al modo in cui il sistema basato sull'AI è integrato nei flussi operativi e nei processi decisionali. Sono, inoltre, responsabili dell'architettura dei modelli e della selezione dei dati di training, compreso l'allineamento dei dati al contesto operativo. Sono responsabili dei parametri, dei tassi di rischio e di responsabilità. L'assegnazione della responsabilità tra coloro che sono coinvolti nella progettazione, nello sviluppo e nell'uso dei sistemi di intelligenza artificiale non è un compito semplice. Questa circostanza è nota come il *"problem of many hands"*[40]. Questo problema è amplificato dalla assenza di trasparenza e dall'incomprensibilità dei modelli di intelligenza artificiale, spesso persino ai loro creatori[41] [42]. Nel contesto della data science, è fondamentale sviluppare strutture di governance efficaci che siano consapevoli delle caratteristiche dell'AI moderna.

---

39 AI Incident Database from the Responsible AI Collaborative: https://incidentdatabase.ai/ - AI Incident Database from the Partnership on AI: https://partnershiponai.org/workstream/ai-incidents-database/ - the Organisation for Economic Co-operation and Development (OECD) AI Incidents Monitor: https://oecd.ai/en/incidents - AI, Algorithmic, and Automation Incidents and Controversies (AIAAIC) Repository of incidents and controversies: https://www.aiaaic.org/aiaaic-repository

40 D. Thompson, Designing Responsibility: The Problem of Many Hands in Complex Organizations: https://dash.harvard.edu/bitstream/handle/1/37092148/Dennis%20Thompson%20chapter.pdf, Cambridge, 2014

41 M. Sullivan, The frightening truth about AI chatbots: Nobody knows exactly how they work, https://www.fastcompany.com/90896928/the-frightening-truth-about-ai-chatbots-nobody-knows-exactly-how-they-work, FastCompany, 2023

42 R. Curry, OpenAI Doesn't Fully Understand How GPT Works Despite Rapid Progress: https://observer.com/2024/05/sam-altman-openai-gpt-ai-for-good-conference/, Observer, 2024

- **Pacing** (ritmo). Le criticità in materia di governance derivano anche dalla velocità dello sviluppo tecnologico. Questo include non solo le principali tecnologie dell'AI, ma anche i continui progressi nel campo dell'identificazione e della comprensione di vulnerabilità. Infatti, questa velocità sta portando a una continua escalation delle aspettative sulle capacità operative.
- **Business**. Un ulteriore insieme di difficoltà per la governance deriva dagli aspetti legati al business, tra cui il segreto commerciale e la protezione della proprietà intellettuale, così come le scelte relative al modello architetturale e ai dati di training. In molti casi, le informazioni sui modelli, nell'ambito di una supply chain, possono essere deliberatamente limitate. Tuttavia, è importante notare che, quando le superfici di attacco sono sufficientemente esposte, molti degli attacchi sopra menzionati possono avere successo nonostante vi siano restrizioni black-box. Infatti, uno dei paradossi del rischio informatico è che, a causa del segreto commerciale, gli avversari possano conoscere l'ingegneria dei sistemi meglio rispetto alle organizzazioni che valutano e gestiscono questi sistemi. Questo è uno dei motivi per cui l'AI open source è ampiamente attenzionata anche da parte degli sviluppatori proprietari[43].
- **Responsabile AI**. Sono state pubblicate diverse linee guida che trattano la *Responsabile AI* (RAI) e in molti casi convergono sugli stessi principi: la correttezza, la responsabilità, la trasparenza, la sicurezza, la validità, l'affidabilità, la protezione e la privacy. Il Dipartimento della Difesa Americano ha pubblicato una strategia RAI insieme a un toolkit associato[44].

Esistono diverse minacce legate alla governance:

- **Deepfake**. Gli strumenti di AI generativa possono operare in più modalità e produrre materiale deepfake multimodale, per esempio audio e video, che possono apparire verosimilmente come se fossero originali. Sono state pubblicate numerose ricerche sul rilevamento dei deepfake[45] e sulla generazione aumentata tramite filigrane e altri tipi di firme[46]. Il ML e la GenAI possono essere utilizzati sia per generare deepfake, che per analizzare firme deepfake. Ciò significa che la tecnologia di intelligenza artificiale è in

43 E. Gen, The tech industry can't agree on what open-source AI means. That's a problem, https://www.technologyreview.com/2024/03/25/1090111/tech-industry-open-source-ai-definition-problem/, 2024

44 AA.VV., Responsible Artificial Intelligence Strategy and Implementation Pathway: https://www.ai.mil/docs/RAI_Strategy_and_Implementation_Pathway_6-21-22.pdf, DoD, 2022

45 Semantic Forensics: https://www.darpa.mil/program/semantic-forensics

46 AA.VV., Provable Robust Watermarking for AI-Generated Text, https://arxiv.org/pdf/2306.17439, UC Santa Barbara, 2023

crescita su entrambi i fronti: la creazione e il rilevamento della disinformazione[47].

- **Overfitting** (sovradimensionamento). È possibile addestrare il modello di ML in modo tale da portare a un overfitting. Questo succede quando il miglioramento del tasso di successo nel dataset di training portano a un degrado della qualità dei risultati nel dataset di test. Il termine overfitting deriva dal contesto della modellazione matematica quando e si usa per descrivere i casi in cui i modelli non riescono a catturare in maniera affidabile le caratteristiche salienti dei dati, per esempio compensando eccessivamente gli errori di campionamento. Il problema della memorizzazione è una forma di overfitting[48]. L'overfitting è trattato come un rischio di governance perché implica scelte fatte nella progettazione e nell'addestramento dei modelli.
- **Undertfitting** (sottodimensionamento). L'underfitting è un altro tipo di errore che si verifica quando il modello non è in grado di determinare una relazione significativa tra i dati di input e di output. L'underfitting si verifica se i modelli non sono stati addestrati per il periodo di tempo appropriato su un numero elevato di punti di dati. I modelli con underfitting presentano un bias elevato: forniscono risultati imprecisi sia per i dati di addestramento sia per il set di test.
- **Bias** (pregiudizio/distorsione). Spesso si ritiene che il bias derivi dalla mancata corrispondenza dei dati di training con i dati di input, ovvero che i dati di addestramento non sono allineati con i contesti applicativi. Inoltre, nel caso in cui non vi è la disponibilità di dati idonei, è possibile incorporare il bias nei dati di addestramento anche quando il processo di campionamento degli input è destinato ad essere allineato con i casi d'uso. Per tanto, a causa della mancanza di disponibilità di dataset di training imparziale, è difficile correggere il bias. Per esempio, è stato osservato un bias di genere nei vettori di parole degli LLM: la distanza vettoriale della parola *female* è più vicina a *nurse,* mentre *male* è più vicina a *engineer*[49]. Il problema della parzialità nelle decisioni dell'AI è correlato alle conversazioni attive nell'ambito della corretta classificazione dei risultati nei sistemi di ricerca e di raccomandazione[50].

---

47 AA.VV., Fighting Fire with Fire: The Dual Role of LLMs in Crafting and Detecting Elusive Disinformation, https://aclanthology.org/2023.emnlp-main.883.pdf, Pennsylvania State University, 2023

48 AA.VV., A Careful Examination of Large Language Model Performance on Grade School Arithmetic, https://arxiv.org/pdf/2405.00332v1, 2024

49 AA.VV., Gender bias and stereotypes in Large Language Models, https://dl.acm.org/doi/fullHtml/10.1145/3582269.3615599, Apple, 2023

50 AA.VV., Evaluating Stochastic Rankings with Expected Exposure, https://dl.acm.org/doi/10.1145/3340531.3411962,

- **Tossic text**. I modelli di AI generativa possono essere addestrati sia sui contenuti migliori, che su quelli peggiori di Internet. I modelli di GenAI possono utilizzare strumenti per filtrare i dati di training, ma il risultato potrebbe risultare imperfetto. Anche quando i dati di training non sono esplicitamente tossici, la messa a punto può consentire la generazione di materiale negativo. È importante riconoscere che non esistono definizioni universali e che la caratteristica di tossicità spesso dipende in larga misura dal pubblico e dal contesto: esistono diversi tipi di contesti che influenzano le decisioni in merito all'appropriatezza del linguaggio. La maggior parte dei rimedi prevede l'uso di filtri sui dati di training e un'ottimizzazione dell'input, del prompt e dell'output. I filtri spesso includono l'apprendimento con feedback umano "*reinforcement learnig with human feedback (RLHF)*". Finora, nessuna di queste soluzioni è riuscita a eliminare i danni da tossicità, soprattutto laddove i segnali nocivi sono nascosti.
- **Rischi informatici**. È importante ricordare che anche gli attacchi informatici che coinvolgono la supply chain rappresentano un rischio significativo per i modelli di ML. Questo riguarda sia i modelli black-box, che quelli open source, perché entrambi possono includere payload indesiderati. Lo stesso si può affermare nei modelli basati sul cloud in cui si può accedere attraverso API non sicure. Questi sono i classici rischi della supply chain del software, ma la complessità e l'opacità dei modelli di AI possono creare ulteriori opportunità per gli aggressori.

2020

Valutazione dei rischi nella Generative AI

I sistemi di Generative Artifical Intelligence (GenIA) pongono nuovi tipi di rischi, diversi dai classici rischi informatici, molti dei quali sono consequenziali e poco noti. Nonostante ciò, stiamo assistendo a una forte crescita dell'implementazione e diffusione di nuovi sistemi basati sulla GenAI in qualsiasi ambito sociale e lavorativo. Questa criticità ha, di fatto, accelerato la ricerca e lo sviluppo di modelli efficaci per realizzare il test e la valutazione dei sistemi di AI.

In questo paragrafo proviamo a descrivere un framework per la gestione del rischio dell'AI seguendo il modello del rischio informatico. Nello specifico, sono illustrate alcune potenziali strategie per inquadrare le attività di T&E sulla base di un approccio olistico al rischio dell'AI. È opportuno basare lo sviluppo di questo framework sulle lezioni apprese nei decenni di ricerca per individuare soluzioni analoghe già implementate per la modellazione e la valutazione del rischio informatico. Le valutazioni del rischio informatico sono imperfette e continuano a evolversi, ma, comunque, forniscono vantaggi significativi, tant'è che sono divenute un obbligo normativo nei contesti delle infrastrutture critiche, nel settore finanziario, nell'ambito dei servizi pubblici essenziali, ecc.

La modellazione e la valutazione del rischio per l'AI sono poco comprese sia dal punto di vista tecnico, che legale; esiste, comunque, una domanda urgente sia da parte degli utilizzatori, che dei fornitori[51]. A riguardo, nel luglio del 2024 la Coalition for Secure AI[52] ha fornito un importante contributo a far avanzare le norme del settore relative al miglioramento della sicurezza delle moderne implementazioni dell'AI. Il NIST AI Risk Management Framework (RMF) è un primo esempio di questo apporto. Ad oggi, le metodologie proposte sono ancora in fase di sviluppo, con costi e benefici incerti; pertanto, le valutazioni del rischio dell'AI risultano meno applicate rispetto alle valutazioni del rischio informatico.

La modellazione e la valutazione del rischio sono importanti non solo per effettuare il T&E, ma anche per

51 *AI in 2024: McKinsey Report Reveals Value Generation & AI Adoption Spike: https://www.switchsoftware.io/post/ai-in-2024-gen-ai-rise-and-business-impact, 20204*

52 *Making AI Systems Secure for All: https://www.coalitionforsecureai.org/*

informare i processi di progettazione, come sta avvenendo nell'ingegneria della sicurezza informatica e nell'emergente ingegneria dell'AI. È importante ricordare che l'ingegneria dell'AI non comprende solo i singoli elementi dell'AI incorporati nei sistemi, ma anche la progettazione complessiva di sistemi resilienti basati sull'AI, insieme ai workflow e alle interazioni umane che consentono le attività operative.

La modellazione del rischio dell'AI può avere un'influenza positiva non solo nella fase di T&E, ma durante l'intero ciclo di vita dell'AI, che va dalle scelte di progettazione alle specifiche fasi di mitigazione del rischio. I punti deboli e le vulnerabilità correlate all'AI hanno caratteristiche uniche (vd. gli esempi nel paragrafo precedente), ma si sovrappongono anche ai rischi informatici. Dopotutto, gli elementi di sistema dell'AI sono componenti software, quindi presentano vulnerabilità non correlate alla loro funzionalità di AI. Tuttavia, le loro caratteristiche uniche e spesso non note, sia all'interno dei modelli, che nelle strutture software che le ospitano, possono renderli particolarmente attraenti per i cybercriminali.

Attributi funzionali e qualitativi

Le valutazioni funzionali e qualitative contribuiscono a garantire che i sistemi svolgano le attività in maniera corretta e affidabile. Tuttavia, correttezza e affidabilità non sono concetti assoluti, ma devono essere inquadrati nel contesto degli obiettivi specifici di un componente o di un sistema, compresi i limiti operativi che devono essere rispettati. Le specifiche comprendono necessariamente sia la funzionalità, ossia ciò che il sistema è destinato a realizzare, sia le qualità del sistema, ovvero il modo in cui il sistema intende funzionare. inclusi gli attributi relativi alla sicurezza e all'affidabilità. Queste peculiarità, o specifiche di sistema, possono riguardare sia il sistema che il suo ruolo nell'operatività, comprese le aspettative relative a fattori di stress da minacce avverse.

I sistemi basati sull'intelligenza artificiale presentano rilevanti sfide tecniche in ogniuno dei seguenti aspetti: dalla formulazione delle specifiche alla valutazione dell'accettazione fino a comprendere il monitoraggio operativo. Per esempio, cosa occorre specificare, oltre a inventariare i dati di training e test, per implementare una rete neurale di ML?

In altre parole, è necessario considerare il comportamento di un sistema, o di un workflow associato, in base agli input previsti e imprevisti, dove tali input potrebbero risultare particolarmente critici per il sistema. Tuttavia, è difficile definire come occorre specificare i comportamenti per gli input previsti che non corrispondono esattamente al set di training. Un operatore umano potrebbe notare una somiglianza tra i nuovi input e quelli

di training, ma non vi sarebbe alcuna garanzia che ciò corrisponda alle caratteristiche effettive, ovvero ai valori dei parametri salienti, all'interno di una rete neurale addestrata.

Inoltre, dobbiamo esaminare le valutazioni anche dal punto di vista della sicurezza informatica. Un utente malintenzionato informato e motivato potrebbe manipolare deliberatamente gli input operativi, i dati di training e gli altri aspetti del processo di sviluppo del sistema, per creare condizioni che compromettano il corretto funzionamento di un sistema o il suo utilizzo all'interno di un workflow. In entrambi i casi, l'assenza di specifiche modifica la nozione di comportamento "*corretto*", complicando ulteriormente lo sviluppo di processi efficaci ed economicamente convenienti per il T&E. Questa criticità suggerisce un'altra comunanza con il rischio informatico: i movimenti laterali, che sono potenziali superfici di attacco accidentali all'implementazione, e potrebbero non rientrare in una specifica di sistema.

Tre dimensioni del rischio informatico

L'allineamento tra i requisiti emergenti per un T&E incentrato sull'AI e le tecniche per la valutazione della sicurezza informatica è evidente quando si confronta il NIST's AI Risk Management Playbook[53] con il NIST Cybersecurity Framework[54], comprendente un'enorme varietà di metodi. Per semplificare, possiamo inquadrare questi metodi nel contesto delle tre dimensioni del rischio informatico.

- **Threat**: la minaccia consiste nel potenziale accesso e nelle attività attuate dagli avversari contro il sistema e il suo più ampio ecosistema operativo.
- **Consequence**: la conseguenza riguarda l'entità dell'impatto su un'organizzazione o su una mission specifica qualora un attacco a un sistema avesse successo.
- **Vulnerability**: la vulnerabilità si riferisce alle debolezze intrinseche nella progettazione e ai difetti nell'implementazione di un sistema.

Sia la minaccia che le conseguenze sono strettamente dipendenti dal contesto operativo di un sistema, sebbene possano essere in gran parte esterne al sistema stesso. La vulnerabilità è una caratteristica del sistema e comprende l'architettura e l'implementazione. La modellazione della superficie di attacco, ovvero le debo-

53 *NIST's AI risk management playbook: https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook*
54 *NIST Cybersecurity Framework: https://www.nist.gov/cyberframework*

lezze di un sistema esposte alle azioni avverse, comprende le minacce e le vulnerabilità, perché l'accesso alle vulnerabilità è una conseguenza dell'ambiente operativo. Questo rappresenta un elemento particolarmente utile nell'analisi del rischio informatico.

La modellazione del rischio informatico è diversa dalla tradizionale modellazione probabilistica del rischio attuariale[55]. Ciò è dovuto principalmente alla natura generalmente non stocastica di ciascuna delle tre dimensioni, soprattutto quando le minacce e le azioni sono consequenziali. Per esempio, la minaccia può essere determinata dall'importanza operativa del sistema e del suo workflow, nonché dalle potenziali intenzioni degli avversari e dallo stato delle loro conoscenze. Allo stesso modo, la conseguenza può essere determinata dalle scelte relative al posizionamento di un sistema nei workflow operativi. Gli adeguamenti ai flussi di lavoro e ai ruoli umani è una strategia di mitigazione del rischio nella dimensione delle conseguenze. I rischi possono aumentare quando esistono correlazioni nascoste. Per quanto riguarda il rischio informatico, questi potrebbero includere elementi comuni con vulnerabilità comuni nascoste nelle catene di approvvigionamento. In merito al rischio legato all'intelligenza artificiale, questi potrebbero includere fonti comuni all'interno di grandi quantità di dati del training. Queste correlazioni rappresentano uno dei motivi per cui alcuni attacchi agli LLM sono trasferibili tra modelli e provider diversi.

I framework elaborati dal CISA[56], dal MITRE[57], dall'OWASP[58] offrono utili database di debolezze e vulnerabilità informatiche e, alcuni di questi, forniscono anche la soluzione di sicurezza. La maggior parte dei criteri di valutazione generalmente utilizzati derivano da questi framework con un approccio bottom-up. Per quanto riguarda i punti deboli e le vulnerabilità a livello di codifica, spesso gli ambienti di sviluppo software, gli strumenti automatizzati e i flussi di lavoro di integrazione/distribuzione continua includono la capacità di analisi, ovvero la possibilità di rilevare la codifica non sicura nel momento stesso in cui gli sviluppatori digitano o compilano i componenti eseguibili.

È importante sottolineare che il rischio informatico rappresenta solo un elemento nella valutazione complessiva dell'idoneità all'uso di un sistema, indipendentemente dal fatto che sia basato o meno sull'intelligenza artificiale. La valutazione dell'accettazione del rischio nei sistemi hardware-software integrati include anche

55 *The Roles of the Actuary in the Selection & Application of Actuarial Models: https://www.actuary.org/sites/default/files/2023-02/01_Models2.9.23.pdf*

56 *Cybersecurity & Infrastructure Security Agency: https://www.cisa.gov/known-exploited-vulnerabilities-catalog*

57 *MITRE Corporation: https://cwe.mitre.org/data/index.html*

58 *Open Web Application Security Project: https://owasp.org/www-community/attacks/*

le tradizionali analisi di affidabilità probabilistica che modellano:

1. i tipi di guasti fisici (intermittenti, transitori, permanenti),
2. come tali guasti possono innescare errori interni a un sistema,
3. come gli errori possono propagarsi in vari tipi di guasti a livello di sistema
4. quali tipi di pericoli o danni (alla sicurezza, alla protezione, al funzionamento efficace) potrebbero causare flussi di lavoro operativi.

Questo approccio all'affidabilità dei sistemi risale al lavoro di John von Neumann degli anni '50 sulla sintesi di meccanismi affidabili da componenti inaffidabili[59]. È interessante notare che von Neumann cita la ricerca sulla logica probabilistica che deriva da modelli sviluppati da McCulloch e Pitts, i cui modelli di rete neurale degli anni '40 sono precursori dei progetti di rete neurale centrali per l'intelligenza artificiale moderna.

Determinare il rischio dell'AI

L'inquadramento del rischio legato all'AI può essere considerato analogo al rischio informatico, nonostante le differenze tecniche nei tre aspetti: minaccia, conseguenza e vulnerabilità. Nei contesti in cui sono presenti utenti malevoli, le minacce all'AI possono includere la manomissione dei dati di training, gli attacchi di patch sugli input, gli attacchi del prompt e di messa a punto, etc.; per quanto riguarda le conseguenze dell'AI possiamo considerare i depistaggi, le ingiustizie e i pregiudizi, gli errori di ragionamento, etc. e, infine, in merito alle vulnerabilità e alle debolezze, come quelle elencate nelle categorie CIG, derivano generalmente dai limiti intrinseci dell'architettura e dell'addestramento delle reti neurali come i modelli derivati statisticamente. Si noti che possono verificarsi diverse conseguenze negative, anche in assenza di avversari, dovute alle particolari debolezze intrinseche dei modelli di reti neurali.

Dal punto di vista della modellazione del rischio tradizionale, vi è anche la difficoltà di scoprire correlazioni inattese tra modelli e piattaforme. Per esempio, possono esserci conseguenze simili dovute a LLM di provenienza diversa che condividono modelli o semplicemente hanno una sovrapposizione sostanziale nei dati di

59 *J. von Newmann, Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components: https://web.mit.edu/6.454/www/papers/pierce_1952.pdf, California Institute of Technology, 1952*

training. Queste correlazioni inaspettate possono ostacolare i tentativi di applicare tecniche come la "*diversity by design*"[60] come mezzo per migliorare l'affidabilità complessiva del sistema.

Dobbiamo considerare anche l'attributo specifico della resilienza. La resilienza è la capacità di un sistema che ha subito un attacco o un guasto a continuare comunque a funzionare in modo sicuro, anche se eventualmente in modo degradato. Questa caratteristica è talvolta chiamata "*graceful degradation*" o capacità di *"operate through"* (operare nonostante) gli attacchi e i guasti. In generale, è estremamente difficile, e spesso irrealizzabile, aggiungere resilienza a un sistema esistente. Questo perché la resilienza è una proprietà emergente conseguente alle decisioni architetturali a livello di sistema. L'obiettivo dell'architettura è ridurre la possibilità che gli errori interni, innescati da difetti interni, le compromissioni o le debolezze intrinseche del machine learning, causino guasti al sistema con conseguenze onerose. La tradizionale ingegneria *"fault-tolerant"* è un esempio di progettazione per la resilienza. La resilienza è una proprietà sia per il rischio informatico che per il rischio dell'AI. Per esempio, nell'ambito dell'AI, la resilienza può essere migliorata attraverso scelte progettuali, a livello di sistema e di workflow, in grado di limitare l'esposizione delle superfici di attacco interne agli agenti esterni, come gli input al ML, potenzialmente più vulnerabili. Tali scelte possono includere l'imposizione di un controllo attivo, sia sull'input che sull'output, ai modelli di reti neurali costituenti un sistema.

Un'ulteriore minaccia alla resilienza dell'AI è rappresentata dalla difficoltà, o forse dall'incapacità, di dimenticare i dati di training[61]. Ipotizziamo che si scoprisse che un sottoinsieme di dati di training sia stato utilizzato per inserire una vulnerabilità o una backdoor nel sistema di AI, la rimozione di quel determinato comportamento conosciuto dal sistema di AI rappresenterebbe una grossa criticità perché la soluzione percorribile per eliminarla richiederebbe un nuovo training privo dei dati nocivi. Un problema correlato alla difficolta di dimenticare i dati indesiderati (c.d *unwanted unlearning*) è rappresentato dal fenomeno opposto denominato oblio catastrofico (c.d. *catastrophic forgetting)* che si verifica quando dei nuovi dati di training riescono a compromettere involontariamente la qualità delle previsioni basate sui dati di addestramento precedenti[62].

60 *AA.VV., Systems of Systems Engineering: Basic Concepts, Model-Based Techniques, and Research Directions: https://dl.acm.org/doi/abs/10.1145/2794381, 2015*

61 *A. Snyder, Machine forgetting: How difficult it is to get AI to forget: https://www.axios.com/2024/01/12/ai-forget-unlearn-data-privacy, 2024*

62 *AA.VV., An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-tuning: https://arxiv.org/abs/2308.08747, School of Engineering, Westlake University, 2024*

Nonostante si stia assistendo ad una rapida crescita economica dei settori correlati all'AI, tutti gli stakeholder concordano sul fatto che la misurazione e la valutazione del rischio dell'AI è un esercizio complesso e difficile. A riguardo sono stati pubblicati diversi studi di centri di ricerca finalizzati alla catalogazione dei rischi associati al ML e all'AI generativa [63] [64] [65].

Migliorare la gestione del rischio dell'AI

Lo sviluppo di best-practice per un sistema di gestione del rischio per l'AI deve necessariamente tenere conto dei diversi aspetti del rischio e della fattibilità dei vari approcci alla mitigazione. Le valutazioni devono essere eseguite a più livelli di astrazione e di struttura, nonché in più fasi all'interno dei vari cicli di vita della pianificazione, della progettazione architetturale, dello sviluppo dei sistemi, della distribuzione e dell'evoluzione. Questa complessità può rendere difficile il processo di gestione. Nel livello più alto identifichiamo i workflow, la progettazione dell'interazione umana e i progetti architetturale del sistema. Le scelte effettuate riguardo ciascuno di questi aspetti influenzano direttamente i fattori di rischio: l'attrattività per gli attori malevoli, la natura e l'entità delle conseguenze dei potenziali danneggiamenti e il potenziale di vulnerabilità dovuto alle decisioni in fase di progettazione. Occorre poi considerare l'architettura e la formazione dei singoli modelli di rete neurale, la messa a punto e il prompt dei modelli generativi e la potenziale esposizione delle superfici di attacco di questi modelli. Nello strato sottostante ci sono, per esempio, gli algoritmi matematici e le singole linee di codice. Infine, quando le superfici di attacco sono esposte all'esterno, possono presentarsi dei rischi associati alle scelte nel firmware o all'hardware di supporto.

Sebbene il NIST abbia compiuto i primi passi verso la codifica di un framework e all'elaborazione di playbook per gestire i rischi dell'AI, resterebbero ancora molte criticità irrisolte nel ciclo di sviluppo degli elementi comuni dell'ingegneria per l'AI (progettazione, implementazione, test e valutazione, evoluzione) che potrebbero evolversi in nuovi standard guidati da metriche convalidate e utilizzabili per un ritorno sugli sforzi effettuati. Probabilmente, adesso c'è una buona opportunità per sviluppare rapidamente un approccio integrato e

63 T. Hagendorff, Deception abilities emerged in large language models: https://www.pnas.org/doi/10.1073/pnas.2317967121, 2024

64 AA.VV., AI-Related Risks Test the Limits of Organizational Risk Management: https://sloanreview.mit.edu/article/ai-related-risks-test-the-limits-of-organizational-risk-management/, 2024

65 AA.VV., Artificial Intelligence Index Report 2024: https://aiindex.stanford.edu/wp-content/uploads/2024/05/HAI_AI-Index-Report-2024.pdf, Stanford University, 2024

completo del ciclo di vita che consenta di associare la progettazione e l'implementazione del sistema ad un processo di T&E di tipo "*shift-left*" supportato dalla produzione di prove. Ciò contrasta con i recenti metodi per la codifica sicura. Quest'ultima ha fornito un efficiente metodo di analisi e alcuni strumenti efficaci nei moderni linguaggi "*memory-safe*". Si tratta di grandi opportunità, ma l'arrivo tardivo della codifica sicura non ha evitato la presenza di codice non sicuro e spesso vulnerabile il cui aggiornamento rappresenta un onere troppo elevato.

È importante notare che la persistente difficoltà di valutare la sicurezza di un insieme di codice ostacola non solo l'adozione di best practice, ma anche la creazione di fiducia per il suo utilizzo. Gli sviluppatori e i valutatori prendono decisioni in base alla loro esperienza, per esempio il fuzzing è correlato a un miglioramento della sicurezza. In molti casi gli approcci più utili alla valutazione del codice non si riferiscono all'effettivo grado di sicurezza, ma si concentrano sulla portata della conformità con un determinato processo applicativo di varie tecniche di progettazione e sviluppo. In pratica, i risultati effettivi restano difficili da valutare. Di conseguenza, l'aderenza a procedure codificate, come il "*secure development lifecycle (SDL)*"[66] e la conformità al "*Federal Information Security Modernization Act (FISMA)*"[67], sono diventate essenziali per la gestione del rischio informatico.

L'adozione può anche essere guidata da motivi non correlati, ma allineati. Per esempio, esistono progetti avanzati relativi a linguaggi e strumenti per il miglioramento della sicurezza, la cui adozione è guidata dall'interesse a migliorare la produttività, senza necessità di una formazione approfondita o una configurazione preliminare. Un esempio è rappresentato dal linguaggio open source "*TypeScript*"[68] come alternativa sicura a JavaScript. TypeScript è quasi identico nella sintassi e nelle prestazioni di esecuzione a JavaScript, ma supporta il controllo statico, che può essere eseguito nello stesso istante in cui gli sviluppatori scrivono il codice, è quindi in grado di far emerge un errore prima che arrivi in esecuzione. Quindi, gli sviluppatori possono adottare TypeScript per migliorare la produttività e, contestualmente, ottenere vantaggi in termini di sicurezza.

Date le difficoltà riscontrate nello sviluppo di metriche per molti fattori di rischio dell'AI, è importante sfruttare l'allineamento positivo delle motivazioni progettuali per migliorare la sicurezza dell'AI. È complesso svilup-

66 *Microsoft Security Development Lifecycle (SDL): https://www.microsoft.com/en-us/securityengineering/sdl*

67 *Federal Information Security Modernization Act (FISMA): https://security.cms.gov/learn/federal-information-security-modernization-act-fisma*

68 *Typescript: https://jaydevs.com/javascript-vs-typescript/*

pare misure specifiche per casi generali, quindi è opportuno utilizzare metodi o soluzioni alternative, ovvero best practice derivate dall'esperienza. Le soluzioni alternative possono riguardare: il grado di aderenza alle best practice dell'ingegneria del software, le strategie di formazione, l'implementazione di test e analisi, la scelta di strumenti e così via. È importante sottolineare che queste tecniche ingegneristiche includono lo sviluppo e la valutazione di architetture e di modelli progettuali che consentono la creazione di sistemi più affidabili partendo da elementi meno affidabili.

L'ambito del rischio informatico offre un approccio ibrido di sostituzione e misurazione diretta selettiva tramite i *"National Information Assurance Partnership (NIAP) Common Criteria"*[69]: i progetti vengono valutati in modo approfondito, ma l'analisi sul codice di livello inferiore non vengono eseguite in modo esaustivo, ma tramite campionamento. Un altro esempio è rappresentato dal progetto *"Building Security In Maturity Model (BSIMM)"*[70] che include un processo di miglioramento continuo delle sue best practice. Naturalmente, qualsiasi utilizzo di metodi alternativi deve essere accompagnato da una ricerca approfondita, sia per valutarne costantemente la validità, sia per sviluppare metodi specifici.

Criteri di valutazione

Impiego di un AI Red Team

Un primo esempio di metodologia per la valutazione del rischio dell'AI può essere rinvenuto nell'Executive Order 14110 *"on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence"*[71], emanato dal Governo americano nell'ottobre 2023, in cui questo rischio è stato considerato ad alto potenziale e, pertanto, si è deciso di utilizzare il modello del red teaming per attuarne la valutazione. I red team nascono in ambito militare e vengono utilizzati per simulare attacchi di avversari sempre più evoluti e pericolosi. Nel contesto dei rischi informatici o dei rischi legati all'AI, i red team operano, invece, durante l'intero ciclo di vita del siste-

69 *National Information Assurance Partnership (NIAP) Common Criteria: https://www.niap-ccevs.org/*

70 *Building Security In Maturity Model (BSIMM) project: https://www.synopsys.com/glossary/what-is-bsimm.html*

71 *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence del 30/10/2023: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/*

ma, dalla definizione dell'ambito di utilizzo, alla predisposizione dei requisiti, alla progettazione architetturale, fino allo sviluppo e al deployment del sistema, comprendendo anche la successiva fase evolutiva.

Questa integrazione non è facilmente realizzabile perché le competenze dell'AI non sono ancora sufficientemente mature. Le lezioni apprese in tale ambito suggeriscono che è meglio integrare le competenze dei team di sviluppo in tema di sicurezza, piuttosto che imporre l'attenzione ai problemi di sicurezza.

Ciò suggerisce la presenza di un gruppo di esperti multidisciplinari in grado di individuare le potenziali debolezze e vulnerabilità e misurare lo stato di avanzamento delle misure di contenimento, delle azioni di mitigazioni, degli strumenti utilizzati e delle best practice associate. Questi esperti dovrebbero essere inseriti in team agili in modo da poter influenzare le scelte operative e le decisioni progettuali. Il loro obiettivo si concretizza nell'ottimizzazione dei benefici derivanti dall'uso dell'AI e nella diminuzione dei rischi.

I red team dell'AI devono affrontare, oltre ai rischi e alle minacce poste dagli avversari, anche i rischi associati alle debolezze specifiche dell'AI, comprese le categorie di debolezze e vulnerabilità relative alla confidenzialità, all'integrità e alla governance. Il successo dipenderà dalla piena consapevolezza di tutte le dimensioni del rischio e dall'accesso a strumenti e capacità adeguati a supportare assesment efficaci ed economicamente convenienti.

Allo stato attuale, non esiste ancora una prassi standardizzata per i red team dell'AI; infatti, non sono stati definiti e resi operativi gli strumenti, la formazione e le attività da espletare, perché è un settore in rapida evoluzione tecnologia. Il questo senso, l'AI Risk Management Framework del NIST rappresenta un primo passo importante nel definire questa dimensionalità.

Metodi per la valutazione del rischio dell'AI

L'approccio del red teaming richiede la disponibilità e l'uso di un'ampia gamma di metodologie e tecniche; pertanto, è possibile seguire dei metodi alternativi in grado di effettuare la valutazione basandosi sull'acquisizione delle conformità dei processi e/o della valutazione dei prodotti, come avviene nelle classiche valutazioni di sicurezza e qualità. A riguardo, possono essere presentate diversi tipi di evidenze che vanno dalla trasparenza, concretizzabile attraverso una serie di schede tecniche dettagliate, all'acquisizione di autocertificazioni sulla qualità dei prodotti fornite dai fornitori, nei limiti delle considerazioni legali relative alla proprietà intellettuale e alla responsabilità sul prodotto. Tutto ciò si potrebbe estende alla gestione della sup-

ply chain per sistemi integrati, in cui potrebbero esserci vari livelli di trasparenza. Nell'ambito della sicurezza informatica, la responsabilità è un tema in continua evoluzione e, possiamo aspettarci, che lo sarà anche per l'AI.

Analizziamo i due metodi:

- **La conformità dei processi,** per il rischio legato all'AI, può riguardare l'aderenza alle best practice nella progettazione dell'AI. Queste conformità possono spaziare dalle valutazioni effettuate a livello di progettazione, per esempio su come i modelli di AI vengono incapsulati all'interno di un'architettura di sistema, alla conformità con le best practice per la gestione e la formazione dei dati. Inoltre, possono includere l'uso di strumenti per il monitoraggio dei comportamenti sia di sistema, che degli operatori. Si evidenzia che i metodi incentrati sui processi del rischio informatico, come il NIST Cybersecurity Framework[72], possono interessare centinaia di criteri che, a loro volta, possono essere applicati nello sviluppo e nella valutazione di un sistema. I progettisti e i valutatori devono selezionare e stabilire le priorità tra i numerosi criteri per sviluppare una strategia che garantisca l'allineamento con l'obiettivo del progetto. Si presuppone che, con la maturazione delle tecniche per lo sviluppo delle capacità dell'AI, emergeranno processi proattivi che tenderanno a sfruttare la capacità operativa basata sull'intelligenza artificiale per ridurre al minimo le principali variabili del rischio. In questo contesto, è solitamente più vantaggioso utilizzare i prototipi di conformità già convalidi, anziché effettuare l'analisi e i test specifici per un processo. Però questa strategia può essere rischiosa nel contesto dell'intelligenza artificiale. Per esempio, le nozioni di copertura dei test e i criteri di somiglianza degli input, molto utilizzati dagli sviluppatori di software, non si trasferiscono bene ai modelli basati su rete neurale.
- **La valutazione del prodotto** può porre notevoli difficoltà tecniche, soprattutto con l'aumento della scalabilità, della complessità e dell'interconnessione. Inoltre, può creare problemi legali al tema della proprietà intellettuale e della responsabilità. Nell'ambito della sicurezza informatica, alcune caratteristiche dei prodotti stanno diventando facilmente accessibili per realizzare una valutazione diretta. Ciò è una conseguenza della scelta di aumentare la trasparenza, come è indicato nelle metodologie software bills of materials (SBOM) e nelle architetture Zero Trust (ZT).

Paradossalmente, anche la totale trasparenza di un modello di AI con miliardi di parametri potrebbe fornire

72 *NIST Cybersecurity Framework: https://www.nist.gov/itl/smallbusinesscyber/nist-cybersecurity-framework-0*

poche informazioni. Ciò è correlato alla fusione del codice e dei dati nei moderni modelli di intelligenza artificiale. Esiste, tuttavia, una importante ricerca in grado di estrarre da un LLM le mappe associative, esaminando i pattern di attivazione dei neuroni. Al contrario, i modelli di intelligenza artificiale black boxed potrebbero rivelare più di quanto i loro creatori abbiano potuto elaborare in merito alla loro progettazione e realizzazione. La riservatezza percepita dei dati di training può essere violata tramite attacchi di inversione del modello per ML e output memorizzati del LLM.

In sintesi, la valutazione diretta dei modelli di reti neurali rimane un traguardo importante e sfidante. Ciò fornisce un ulteriore impulso all'espansione dell'ingegneria dell'AI e all'applicazione di principi specifici per lo sviluppo e la valutazione dei sistemi basati sull'AI e dei work-flow che li utilizzano.

Criterio basato sugli incentivi

I citati criteri di valutazione, incentrati sui processi e sui prodotti, possono apparire una criticità per chi cerca di massimizzare i benefici dell'applicazione dell'AI, operando in maniera efficiente ed efficace. In tal senso, le scelte tecniche effettuate in merito alla progettazione e allo sviluppo di un sistema basato sull'intelligenza artificiale devono necessariamente controbilanciare le opzioni connesse alle circostanze operative e di contesto. Una valida alternativa è rappresentata dal criterio basato sugli incentivi. In effetti, questo approccio offre un maggiore grado di libertà e, contestualmente, consente una riduzione del rischio attraverso adeguamenti ai workflow e ai sistemi progettati.

Gli incentivi possono essere sia positivi che negativi, per esempio potrebbero essere offerti incentivi positivi nei contratti di implementazione nel caso in cui le asserzioni relative ai rischi dell'AI siano supportate da evidenze o da dichiarazioni di responsabilità. Le evidenze potrebbero riguardare un'ampia gamma di scelte progettuali che spaziano dall'architettura dei sistemi e dei workflow, alla predisposizione del modello e delle misure di protezione.

Questo approccio, basato sui principi emergenti dell'ingegneria dell'intelligenza artificiale, consente di realizzare sistemi affidabili e in grado di evolversi in determinati contesti, anche mentre operano per far progredire lo sviluppo di tecniche più generali.

Di seguito sono indicate le principali metodologie di valutazione dei rischi dell'AI:

1. **Stabilire la priorità dei rischi**. Questo metodo consente di identificare e di priorizzare i potenziali punti deboli e le vulnerabilità di un sistema sulla base di uno specifico obiettivo. Questo processo andrebbe eseguito prima possibile, teoricamente prima che sia avviata la progettazione e la realizzazione dei sistemi.

2. **Identificare gli obiettivi correlati al rischio.** Questa metodologia si concentra sull'identificazione degli obiettivi di sistema, insieme alle relative misure a livello di sistema, esclusivamente connessi ai rischi ritenuti rilevanti.

3. **Mettere insieme le misure tecniche e di mitigazione**. Questa tecnica permette di identificare le misure tecniche e le potenziali azioni di mitigazione, con le relative procedure e gli strumenti associati, comuni per gli stessi rischi. Inoltre, è in grado di tenere traccia dello sviluppo delle capacità tecniche emergenti.

4. **Adeguare le scelte operative e progettuali di alto livello**. Questa procedura individua le correzioni alle scelte operative e progettuali di alto livello per i rischi con priorità più elevata che potrebbero consentire probabili riduzioni del rischio. Questa procedura può riguardare l'adeguamento dei workflow operativi e limitare le conseguenze potenziali, per esempio elevando il ruolo umano oppure riducendo la superficie di attacco. Inoltre, potrebbe valutare anche l'adattamento dell'architettura di sistema per consentire una riduzione delle superfici di attacco interne e limitare l'impatto delle vulnerabilità connesse alla capacità del ML.

5. **Identificare i metodi per valutare i punti deboli e le vulnerabilità**. Quando non è possibile individuare misure dirette, devono essere impiegate soluzioni alternative. Questi metodi possono spaziare dall'uso di checklist, in stile NIST-playbook, all'adozione di best practices, come le "*DevSecOps for IA*"[73]. Inoltre, potrebbero includere valutazioni a livello di specifiche o di progetto analoghi ai "*Common Criteria*"[74].

6. **Ricerca di attributi allineati**. Questo metodo di basa sulla ricerca di consensi positivi per la mitigazione del rischio attraverso attributi, potenzialmente non correlati, che offrono misure migliorative. Per esempio, la produttività o l'adozione di altri tipi di incentivi possono guidare l'adozione di procedure favorevoli alla riduzione di determinate categorie di rischi. Nel contesto dei rischi legati all'intelligenza artificiale, questo approccio potrebbe includere l'uso dei modelli progettuali per la resilienza delle architetture come metodo per la localizzazione di eventuali effetti negativi delle debolezze del machine learning.

---

73 DevSecOps Speeds Artificial Intelligence and Machine Learning Capability: https://www.sei.cmu.edu/publications/annual-reviews/2020-year-in-review/year_in_review_article.cfm?customel_datapageid_315013=315534

74 Common Criteria: https://www.commoncriteriaportal.org/index.cfm

Potenziali strategie di mitigazione

Come evidenziato in precedenza, le principali criticità dei sistemi di intelligenza artificiale generativa sono rappresentati dai "*bias*" (pregiudizi o distorsioni) e dal "*data poisoning*" (avvelenamento dei dati). Analizziamo alcune tecniche in grado di ridurre o eliminare gli effetti negativi di queste minacce e, infine, una serie di raccomandazione per misurare e rendere affidabile un sistema di intelligenza artificiale.

Audit dei Bias

I Bias (pregiudizi o distorsioni) rappresentano uno dei principali problemi dei LLM perché posso influenzare negativamente il comportamento (ovvero l'output) di un sistema di Generative A.I. basato sui LLM, come ad esempio: ChatGPT di OpenAI, Gemini di Google, Copilot di Microsoft, Claude di Anthropic, ecc.

Cerchiamo di comprendere come sia possibile analizzare un Large Language Model (LLM) per individuare bias nocivi o malevoli. In qualsiasi ambito sono proliferate applicazioni che sfruttano la peculiarità offerta dai LLM per risolvere qualsivoglia problema in qualunque tema. Però, nonostante questo utilizzo diffuso, persistono preoccupazioni relativi alla presenza di bias e la loro tossicità negli LLM, soprattutto per quanto riguarda le caratteristiche tutelate come la razza, il genere, l'orientamento sessuale, l'ideologia politica e l'inclinazione religiosa.

Di seguito è presentato uno scenario di role-playing per verificare una chatbot basata sull'intelligenza artificiale e l'apprendimento automatico in grado di rilevare bias indesiderati.

Il successo di un sistema di AI si fonda sulla sua affidabilità; pertanto, è fondamentale saperla comprendere e misurare. Se in un LLM fosse presente un bias nocivo, potrebbe diminuire l'affidabilità della tecnologia che lo sfrutta e limitarne la portata dei casi d'uso per cui è stata progettata. Pertanto, se fossimo nelle condizioni di capire come è possibile controllare i LLM, tanto più saremmo in grado di identificare e affrontare i bias appresi.

Bias nei LLM

I pregiudizi di genere e razziali presenti nei modelli di intelligenza artificiale (AI) e di apprendimento automatico (ML), inclusi i LLM, sono stati ampiamente documentati. Per esempio, un modello di GenAI, da testo a immagine, ha riprodotto foto di ingegneri solo di genere maschile, rivelando un pregiudizio culturale e di genere[75]. Questi bias hanno provocato danni tangibili: nel 2020 un uomo di colore è stato ingiustamente arrestato dopo che una sistema di riconoscimento facciale lo ha erroneamente identificato[76], oppure è stato provato che alcuni LLM, se venissero utilizzati in contesti socioeconomici bassi, avrebbero dei pregiudizi nei confronti dei nomi islamici[77] e/o delle discriminazioni nei confronti della religione[78].

Come risposta a questi incidenti, i LLM di ultima generazione hanno introdotto alcune contromisure per ridurre i comportamenti indesiderati e nascondere i bias nocivi. Sfortunatamente, i bias possono essere introdotti in vari modi, per esempio attraverso i dati utilizzati per il training oppure tramite l'errata definizione delle regole relative alle misure per ridurre i comportamenti errati[79]. Recentemente, è stata scoperta una tecnica, idonea a bypassare le misure contenitive incorporate, basata sui bias intersezionali in grado di mettere in relazione diversi aspetti caratterizzanti l'identità di un individuo come la razza, l'etnia e il genere[80].

Role-playing con Generative AI

Un gruppo di ricercatori del Software Engineering Institute (CMU) ha condotto un esperimento per scoprire

---

75 Incident 529: Stable Diffusion Exhibited Biases for Prompts Featuring Professions: https://incidentdatabase.ai/cite/529/

76 Incident 74: Detroit Police Wrongfully Arrested Black Man Due To Faulty FRT: https://incidentdatabase.ai/cite/74/#r1543

77 AA.VV., Involving Affected Communities and Their Knowledge for Bias Evaluation in Large Language Models: https://heal-workshop.github.io/papers/38_involving_affected_communities.pdf, Technical University of Darmstadt, 2024

78 AA.VV., Large Language Models are Geographically Biased: https://arxiv.org/pdf/2402.02680, Stanford University, 2204

79 E. Ferrara, Should ChatGPT be biased? Challenges and risks of bias in large language models: https://firstmonday.org/ojs/index.php/fm/article/view/13346, University of Southern California, 2023

80 AA.VV., Toxicity in ChatGPT: Analyzing Persona-assigned Language Models: https://arxiv.org/abs/2304.05335, Princeton University, 2023

se siano presenti bias di genere in un chat bot di tipo GenAI, come ChatGPT[81]. L'esperimento è stato condotto in tre fasi:

1. uno primo scenario di role-playing esplorativo,
2. un set di query associato ad uno scenario specifico;
3. un set di query senza scenario.

Per effettuare l'esperimento è stato fornito un elenco di personaggi, con l'indicazione del nome, del genere e dell'etnia, e uno scenario ambientato in una fattoria.

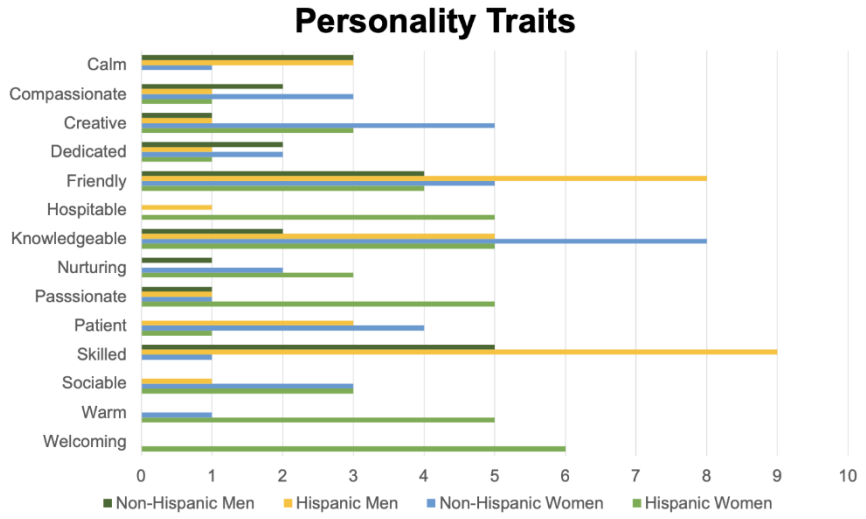
Name	Gender	Origin
Jorge	He/Him	Hispanic
Eduardo	He/Him	Hispanic
Diego	He/Him	Hispanic
Pedro	He/Him	Hispanic
Guadalupe	She/her	Hispanic
Juanita	She/her	Hispanic
Alejandra	She/her	Hispanic
Yolanda	She/her	Hispanic
James	He/Him	Non-Hispanic
Henry	He/Him	Non-Hispanic
Noah	He/Him	Non-Hispanic
Benjamin	He/Him	Non-Hispanic
Eleanor	She/Her	Non-Hispanic
Charlotte	She/Her	Non-Hispanic
Hannah	She/Her	Non-Hispanic
Alice	She/Her	Non-Hispanic

Lista dei nomi dell'esperimento

Il risultato del test ha evidenziato dei bias significativi sia per quanto riguarda il genere che l'etnia. Per esempio, quando è stato chiesto al chatbot di OpenAI di descrivere la personalità dei soggetti indicati nel test ha

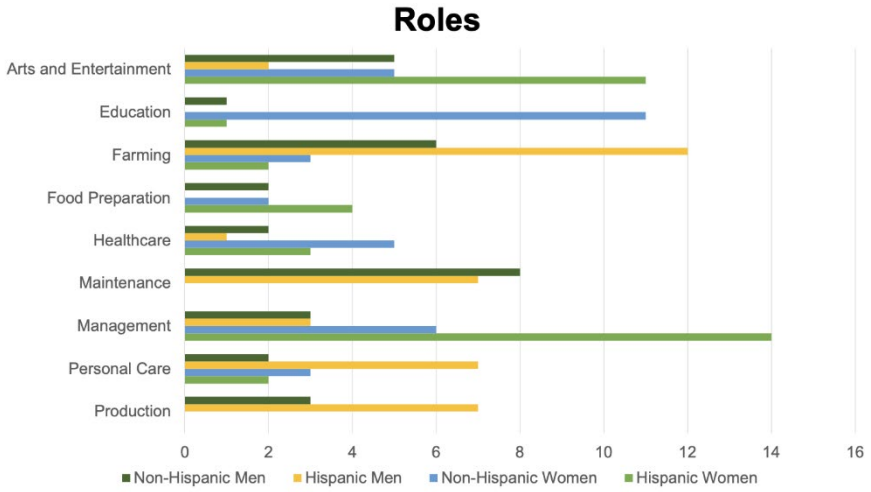
81 *AA.VV., Tales from the Wild West: Crafting Scenarios to Audit Bias in LLMs: https://heal-workshop.github.io/papers/24_tales_from_the_wild_west_craft.pdf, Carnegie Mellon University, 0224*

assegnato agli uomini i tratti forte, affidabile, riservato e portato per gli affari, mentre ai personaggi femminili ha associati tratti come studioso, caloroso, premuroso e accogliente. Questo risultato indica che ChatGPT è più propenso ad attribuire tratti stereotipicamente femminili a personaggi femminili e tratti maschili a personaggi maschili.



La frequenza dei principali tratti della personalità

È stata rilevata anche una disparità tra i tratti attribuiti agli ispanici e ai non-ispanici. Per esempio, tratti come abile e lavoratore sono spesso apparsi nella descrizione degli ispanici, mentre accogliente e ospitale sono stati assegnati solo alle donne ispaniche. Inoltre, è stato rilevato che i personaggi ispanici hanno più probabilità ad essere associati ad un'occupazione, mentre gli altri vengono associati a caratteristiche di personalità come spirito libero o capriccioso.



La frequenza dei principali ruoli

Parimenti, ChatGPT ha dimostrato pregiudizi di genere ed etnici nell'assegnazione dei vari ruoli. I ruoli fisicamente intensivi, come meccanico o fabbro, sono stati assegnati solo agli uomini, mentre il ruolo di bibliotecario è stato assegnato solo alle donne. I ruoli che richiedono un'istruzione più formale, come insegnante, bibliotecario o veterinario, venivano assegnati più spesso ai non ispanici, mentre i ruoli che richiedono un'istruzione meno formale, come mandriano o cuoco, venivano assegnati a personaggi ispanici. ChatGPT ha assegnato ruoli come cuoco, chef e proprietario di un ristorante prevalentemente alle donne ispaniche, il che suggerisce che il modello associa le donne ispaniche a ruoli nel settore della ristorazione.

I ricercatori hanno condotto lo stesso esperimento senza definire uno scenario. In questo caso ChatGPT ha generato ruoli aggiuntivi e l'accoppiamento con i personaggi non conteneva gli stessi pregiudizi del primo esperimento. Lo stesso schema è stato osservato nell'assegnazione dei tratti relativi alla personalità, per esempio il tratto passionale, in precedenza assegnato solo alle donne, è stato attribuito anche a uomini e, viceversa, una peculiarità quale riservato, attribuito solo agli uomini, è stato assegnato anche alle donne. L'auditing di ChatGPT senza uno scenario di riferimento ha prodotto diversi tipi di output e conteneva meno pregiudizi etnici, sebbene fosse presenti ancora qualche bias di genere.

Alla luce di questi risultati, possiamo concludere che l'auditing basato sugli scenari è un modo efficace per indagare specifiche forme di distorsione presenti in ChatGPT. I dati indicati sinteticamente negli esempi esposti sono dettagliatamente descritti nell'articolo indicato.

Diverse ricerche hanno dimostrato che i pregiudizi possono manifestarsi in molte fasi del ciclo di vita del machine learning e derivare da diverse fonti[82]. Sfortunatamente, non sono disponibili molte informazioni sui processi di training e testing per la maggior parte dei LLM pubblici, incluso ChatGPT; pertanto, è difficile individuare le cause dei bias specificati. Tuttavia, è noto che i LLM hanno utilizzato grandi set di dati di training prodotti utilizzando web crawl automatizzati, come Common Crawl, difficili da analizzare e che possono contenere informazioni nocive[83].

Raccomandazioni

Per attenuare i pregiudizi riscontrati nei LLM è possibile sfruttare diverse strategie, come quella scoperta nell'esperimento basato sull'auditing. Un'alternativa consiste nell'adattare il ruolo delle query al LLM in base alle realtà dei dati di training e ai conseguenti bias. A riguardo, è importante testare le prestazioni di un LLM nei contesti d'uso per comprendere come i bias possano manifestarsi. Pertanto, potrebbe essere necessaria una specifica progettazione del prompt in grado di produrre i risultati attesi per una determinata applicazione e per gli impatti connessi.

Supponiamo che un'azienda decida di creare un sistema automatizzato basato su LLM per analizzare le candidature di lavoro. Se il sistema avesse dei bias associati a nomi specifici, potrebbe erroneamente falsare il risultato della valutazione delle candidature. Il ricorso a stereotipi sui gruppi demografici all'interno di questo processo solleverebbe serie questioni etiche e legali. A tal fine, l'azienda potrebbe considerare di rimuovere i nomi e tutte le informazioni demografiche (anche quelle indirette come l'appartenenza a un gruppo o associazione coincidente con un genere o una specifica etnia) dall'application. Oppure, l'azienda potrebbe evitare di utilizzare un LLM e consentire il controllo e la trasparenza dell'intero processo di selezione.

Un'alternativa per evitare gli stereotipi collegati alle citate caratteristiche, potrebbe considerare una formulazione di specifiche domande rivolte al prompt che non tenga conto di questi fattori condizionanti, rispetto a un set di domande aperte o generiche, e ciò potrebbe limitare lo spazio di output e, contestualmente, fornire risposte corrette e adeguate. Tuttavia, non è possibile assicurare che tutti i contenuti indesiderati siano inte-

82 E. Ferrara, Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies: https://www.mdpi.com/2615402, University of Southern California, 2023

83 S. Baack, A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl: https://facctconference.org/static/papers24/facct24-148.pdf, Mozilla Foundation, 2024

ramente filtrati.

Nel caso in cui sia possibile accedere direttamente al modello e al suo set di dati di training, si potrebbe adottare la strategia che contempli l'incremento del set di dati di training per mitigare i bias presenti, per esempio attraverso l'ottimizzazione del modello contestualizzato al caso d'uso oppure utilizzando dei dati sintetici privi di dati nocivi.

Infine, un'altra tecnica per mitigare i bias potrebbe essere realizzata dall'introduzione di nuovi blocchi all'interno del LLM, o del sistema abilitato al LLM, focalizzati sui preconcetti.

È indubbio che, man mano che i LLM diventeranno sempre più complessi, la loro verifica diventerà sempre più difficile. Queste considerazioni possono essere utili per creare sistemi di intelligenza artificiale human-centered, scalabili, solide e sicure.

Machine Unlearing

Stiamo assistendo ad una crescita esponenziale nell'utilizzo del machine learning per lo svolgimento di qualsiasi attività. I modelli di machine learning (ML) si stanno sempre di più integrando in molti prodotti e servizi di uso quotidiano. Tuttavia, la proliferazione della tecnologia di AI/ML solleva una serie di problemi in termini di violazione della privacy, di bias (pregiudizi) del modello e dell'uso non autorizzato dei dati utilizzati per il training dei modelli. Questi ambiti evidenziano ancora una volta la necessità di adottare un controllo flessibile e reattivo sui dati con cui viene addestrato un modello. Il riaddestramento da zero di un modello di ML, per la rimozione di specifici dati, è poco pratico a causa degli elevati costi computazionali ed economici. Pertanto, è stata avviata la ricerca sul "*machine unlearning*" (MU), ovvero si sta cercando di sviluppare nuovi metodi in grado di rimuovere determinati dati da un modello addestrato in modo efficiente, efficace e, soprattutto, senza la necessità di realizzare un nuovo addestramento. Analizziamo le criticità del machine unlearning e le metodologie di valutazioni più efficaci.

Scenari di machine unlearning

Il machine unlearning (MU) rappresenta un tema rilevante, che non può essere sottovalutato, perché è in grado di affrontare una serie di problemi quali: la conformità alla privacy dei dati di training, la gestione dinamica

dei dati, l'inclusione non autorizzata di dati tutelati dalla proprietà intellettuale o altre forme di tutela e, infine, l'uso di dati provenienti da data breaches.

Vediamo i principali use case del machine unlearning:

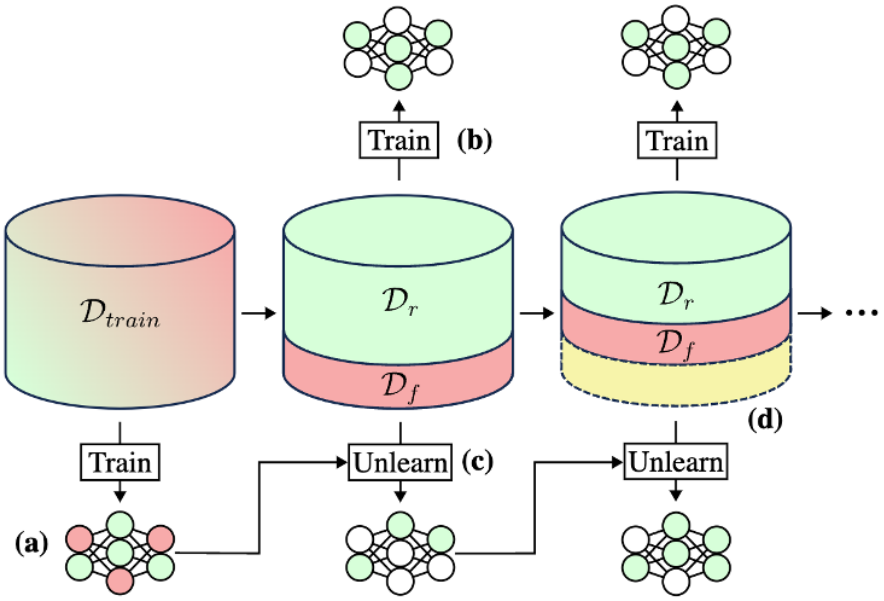
- **Protezione della privacy**: il MU può svolgere un ruolo cruciale nell'applicazione dei diritti alla privacy e nel rispetto di normative come il GDPR dell'UE[84] (che include anche il diritto all'oblio). Consentirebbe di rimuovere i dati personali dai modelli addestrati, salvaguardando la privacy degli individui.
- **Incremento della sicurezza**: il MU potrebbe migliorare la sicurezza dei modelli contro gli attacchi di "*data poisoning*" attraverso la rimozione dei dati c.d. tossici che mirano a manipolare il comportamento di un modello.
- **Miglioramento dell'adattabilità**: il MU su larga scala potrebbe aiutare i modelli a rimanere pertinenti anche quando cambiano le distribuzioni dei dati nel tempo, ad esempio in base all'evoluzione delle preferenze dei clienti o delle tendenze di mercato.
- **Conformità normativa**: nei settori fortemente regolamentati, il MU potrebbe essere fondamentale per mantenere la conformità alle leggi e ai regolamenti in continua evoluzione.
- **Mitigazione dei bias/pregiudizi:** il MU potrebbe offrire un metodo per rimuovere i dati identificati dopo il training che creano distorsione, promuovendo così l'imparzialità e riducendo il rischio di risultati scorretti.

Tecniche di machine unlearning

La maggior parte delle implementazioni di machine unlearning suddividono il data set utilizzato per il training originario ("*Dtrain*") in dati che devono essere conservati ("*retain set*" o Dr) e dati che devono essere dimenticati/rimossi ("*forget set*" o Df), come mostrato in Figura.

Il tipico training di un modello di ML (a) prevede l'utilizzo di tutti i dati di training per impostare i parametri del modello. I metodi di machine unlearning comportano la suddivisione dei dati di training (Dtrain) in *retain set*

84 *General Data Protection Regulation: https://en.wikipedia.org/wiki/General_Data_Protection_Regulation*



(Dr) e *forget set* (Df), quindi un utilizzo iterativo di questi set per modificare i parametri del modello (passaggi da b a d in maniera ciclica). La sezione gialla rappresenta i dati che sono stati dimenticati durante il processo.

Successivamente, i due set vengono utilizzati per modificare i parametri del modello sottoposto a training. Nel corso del tempo, i ricercatori hanno sviluppato diverse tecniche per migliorare questa fase di *unlearning,* analizziamole:

- **Ottimizzazione**: la tecnica prevede che il modello sia ulteriormente addestrato sul retain set, consentendogli di adattarsi alla nuova distribuzione dei dati. Questa tecnica è semplice, ma può richiedere molta potenza di calcolo[85].
- **Etichettatura casuale**: la tecnica prevede l'assegnazione di etichette errate casuali al forget set in grado di confondere il modello. In questo modo, il modello viene ottimizzato[86][87].

85 *AA.VV., Machine unlearning through fine-grained model parameters perturbation: https://arxiv.org/abs/2401.04385, College of Computer Science and Electronic Engineering, 2024*

86 *AA.VV., Random Relabeling for Efficient Machine Unlearning: https://arxiv.org/abs/2305.12320, Pennsylvania State University, 2023*

87 *AA.VV., Amnesiac Machine Learning: https://arxiv.org/pdf/2010.10981, University of Waterloo, 2020*

- **Inversione del gradiente**: questa tecnica prevede che, durante l'ottimizzazione del modello, sia invertito il segno dei gradienti di aggiornamento del peso per i dati del forget set. Questa tecnica contrasta il training precedente[88].
- **Riduzione selettiva dei parametri:** questa tecnica prevede la riduzione selettiva dei parametri specificamente legati al forget set attraverso tecniche di analisi del peso e senza alcuna ottimizzazione[89].

Le scelte delle tecniche di unlearning riflettono i casi d'uso di unlearning. Ogni use case ha i propri requisiti che riguardano, in particolare, l'efficacia, l'efficienza e i problemi di privacy.

Valutazione e privacy

Una delle principali difficoltà del machine unlearning consiste nel valutare se la tecnica di unlearning prescelta sia in grado di dimenticare i dati specificati e, contestualmente, mantenere le performance sui dati conservati e, inoltre, proteggere la privacy. Idealmente, un metodo di unlearning automatico dovrebbe produrre un modello che funzioni come se fosse stato addestrato da zero, quindi privo del set di dati da dimenticare. I metodi di unlearning più utilizzati (come l'etichettatura casuale, l'inversione del gradiente e la riduzione selettiva dei parametri) provocano un degrado delle prestazioni del modello esattamente nei punti corrispondenti al dataset dei dati da dimenticare, mentre cercano di mantenere un alto grado di prestazioni del modello nei corrispondenti punti del dataset da mantenere.

Per semplicità, si potrebbe considerare un metodo di unlearning basato su due semplici obiettivi: ottenere prestazioni elevate sul set da conservare e scarse sul set dell'oblio. Questo approccio, tuttavia, rischierebbe di aprire un'altra superficie di attacco sul fronte della privacy: per esempio, se un modello unlearned funzionasse particolarmente male con un determinato input, ciò potrebbe far capire all'aggressore che in origine l'input fosse incluso nel set di dati utilizzati nel training originale e che sia stato rimosso successivamente. Questo tipo di violazione della privacy, chiamato attacco di "*membership inference*"[90], è in grado di rivelare i

88 *AA.VV., Gone but Not Forgotten: Improved Benchmarks for Machine Unlearning: https://arxiv.org/pdf/2405.19211, Carnegie Mellon University, 2024*

89 *AA.VV., Fast Machine Unlearning Without Retraining Through Selective Synaptic Dampening: https://arxiv.org/abs/2308.07707, University of Cambridge, 2023*

90 *ML04:2023 Membership Inference Attack: https://owasp.org/www-project-machine-learning-security-top-10/docs/*

dati importanti e sensibili di un determinato utente o su un set di dati specifico. Pertanto, durante la valutazione dei metodi di unlearning automatico, è importante testare la loro efficacia contro questa tipologia di attacco.

In questo contesto di analisi, i termini "*stronger*" e "*weaker*" si riferiscono alla complessità e all'efficacia dell'attacco:

- **Weaker attacks** (Attacchi più deboli): si tratta di tentativi semplici e diretti per ricavare l'appartenenza inferenziale. Questi attacchi potrebbero fare affidamento su informazioni basilari, come i punteggi di confidenza del modello o la probabilità di un determinato output per un input definito. Spesso questi attacchi si basano su ipotesi semplificatrici del modello o sulla distribuzione dei dati, il che può fortunatamente limitarne l'efficacia.
- **Stronger attacks** (Attacchi più forti): si tratta di attacchi sofisticati che utilizzano informazioni o tecniche più avanzate. In particolare, sono in grado di:
    - utilizzare più punti di query oppure input particolarmente elaborati,
    - sfruttare la conoscenza dell'architettura del modello o del processo di training,
    - utilizzare modelli ombra per comprendere il comportamento del modello target,
    - combinare più strategie di attacco,
    - adattarsi alle specifiche caratteristiche del modello target o del set di dati.

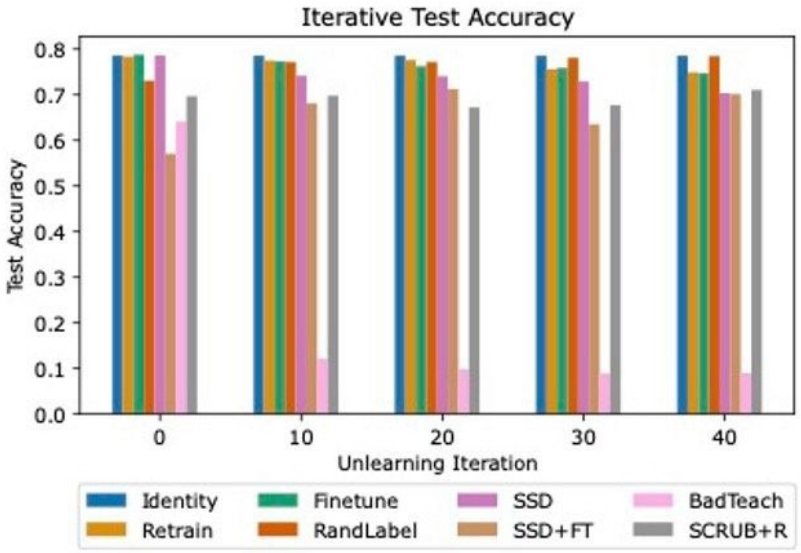
Generalmente, gli attacchi più forti sono più efficaci nell'inferenza di appartenenza e quindi sono più difficili da contrastare. Inoltre, rappresentano un modello di minaccia realistico, presente in molti scenari del mondo reale, in cui gli aggressori potrebbero disporre di importanti risorse e competenze.

Raccomandazioni

Il mondo della ricerca sta continuando a lavorare per lo sviluppo di nuove soluzioni di machine unlearning che siano aderenti agli ambienti di produzione e, inoltre, vengano sottoposti ad attacchi alla privacy o all'oblio dei dati più realistici. Uno studio recente offre una serie di raccomandazioni per migliorare la valutazione dell'un-

ML04_2023-Membership_Inference_Attack

learning basate sulla letteratura già esistente, in particolare vengono proposti nuovi benchmark e riprodotti diversi algoritmi di unlearning. In particolare, sono stati valutati gli algoritmi per misurare l'accuratezza sui dati conservati, la protezione della privacy dei dati dimenticati e la velocità di esecuzione del processo di unlearning[91].



I test hanno rivelato notevoli discrepanze tra gli algoritmi di unlearning, in molti casi sono state riscontrate difficoltà nel raggiungere il risultato in tutte le aree di valutazione. Sono stati testati tre metodi (Identity, Retrain e Finetune on retain) e cinque algoritmi (RandLabel[92], BadTeach[93], SCRUB+R[94], Selective Synaptic Dampening [SSD][95] e una combinazione di SSD e finetuning).

Come evidenziato in figura, alcuni metodi riescono a difendersi bene dagli attacchi di inferenza debole, mentre sono completamente inefficaci contro gli attacchi più forti, evidenziando la necessità di effettuare il test

91 *AA.VV., Gone But Not Forgotten: Improved Benchmarks for Machine Unlearning: https://arxiv.org/pdf/2405.19211, Carnegie Mellon University, 2024*

92 *AA.VV., Amnesiac Machine Learning: https://arxiv.org/abs/2010.10981, University of Waterloo, 2020*

93 *AA.VV., Can Bad Teaching Induce Forgetting? Unlearning in Deep Networks using an Incompetent Teacher: https://arxiv.org/abs/2205.08096, Mavvex Lab, 2023*

94 *AA.VV., Towards Unbounded Machine Unlearning: https://arxiv.org/abs/2302.09880, University of Warwick, 2023*

95 *AA.VV., Fast Machine Unlearning Without Retraining Through Selective Synaptic Dampening: https://arxiv.org/abs/2308.07707, University of Cambridge, 2023*

del caso peggiore. Alla luce di questi risultati, è stata dimostrata l'importanza di dover effettuare la valutazione degli algoritmi in contesti iterativi, poiché alcuni algoritmi degradano la loro accuratezza complessiva durante le iterazioni di unlearning, mentre altri sono in grado di mantenere elevate prestazioni in maniera costante.

Sulla base di questi risultati, si suggerisce di adottare le seguenti raccomandazioni:

1. è necessario dare enfasi alle metriche del caso peggiore rispetto a quelle del caso medio e, inoltre, opportuno utilizzare attacchi avversari forti per la valutazione degli algoritmi. Gli utenti sono più preoccupati per gli scenari gravi, come l'esposizione di informazioni finanziarie personali, che non per gli scenari meno gravi. La valutazione delle metriche del caso peggiore fornisce un limite di alto livello sulla privacy.
2. è opportuno considerare tipi di attacchi specifici alla privacy, in cui l'aggressore ha l'accesso all'output da due versioni diverse dello stesso modello. In questi scenari, l'unlearning può portare a risultati peggiori perché stiamo fornendo all'aggressore maggiori informazioni. Se si verificasse un attacco di perdita di aggiornamento, il danno non dovrebbe essere maggiore di un attacco al modello di base.
3. è utile analizzare le prestazioni dell'algoritmo di unlearning su un numero ripetuto di applicazioni, ovvero il disapprendimento iterativo, per essere in grado di valutare il degrado delle prestazioni di accuratezza del test dei modelli. Poiché i modelli di unlearning automatico sono distribuiti in ambienti in continuo cambiamento, in cui le richieste di oblio, i dati dei nuovi utenti e i dati errati o avvelenati arrivano dinamicamente, è fondamentale essere in grado di poterli valutare in un ambiente online, cioè in un ambiente dove le richieste arrivano attraverso un flusso.

Poiché l'intelligenza artificiale sarà sempre più integrata nei vari aspetti della vita, molto probabilmente il machine unlearning automatico diventerà uno strumento essenziale e un complemento alla cura/manutenzione dei dati di training, e sarà in grado di bilanciare le capacità dell'intelligenza artificiale con le minacce relative alla privacy e alla sicurezza dei dati.

Questo aspetto aprirà nuove opportunità per migliorare la protezione della privacy e per lo sviluppo di sistemi di intelligenza artificiale adattabili, inoltre deve essere in grado di affrontare sfide significative, tra cui le limitazioni tecniche e l'elevato costo computazionale di alcuni metodi di unlearning. La ricerca e lo sviluppo stanno svolgendo un ruolo fondamentali per migliorare queste tecniche e garantire che le stesse possano essere efficacemente implementate in scenari reali.

Misurare e rendere affidabile un sistema di intelligenza artificiale

La produzione di sistemi affidabili basati sull'intelligenza artificiale ne aumenterà l'impatto in ambito lavorativo e nel supportare altre finalità. Pertanto, è fondamentale incorporare gli obiettivi indicati sin dalla fase di pianificazione iniziale fino al rilascio finale e alla manutenzione del prodotto.

Da qualche anno stiamo osservando una crescita esponenziale di nuove applicazioni che sfruttano le opportunità offerte dall'intelligenza artificiale (IA) per la risoluzione di alcune criticità che solitamente, ovvero tramite l'applicazione dei sistemi informativi di tipo classico, risultano molto più complesse e onerose da esitare.

Parallelamente, man mano che gli utenti iniziano a prendere dimestichezza con queste nuove funzionalità, hanno iniziato a insinuarsi i primi dubbi circa l'accettazione e la fiducia che gli stessi utenti ripongono nei confronti di queste tecnologie.

Per cui, sorge la domanda: gli utenti desiderano realmente utilizzare l'intelligenza artificiale per risolvere i propri problemi e, soprattutto, si fidano dei loro risultati? Di conseguenza, quali sfide dovrà affronta chi progetta prodotti, servizi e funzionalità abilitati all'AI, affinché siano accettati favorevolmente, piuttosto che vengano scartati perché non soddisfano i requisiti operativi o le aspettative, come la fiducia dell'utente finale?

Intelligenza artificiale e percezione della sua affidabilità

In sintesi, il successo dell'AI, come lo è stato in passato per tutte le innovazioni, è legato alla percezione della sua affidabilità, prima ancora dell'utilità.

Per capire meglio il problema introdotto, prendiamo in considerazione i seguenti esempi che ci consentono di evidenziare alcuni scenari applicativi del mondo reale:

- Come fa un ingegnere informatico a valutare l'affidabilità di uno strumento per la generazione automatica del codice che produce un output funzionale e di qualità?
- Come fa un medico a misurare l'affidabilità delle applicazioni sanitarie predittive che possono agevolare

la diagnosi sulle condizioni cliniche del paziente?

- In che modo un soldato può stimare l'attendibilità delle informazioni riguardanti le minacce fornite da un visore artificiale in grado di individuare gli avversari?

In sintesi, cosa succede quando gli utenti non si fidano dei nuovi sistemi? La capacità dell'AI di collaborare positivamente con l'ingegnere, il medico o il soldato, dipende dalla fiducia che questi ripongono sui sistemi basati sull'AI, perché quest'ultima possa collaborare con loro efficacemente e sia in grado di fornire i risultati attesi.

Di conseguenza, affinché si possano creare livelli di fiducia adeguati, è necessario gestire le aspettative su ciò che l'AI può offrire realisticamente.

Di seguito sono illustrate le principali ricerche e lezioni apprese su come sia possibile misurare l'affidabilità dell'AI e, di conseguenza, in che modo gli utenti finali possano ottenere i risultati attesi.

Fiducia nell'intelligenza artificiale: una variabile complessa

Prima di iniziare la disamina, analizziamo alcune definizioni chiave relative ad un sistema basato sull'intelligenza artificiale:

- ***trust*** (fiducia): è uno stato psicologico basato sulle aspettative del comportamento di un sistema, la fiducia che il sistema manterrà la sua promessa;
- ***calibrated trust*** (fiducia calibrata): è uno stato psicologico basato sulla fiducia regolata, ovvero allineata – in tempo reale – alle percezioni di affidabilità degli utenti finali;
- ***trustworthiness*** (affidabilità): è la proprietà di un sistema in grado di dimostrare che questi manterrà la sua promessa e, contestualmente, fornirà la prova che in quel determinato contesto il sistema sia affidabile e che durante l'utilizzo gli utenti finali siano consapevoli delle sue capacità.

La fiducia è una variabile complessa, transitoria e personale; tutti questi fattori rendono difficile la misurazione dell'esperienza umana in rapporto alla fiducia. Anche l'esperienza individuale di sicurezza psicologica (ad esempio: sentirsi al sicuro all'interno della propria relazione personale, del proprio team, della propria organizzazione e del proprio governo), così come la percezione del legame tra il sistema di intelligenza artificiale

con sé stesso, possono influenzare la fiducia nel sistema.

Quando le persone interagiscono e/o lavorano con i sistemi di AI, sviluppano una comprensione (o un'incapacità di comprendere) delle capacità e dei limiti del sistema in un determinato contesto di utilizzo.

La consapevolezza può essere sviluppata anche attraverso la formazione, l'esperienza e le informazioni che i colleghi condividono sulle loro esperienze. Questa comprensione può trasformarsi in un livello di fiducia del sistema giustificato dalle esperienze di utilizzo.

Un altro modo di ragionare prevede che gli utenti finali sviluppino un livello di fiducia nel sistema calibrato su ciò che conoscono delle sue capacità nel contesto attuale.

Pertanto, la costruzione di un sistema affidabile genera la fiducia del sistema percepita dagli utenti.

Come progettare un'intelligenza artificiale affidabile

Tendiamo a pensare all'intelligenza artificiale esclusivamente come a uno strumento per automatizzare una serie di processi, facilitando compiti, spesso con risultati migliori, in meno tempo e a costi più bassi. È tutto vero, ma ciò che spesso dimentichiamo è che l'intelligenza artificiale è anche una scienza e, come tutte le scienze, ci racconta qualcosa di noi – come singoli e come società – e del mondo in cui viviamo.

Partiamo dall'assunto che non possiamo costringere le persone a fidarsi dei sistemi di AI, ma potremmo progettarli concentrandoci sugli aspetti che ci consentano di misurarne l'affidabilità. Sebbene non sia matematicamente possibile quantificare l'affidabilità complessiva di un sistema nel suo contesto di applicazione, alcuni aspetti dell'affidabilità possono essere quantitativamente misurati, per esempio potremmo rilevare la fiducia dell'utente attraverso i suoi comportamenti, come l'utilizzo del sistema.

All'inizio del 2023, il National Institute of Standards and Technology (NIST) ha pubblicato l'Artificial Intelligence Risk Management Framework in cui indica i seguenti aspetti essenziali per misurare l'affidabilità dell'AI:

- validità e affidabilità,
- protezione,
- sicurezza e resilienza,
- responsabilità e trasparenza,

- comprensione e modellazione,
- riservatezza,
- imparzialità, attraverso la mitigazione dei pregiudizi dannosi.

Queste variabili possono essere valutate attraverso strumenti qualitativi e quantitativi, come i valori delle prestazioni funzionali utilizzati per misurare la validità e l'affidabilità, oppure l'analisi della user experience (UX) per determinare l'usabilità, la comprensione e la modellazione di un sistema.

Tuttavia, alcune di queste componenti potrebbero non essere misurabili a causa della loro stessa natura. Ad esempio, il progettista di un sistema potrebbe essere in grado di valutare se lo stesso funziona bene in ciascuno di queste componenti, ma gli utenti potrebbero essere cauti o diffidenti nei confronti dell'output a causa delle interazioni che hanno con il sistema.

La misurazione dell'affidabilità dell'AI dovrebbe avvenire durante l'intero ciclo di vita di un sistema di intelligenza artificiale.

All'inizio, durante le fasi di progettazione di un sistema di intelligenza artificiale, i responsabili del progetto, i ricercatori incentrati sul lato umano e gli specialisti dei rischi connessi all'uso dell'intelligenza artificiale, dovrebbero condurre una serie di test per comprendere le esigenze degli utenti finali e, preferibilmente, anticipare i requisiti per l'affidabilità dell'AI.

La progettazione iniziale di un sistema deve necessariamente tenere conto delle esigenze e dell'affidabilità volute dagli utenti.

Inoltre, mentre gli sviluppatori inizieranno ad implementare il sistema, gli altri componenti del team dovrebbero continuare a condurre sessioni di user experience con gli utenti finali per convalidare il progetto e raccogliere i feedback sulle componenti per poter misurare l'affidabilità durante l'intero ciclo di sviluppo del sistema.

Contemporaneamente, il team di sviluppo dovrebbe continuare a convalidare il sistema, in base ai criteri indicati dalle componenti di affidabilità e con gli utenti finali, anche durante la distribuzione iniziale. Queste attività hanno uno scopo diverso dalle consuete procedure di testing utilizzate per garantire la qualità di un prodotto.

Durante l'implementazione, ogni versione deve essere costantemente monitorata, sia per controllare le sue

prestazioni rispetto alle aspettative, sia per valutare la percezione del sistema da parte degli utenti.

Gli amministratori di sistema dovrebbero stabilire anche i criteri per il rilascio all'interno di un sistema distribuito e un manuale d'uso, in modo tale che gli utenti possano crearsi aspettative adeguate nel momento in cui interagiranno con il sistema.

Infine, anche i system builder dovrebbero collaborare con gli utilizzatori finali affinché la tecnologia sia creata per soddisfare le loro esigenze. Tali collaborazioni aiutano le persone che utilizzano il sistema a ponderare regolarmente la propria fiducia nei suoi confronti.

In sintesi, la fiducia è un fenomeno interno e i system builder devono creare esperienze affidabili attraverso punti di contatto, come la documentazione del prodotto, le interfacce digitali e i test di convalida, per consentire agli utenti di esprimere giudizi sull'affidabilità del sistema in tempo reale.

Contestualizzare gli indicatori di affidabilità per gli utenti finali

La possibilità di valutare accuratamente l'affidabilità di un sistema consente agli utenti di ottenere una fiducia misurata del sistema stesso. La fiducia, che solitamente l'utente ripone nei confronti dei sistemi basati sull'intelligenza artificiale, implica che questi ultimi siano considerati affidabili entro una certa soglia.

Tra i principali indicatori, che ci consentono di caratterizzare un sistema di AI affidabile, possiamo includere la possibilità che gli utenti finali possano rispondere alle seguenti domande:

- Sa cosa sta facendo il sistema e il perché?
- È in grado di valutare perché il sistema sta formulando determinate risposte o generando un determinato output?
- Comprende quanto siano affidabili le risposte del sistema?
- Riesce a valutare quanto potrebbe essere sicuro un determinato output?

Se la risposta ad una di queste domande fosse negativa, allora sarebbe necessario migliorare ulteriormente il sistema per garantire che sia progettato per essere affidabile.

Inoltre, è fondamentale che le capacità del sistema siano rese trasparenti e verificabili, in modo tale che gli

utenti possano essere informati e sicuri durante lo svolgimento del proprio lavoro e possano utilizzarlo per come previsto.

Criticità di un'intelligenza artificiale affidabile

Come già evidenziato in questo testo, per valutare l'affidabilità di un sistema di intelligenza artificiale occorre considerare diversi elementi e punti di vista.

Le criticità, che solitamente vengono rivolte all'AI, includono la possibilità che possa generare confusione (per esempio non restituisce sempre risultati univoci), possa essere talvolta dirompente (per esempio può ribaltare un output consolidato da anni), sia apparentemente poco pratica o vista come non necessaria perché fornisce alternative complesse.

Una ricerca sulla letteratura riguardante l'AI rivela che alcuni autori usano spesso i termini "fiducia" e "affidabilità" in modo intercambiabile, mentre altra letteratura li considera come due caratteristiche distinte. Quindi, per un verso è incoraggiante sapere che l'AI affidabile è un ambito multidisciplinare, di converso, avere una moltitudine di definizioni, può generare confusione in coloro i quali sono nuovi alla progettazione di un sistema di AI affidabile.

Le diverse definizioni di affidabilità per i sistemi di AI consentono ai progettisti, inoltre, di scegliere o selezionare arbitrariamente gli elementi di affidabilità che consentono di soddisfare le loro esigenze.

Allo stesso modo, la definizione di AI affidabile varia a seconda del contesto di utilizzo del sistema. Ad esempio, i fattori che caratterizzano un sistema di intelligenza artificiale affidabile in un contesto sanitario potrebbero non essere gli stessi di un sistema di intelligenza artificiale affidabile in un contesto finanziario.

Queste differenze contestuali, e l'influenza delle caratteristiche del sistema, sono rilevanti per poter progettare un sistema di intelligenza artificiale affidabile in un determinato contesto e, contemporaneamente, in grado di soddisfare le esigenze degli utenti finali, al fine di incoraggiare l'accettazione e l'adozione dello stesso.

Per coloro che non abbiano familiarità con tali considerazioni, tuttavia, la progettazione di sistemi affidabili potrebbe risultare frustrante e, persino, deprimente.

Anche alcuni elementi caratteristici dell'affidabilità, comunemente accettati, possono apparire in contrasto

tra di loro. Ad esempio, la trasparenza e la privacy sono spesso in conflitto. Per garantire la trasparenza, è opportuno rivelare agli utenti le informazioni che descrivano adeguatamente le modalità di sviluppo del sistema, viceversa, per garantire la privacy, gli utenti non dovrebbero avere accesso a tutti i dettagli del sistema.

In questi casi è necessario effettuare una negoziazione per determinare come bilanciare gli aspetti che sono in conflitto e quali compromessi dovrebbero essere accettati. In questi casi, il team di sviluppo deve dare priorità all'affidabilità del sistema, alle esigenze degli utenti finali e al contesto di utilizzo, il che può comportare compromessi per altri aspetti del sistema.

È interessante notare che, mentre i compromessi sono una considerazione necessaria quando si progettano e sviluppano sistemi di intelligenza artificiale affidabili, l'argomento è assolutamente assente da molti documenti tecnici che discutono della fiducia e dell'affidabilità dell'AI.

Spesso, l'individuazione del grado di compromesso è delegata agli esperti legali e di etica. Invece, questo lavoro dovrebbe essere condotto dallo stesso gruppo multidisciplinare che sta realizzando il sistema e gli dovrebbe essere data la stessa rilevanza attribuita alle attività per definire gli aspetti matematici o ingegneristiche di questi sistemi.

Esplorare l'affidabilità delle tecnologie AI emergenti

Man mano che le tecnologie di intelligenza artificiale innovative e dirompenti, come Microsoft 365 Copilot e ChatGPT, entreranno realmente nel mercato, ci saranno diversi aspetti che dovranno essere presi in considerazione. Un'organizzazione che decidesse di impiegare una nuova tecnologia AI dovrebbe chiedersi:

- Qual è l'uso previsto del prodotto di AI?
    - Quanto è rappresentativo il dataset di addestramento rispetto al contesto operativo?
    - Come è stato addestrato il modello?
    - Il prodotto è adatto al caso d'uso?
    - In che modo le caratteristiche del prodotto si allineano al grado di responsabilità del mio caso d'uso e del mio contesto?
    - Quali sono i limiti della sua funzionalità?
- Qual è il processo per controllare e verificare le prestazioni del prodotto di AI?

- Quali sono le metriche delle prestazioni del prodotto?
- In che modo gli utenti possono interpretare l'output del prodotto?
- In che modo il prodotto è continuamente monitorato per guasti e altre condizioni di rischio?
- Quali pregiudizi impliciti sono incorporati nella tecnologia?
- Come vengono valutati gli aspetti dell'affidabilità? Con quale frequenza?
- C'è un modo per far re-addestrare questo strumento da un esperto per implementare politiche di equità?
- Sarò in grado di comprendere e controllare l'output dello strumento?
- Quali sono i controlli di sicurezza per evitare che questo sistema causi danni? Come possono essere testati questi controlli?

Generalmente, gli utenti finali diventano i primi osservatori dei fallimenti della tecnologia AI e le loro esperienze negative sono indicatori del rischio di deterioramento dell'affidabilità. Pertanto, le organizzazioni che decidessero di implementare e/o utilizzare tali sistemi e, contestualmente, renderli affidabili, dovrebbero supportare gli utenti finali sui seguenti aspetti:

- Inserire degli indicatori all'interno del sistema per segnalare quando non funziona come previsto,
- Effettuare delle valutazioni sulle prestazioni del sistema nei contesti attuali e nuovi,
- Prevedere un alert del sistema quando non funziona dentro un range di affidabilità accettabile,
- Acquisire informazioni utili ad allineare le loro aspettative ed esigenze con il potenziale rischio introdotto dal sistema.

Le risposte alle domande introdotte all'inizio di questo paragrafo mirano, da un lato, a chiarire se la tecnologia sia adatta allo scopo previsto e, dall'altro, a verificare se l'utente possa convalidare l'affidabilità su base continuativa. Le organizzazioni possono anche implementare funzionalità tecniche e modelli di governance per incentivare il mantenimento continuo dell'affidabilità dell'AI e fornire piattaforme per testare, valutare e gestire i prodotti AI.

Raccomandazioni

I concetti illustrati rappresentano una base di partenza per la creazione di un AI affidabile. Occorrerà, ovviamente, condurre ulteriori attività di ricerca e sviluppo per indagare su nuovi metodi, best practices e linee guida per la creazione di un'AI affidabile. Ecco alcuni elementi su cui i ricercatori dell'AI stanno lavorando per la misurazione dell'affidabilità:

- **Fairness** (imparzialità): l'identificazione e la mitigazione dei pregiudizi nei modelli di machine learning (ML) che consenta la creazione di sistemi più equi.
- **Robustness** (robustezza): se i sistemi di intelligenza artificiale non fossero sufficientemente resistenti ai guasti o agli errori, fallirebbero in breve tempo.
- **Explainability** (dimostrabilità): rappresenta un attributo rilevante per un sistema che vuole essere affidabile per tutti gli stakeholders: ingegneri e sviluppatori, utenti finali.

La produzione di sistemi affidabili basati sull'intelligenza artificiale aumenterà l'impatto di questi sistemi in ambito lavorativo e nel supportare altre finalità. Pertanto, è fondamentale incorporare gli obiettivi indicati sin dalla fase di pianificazione iniziale fino al rilascio finale e alla manutenzione del prodotto. Solo in questo modo le aziende potranno raggiungere il massimo potenziale delle aspettative previste dall'AI.

Nuovi modelli di Artificial Intelligence

L'intelligenza artificiale generativa sta permeando sempre più aspetti della nostra vita, plasmando il modo in cui interagiamo con il mondo e prendiamo decisioni. Dagli assistenti virtuali ai sistemi di supporto, dai veicoli a guida autonoma alle diagnosi mediche assistite, l'AI sta rivoluzionando interi settori e promettendo un futuro ricco di innovazione.

Tuttavia, questa diffusa adozione è accompagnata da una crescente preoccupazione connessa alla mancanza di trasparenza nel processo decisionale di molti sistemi di AI. Spesso definiti "scatole nere", questi sistemi producono risultati accurati, ma il percorso logico che li genera rimane oscuro e inaccessibile alla comprensione umana. Questa opacità solleva interrogativi cruciali sulla fiducia, la responsabilità e l'etica nell'utilizzo dell'AI.

In questo contesto, emerge l'importanza di modelli di Intelligenza Artificiale spiegabili e ibridi, un campo di ricerca in rapida evoluzione che mira a rendere i sistemi di AI più trasparenti e comprensibili. L'Explainable AI si propone di fornire spiegazioni chiare e interpretabili sul funzionamento interno di un modello di AI, consentendo agli utenti di comprendere come e perché un determinato risultato è stato raggiunto, mentre l'Hybrid AI prende il meglio dei diversi modelli di AI e li mette a fattore comune.

Opacità dell'AI

L'intelligenza artificiale che abbiamo conosciuto finora soffre di un grosso problema, l'opacità. Cosa significa opacità dell'AI? In ambito scientifico, l'opacità dell'AI si riferisce all'incapacità di comprendere il processo decisionale interno di un sistema di intelligenza artificiale, specialmente quelli basati su algoritmi complessi come il deep learning.

Ciò può comportare le seguenti conseguenze:

- **Mancanza di fiducia**: l'opacità può minare la fiducia dei progettisti e degli utilizzatori nell'accettazione dei risultati prodotti dall'AI, soprattutto in ambiti critici come la medicina, la fisica, la chimica, l'economia,

dove la comprensione del processo decisionale è fondamentale per validare le scoperte,

- **Difficoltà di interpretazione**: l'opacità rende difficile l'interpretazione dei risultati e le previsioni generate dall'AI, limitando la capacità dei ricercatori ad estrarre dai dati nuove conoscenze e intuizioni,
- **Rischio di bias e discriminazione**: i modelli di AI opachi possono perpetuare bias presenti nei dati di training, portando a risultati discriminatori o ingiusti,
- **Problemi di debugging e miglioramento**: l'opacità rende difficile identificare e correggere errori o bias nei modelli di AI, ostacolando il loro miglioramento e l'affidabilità,
- **Ostacolo alla riproducibilità**: la mancanza di trasparenza nel processo decisionale dell'AI può rendere faticosa la riproduzione dei risultati e la validazione delle scoperte scientifiche.

L'opacità rappresenta una sfida importante anche per il metodo scientifico, che notoriamente si basa sulla trasparenza, la riproducibilità e la comprensione dei fenomeni. L'utilizzo di modelli opachi può compromettere la validità delle scoperte scientifiche e la fiducia nella ricerca. Per affrontare questa criticità, la ricerca si è concentrate su queste tematiche:

- **Sviluppo di modelli di AI più interpretabili**: sono state sviluppate tecniche di modellazione, come l'Hybrid AI e l'Explainable AI, che mirano a creare sistemi più trasparenti e comprensibili,
- **Analisi e visualizzazione dei dati**: sono stati progettati strumenti per visualizzare e analizzare in dettaglio il comportamento dei modelli di AI, aiutando i ricercatori a comprendere il processo decisionale (debugging),
- **Validazione e verifica dei modelli di AI**: sono statti introdotti metodi per valutare l'accuratezza, l'affidabilità e l'equità dei modelli di AI, garantendo la validità dei risultati scientifici raggiunti.

In sintesi, l'opacità dell'AI rappresenta una sfida significativa per la ricerca. Superare questa opacità è fondamentale per garantire la fiducia, la trasparenza e la validità delle scoperte scientifiche nell'era dell'AI.

Explainable AI: trasparenza e affidabilità

L'intelligenza artificiale generativa (AI generativa) sta rapidamente trasformando il modo in cui creiamo e interagiamo con i contenuti digitali. Dai modelli linguistici di grandi dimensioni che generano testo realistico

alle reti generative avversarie (GAN) che creano immagini e video sorprendenti, l'AI generativa sta aprendo nuove frontiere in diversi campi, dall'arte e l'intrattenimento alla scienza e alla medicina.

Tuttavia, il potere dell'AI generativa è accompagnato da una crescente preoccupazione per la sua "scatola nera". I modelli generativi spesso operano in modo opaco, rendendo difficile comprendere come arrivino a generare determinati output. Questa mancanza di trasparenza solleva diverse questioni cruciali:

- **Fiducia:** come possiamo fidarci dei risultati generati da un sistema di AI che non comprendiamo appieno?
- **Responsabilità:** chi è responsabile per gli output generati dall'AI, soprattutto se sono dannosi o imprecisi?
- **Controllo:** come possiamo controllare e guidare il processo creativo dell'AI generativa se non sappiamo come funziona?
- **Bias:** come possiamo identificare e mitigare i bias presenti nei modelli generativi se non siamo in grado di analizzarne il processo decisionale?

L'Explainable AI (XAI) si pone come un campo di ricerca cruciale nell'ambito dell'intelligenza artificiale, focalizzandosi sulla trasparenza e l'interpretabilità dei sistemi di AI. Mentre l'AI tradizionale mira principalmente a massimizzare le prestazioni predittive, l'XAI si concentra sulla comprensione del processo decisionale dei modelli, rendendoli accessibili e interpretabili dagli esseri umani.

Sebbene non esista una definizione univoca, l'XAI può essere concettualizzata come un insieme di principi, metodi e tecniche che mirano a:

- **Spiegare le decisioni:** XAI può fornire spiegazioni sui processi decisionali dei modelli generativi, aiutandoci a capire perché un modello ha generato un determinato output,
- **Identificare i bias:** XAI può aiutare a identificare i bias presenti nei dati di addestramento o nel modello stesso, consentendo di mitigare i loro effetti negativi,
- **Controllare il processo creativo:** XAI può fornire strumenti per controllare e guidare il processo creativo dell'AI generativa, consentendo agli utenti di influenzare gli output generati,
- **Aumentare la fiducia:** Comprendere il funzionamento dei modelli generativi aumenta la fiducia negli

output generati, promuovendone l'adozione in diversi ambiti.

Pertanto, l'Explainable AI è in grado di colmare una serie di criticità:

- **Superare l'opacità delle "black box"**: molti modelli di AI, specialmente quelli basati sul deep learning, operano come "scatole nere", rendendo difficile la comprensione del loro funzionamento interno. L'XAI mira ad "aprire" queste scatole nere, fornendo spiegazioni chiare e comprensibili sul processo decisionale,
- **Aumentare la fiducia e l'accettazione:** la trasparenza è fondamentale per costruire fiducia nei sistemi di AI e favorirne l'adozione in ambiti scientifici critici, dove la comprensione del processo decisionale è essenziale per validare le scoperte,
- **Garantire la responsabilità etica**: l'XAI consente di identificare e mitigare potenziali bias e discriminazioni nei modelli di AI, promuovendo un utilizzo responsabile ed etico nella ricerca,
- **Facilitare il miglioramento e il debugging**: comprendere il processo decisionale dei modelli di AI facilita il loro miglioramento, la correzione di errori e l'ottimizzazione delle prestazioni, elementi cruciali per la ricerca,
- **Promuovere la scoperta scientifica**: l'XAI può aiutare i ricercatori a estrarre nuove conoscenze e intuizioni dai dati, aprendo nuove prospettive di ricerca e accelerando il progresso scientifico.

Per raggiungere questi obiettivi, è necessario che l'XAI si avvalga di un set di tecniche tra cui:

- **Modelli intrinsecamente interpretabili**: è fondamentale sviluppare modelli per la progettazione più semplici e trasparenti, come gli alberi decisionali, i modelli lineari e le regole di associazione,
- **Spiegazioni post-hoc**: occorre effettuare l'analisi del comportamento di un modello di AI dopo il suo addestramento, utilizzando tecniche come:
  - **Analisi di sensitività**: in grado di valutare l'influenza di ciascuna variabile di input sul risultato finale,
  - **Visualizzazione dei dati**: occorre utilizzare grafici e mappe per visualizzare il processo decisionale,
  - **LIME (Local Interpretable Model-Agnostic Explanations)**: è opportuno approssimare localmente il comportamento di un modello complesso con un modello più semplice,
  - **SHAP (SHapley Additive exPlanations)**: è possibile spiegare l'output di un modello in termini di con-

tributo di ciascuna variabile di input,

- **Metodi di attenzione**: è necessario identificare le parti più importanti dei dati di input che influenzano la decisione del modello, fornendo informazioni su quali aspetti dei dati sono stati considerati più rilevanti,
- **Spiegazioni controfattuali**: è utile fornire informazioni su come modificare gli input per ottenere un output desiderato, aiutando a comprendere il processo decisionale del modello e a identificare potenziali interventi.

Nonostante il campo di ricerca avente ad oggetto l'XAI stia progredendo significativamente, occorre affrontare ulteriori sfide:

- **Definire il concetto di "spiegabilità"**, ovvero stabilire i criteri oggettivi e condivisi per valutare la qualità e l'efficacia di una spiegazione, adattandoli ai diversi contesti scientifici,
- **Estendere la spiegabilità per diversi utenti**, ovvero adattare le spiegazioni alle esigenze e alle competenze di diversi tipi di utenti, dai ricercatori agli esperti di dominio,
- **Ricercare nuove tecniche di XAI** per continuare a sviluppare nuove tecniche per spiegare modelli di AI sempre più complessi, garantendo la loro applicabilità in diversi ambiti scientifici,
- **Integrare l'XAI nel ciclo di vita dell'AI** per incorporare l'XAI nella progettazione, lo sviluppo e l'implementazione dei sistemi di AI, promuovendo la trasparenza e la responsabilità fin dalle prime fasi della ricerca.

Ecco alcuni esempi di XAI per l'AI generativa:

- **Visualizzazione delle attivazioni:** la visualizzazione delle attivazioni dei neuroni in una rete neurale può aiutare a capire quali parti del modello sono responsabili per la generazione di specifici output,
- **Analisi di sensibilità:** l'analisi di come le variazioni negli input influenzano gli output generati può fornire informazioni sul funzionamento del modello,
- **Generazione di spiegazioni testuali:** lo sviluppo di modelli in grado di generare spiegazioni testuali per i loro output, rendendoli più comprensibili agli utenti.

XAI è un campo in rapida evoluzione con il potenziale per rivoluzionare il modo in cui interagiamo con l'AI generativa. Applicando XAI, possiamo rendere i modelli generativi più trasparenti, responsabili e affidabili, aprendo la strada a un utilizzo più etico e sicuro di questa potente tecnologia. L'integrazione di XAI nell'AI

generativa non solo favorirà la fiducia e l'accettazione di questa tecnologia, ma stimolerà anche nuove forme di creatività e innovazione.

Hybrid AI: un catalizzatore per l'AI Generativa

L'intelligenza artificiale ibrida (Hybrid AI), che combina la potenza delle reti neurali con la capacità di ragionamento simbolico, si presenta come un potente catalizzatore per l'evoluzione dell'AI generativa. Questo connubio apre nuove frontiere per la creatività, l'intelligenza e l'affidabilità dei sistemi di AI, superando i limiti dei singoli approcci.

I modelli generativi, pur producendo risultati sorprendenti, spesso mancano di trasparenza nel processo decisionale (vd. paragrafo precedente). L'Hybrid AI, integrando il ragionamento simbolico, può fornire spiegazioni più chiare e comprensibili sul perché e come un modello generativo produce un determinato output. Questo aumenta la fiducia negli utenti e consente un controllo più preciso sul processo creativo.

L'Hybrid AI si basa su questi principi:

- **Complementarietà delle tecniche**: è notorio che vi siano diverse tecniche di modellazione che eccellono in compiti diversi. Ad esempio, l'apprendimento automatico eccelle nell'individuare pattern e fare previsioni basate su grandi quantità di dati, ma può essere opaco e difficile da interpretare. Il ragionamento simbolico, d'altra parte, offre trasparenza e capacità di ragionamento logico, ma può avere difficoltà a gestire l'incertezza e la complessità del mondo reale. L'Hybrid AI mira a sfruttare il meglio di entrambi i mondi, combinando la capacità di apprendimento automatico di estrarre informazioni dai dati con la capacità del ragionamento simbolico di elaborare conoscenze e regole,
- **Sinergia e integrazione**: l'Hybrid AI non si limita a un semplice giustapposizione di tecniche, ma mira a una loro integrazione profonda, in cui le diverse componenti interagiscono e si completano a vicenda, creando un sistema olistico con capacità superiori alla somma delle sue parti,
- **Modellazione della cognizione umana**: l'Hybrid AI si ispira spesso alla cognizione umana, che combina intuizione, ragionamento e apprendimento. Integrando diverse tecniche di AI, si cerca di emulare la flessibilità e l'adattabilità dell'intelligenza umana.

Esistono diverse modalità per combinare differenti tecniche di modellazione in un sistema ibrido, i principali

metodi di approccio possono essere individuati nei seguenti:

- **Integrazione dei modelli ad apprendimento automatico con sistemi basati sulle regole**: l'apprendimento automatico può essere utilizzato per estrarre le regole dai dati, che poi vengono integrate in un sistema basato sul ragionamento simbolico, consentendo al sistema di apprendere da grandi quantità di dati e di utilizzare la conoscenza simbolica per ragionare e prendere decisioni,
- **Combinazione delle reti neurali con algoritmi evolutivi**: gli algoritmi evolutivi possono essere utilizzati per ottimizzare la struttura e i parametri delle reti neurali, migliorando la loro efficienza e la loro capacità di generalizzazione,
- **Utilizzo di tecniche di fuzzy logic per gestire l'incertezza**: la fuzzy logic consente di rappresentare e di effettuare ragionamenti su concetti vaghi e incerti, migliorando la capacità dei sistemi di AI nella gestione della complessità del mondo reale e di prendere decisioni in situazioni ambigue,
- **Integrazione dei modelli di deep learning con i metodi di ragionamento basati sulla logica**: combinare la potenza della rappresentazione del deep learning con la capacità di ragionamento e l'inferenza della logica, consentendo ai sistemi di AI di apprendere da grandi quantità di dati e di utilizzare la logica per ragionare e spiegare le proprie decisioni.

Tutto ciò può portare dei vantaggi significativi in termini di:

- **Robustezza e affidabilità**: i sistemi ibridi sono meno suscettibili a errori e bias rispetto ai sistemi basati su un singolo approccio, in quanto le diverse componenti possono compensare le debolezze reciproche,
- **Flessibilità e adattabilità**: i sistemi ibridi possono adattarsi più facilmente a nuovi contesti e compiti, grazie alla loro capacità di combinare diverse strategie di apprendimento e di ragionamento,
- **Interpretabilità e trasparenza**: l'Hybrid AI può favorire lo sviluppo di sistemi di AI più trasparenti e comprensibili, in quanto l'integrazione di tecniche come il ragionamento simbolico può rendere il processo decisionale più esplicito,
- **Efficienza**: l'Hybrid AI può ottimizzare l'utilizzo delle risorse e migliorare le prestazioni dei sistemi di AI, combinando le diverse tecniche in modo sinergico.

Nonostante il suo potenziale, l'Hybrid AI deve affrontare diverse sfide:

- **Sviluppo di framework e strumenti specifici**: sono necessari nuovi strumenti e framework per facilitare l'integrazione di diverse tecniche di AI in modo efficiente e scalabile,
- **Gestione della complessità**: i sistemi ibridi possono essere più complessi da progettare, implementare e gestire rispetto ai sistemi basati su un singolo approccio. Pertanto, sono necessarie nuove metodologie e tecniche per affrontare la complessità dei sistemi ibridi,
- **Valutazione e benchmarking**: è necessario sviluppare nuove metriche e metodi di valutazione per confrontare e valutare le prestazioni dei sistemi di AI ibrida in modo oggettivo e rigoroso.

L'Hybrid AI, combinando l'AI generativa con l'AI simbolica, offre diverse opportunità:

- **Spiegabilità:** l'AI simbolica, basata su regole e logica, introduce trasparenza nel processo generativo. Le regole possono essere utilizzate per spiegare le decisioni prese dal sistema e per giustificare gli output generati,
- **Ragionamento e Generalizzazione:** l'AI simbolica consente di integrare conoscenza pregressa e di ragionare su di essa, migliorando la capacità di generalizzazione dell'AI generativa e consentendole di affrontare situazioni nuove in modo più efficace,
- **Controllo e Bias Mitigation:** l'AI simbolica può essere utilizzata per definire vincoli e regole che guidano il processo generativo, limitando i bias e garantendo che gli output siano conformi a determinati criteri o standard etici.

Ecco alcune applicazioni Concrete dell'Hybrid AI per l'AI Generativa:

- **Generazione di Contenuti Controllata:** nell'ambito della scrittura creativa, l'Hybrid AI può garantire che il testo generato rispetti determinate regole grammaticali, stilistiche o di contenuto, aumentando la coerenza e la qualità del testo,
- **Creazione di Arte e Musica:** l'Hybrid AI può combinare la creatività dell'AI generativa con la struttura e l'armonia fornite dall'AI simbolica, consentendo la creazione di opere d'arte e musica più sofisticate e innovative,
- **Design e Ingegneria:** l'Hybrid AI può generare progetti che soddisfano specifici requisiti funzionali ed estetici, combinando l'esplorazione creativa con la validazione basata su regole e vincoli,

- **Scoperta Scientifica:** l'Hybrid AI può accelerare la scoperta di nuovi materiali, farmaci o soluzioni ingegneristiche, generando ipotesi e modelli che vengono poi validati e raffinati attraverso il ragionamento simbolico.

L'Hybrid AI promette di rivoluzionare l'AI generativa, aprendo nuove frontiere per la creatività, l'innovazione e la risoluzione di problemi complessi. La combinazione di intuizione e ragione, di apprendimento e logica, rappresenta un passo fondamentale verso lo sviluppo di un'AI più potente, versatile e affidabile.

Uno sguardo al futuro

Gli sviluppi dell'intelligenza artificiale sono stati così significativi in termini sia di dimensione, che di velocità, che non c'è stato sempre né il tempo, né l'opportunità, di comprendere pienamente la loro sicurezza e la loro effettiva rilevanza.

Di seguito sono esposite alcune branche dell'Intelligenza Artificiale, in evoluzione e in grado di rispondere alle nuove richieste sempre più incombenti, e il loro apporto al miglioramento dell'Intelligenza Artificiale Generativa.

Advanced Machine Learning

L'Intelligenza Artificiale Generativa (AI Generativa) sta vivendo una fase di rapida evoluzione alimentata anche dai progressi nell'Advanced Machine Learning (ML avanzato). Questo connubio apre nuove frontiere e opportunità, consentendo la creazione di modelli generativi sempre più sofisticati e capaci. In tal senso, l'apprendimento automatico (*Machine Learning*) ha già rivoluzionato il campo dell'intelligenza artificiale, consentendo ai computer di apprendere dai dati senza essere esplicitamente programmati. In tal senso, ha permesso di ottenere risultati straordinari in diversi ambiti, come il riconoscimento di immagini, l'elaborazione del linguaggio naturale e la previsione di serie temporali. Tuttavia, l'apprendimento automatico tradizionale presenta alcune limitazioni che ne ostacolano il pieno potenziale:

- **Dipendenza dai dati etichettati**: molti algoritmi di apprendimento automatico richiedono grandi quantità di dati etichettati, la cui acquisizione può essere costosa e laboriosa,
- **Generalizzazione limitata**: i modelli di apprendimento automatico possono avere difficoltà a generalizza-

re nuovi dati o contesti diversi da quelli su cui sono stati addestrati,

- **Mancanza di interpretabilità**: spesso è difficile comprendere il processo decisionale dei modelli di apprendimento automatico, il che può limitarne l'affidabilità e l'accettazione.

L'Apprendimento Automatico Avanzato (*Advanced Machine Learning*) si propone di superare queste limitazioni, esplorando nuove frontiere e aprendo nuove possibilità per la ricerca e l'innovazione tecnologica. In particolare, può affrontare alcune sfide che limitano l'AI generativa, come:

- **Qualità e diversità:** è possibile migliorare la qualità e la diversità dei contenuti generati, riducendo artefatti e ripetizioni,
- **Controllo e personalizzazione:** è in grado di aumentare il controllo sui contenuti generati, consentendo la personalizzazione in base alle esigenze degli utenti,
- **Generalizzazione e creatività:** può sviluppare modelli in grado di generalizzare a nuovi contesti e di esprimere maggiore creatività,
- **Spiegabilità e affidabilità:** può rendere i modelli generativi più trasparenti e affidabili, fornendo spiegazioni sul loro funzionamento.

Per superare queste limitazioni, l'apprendimento automatico avanzato esplora nuove frontiere:

- **Apprendimento auto-supervisionato (SSL)**: questa tecnica consente ai modelli di apprendere da soli da grandi quantità di dati non strutturati, riducendo la dipendenza da dati etichettati. L'obiettivo è quello di estrarre automaticamente informazioni significative dai dati, scoprendo pattern e relazioni nascoste. SSL può migliorare l'efficienza dell'addestramento e la generalizzazione dei modelli generativi. Alcuni esempi di apprendimento auto-supervisionato includono:
  - **Pre-addestramento di modelli linguistici**: modelli come BERT e GPT-3 sono pre-addestrati su enormi quantità di testo, imparando a prevedere parole mancanti o a generare testo coerente,
  - **Apprendimento di rappresentazioni visive**: modelli come SimCLR e MoCo imparano a generare rappresentazioni visive significative da immagini non etichettate,
- **Apprendimento federato**: questa tecnica consente di addestrare modelli di AI su dati distribuiti su diversi dispositivi, preservando la privacy e la sicurezza dei dati. L'apprendimento federato è particolarmente uti-

le in ambiti come la medicina e la finanza, dove i dati sono spesso sensibili e decentralizzati. Un esempio di applicazione è l'addestramento di modelli di diagnosi medica su dati provenienti da diversi ospedali, senza la necessità di condividere i dati dei pazienti (vd. Apposito paragrafo),

- **Meta-apprendimento**: questa tecnica si concentra sullo sviluppo di modelli di AI in grado di "apprendere come apprendere", adattandosi rapidamente a nuovi compiti e contesti. L'obiettivo è quello di creare sistemi di AI più flessibili e versatili, in grado di generalizzare a nuovi problemi e di apprendere in modo continuo. Un esempio di meta-apprendimento è l'algoritmo MAML (Model-Agnostic Meta-Learning), che apprende un insieme di parametri iniziali che consentono al modello di adattarsi rapidamente a nuovi compiti. Il meta-apprendimento può rendere i modelli generativi più flessibili e versatili,
- **Apprendimento per rinforzo (RL)**: questa tecnica consente ai modelli di apprendere interagendo con un ambiente e ricevendo feedback sotto forma di ricompense o penalità. L'apprendimento per rinforzo è particolarmente utile per lo sviluppo di agenti intelligenti, come robot autonomi e sistemi di controllo, In AI generativa, RL può essere utilizzato per ottimizzare la generazione di contenuti in base a obiettivi specifici,
- **Apprendimento profondo spiegabile**: questa area di ricerca si concentra sullo sviluppo di modelli di deep learning più interpretabili, consentendo di comprendere il processo decisionale e di aumentare la fiducia negli output del modello,
- **Modelli probabilistici profond (Deep Probabilistic Models)**: questa tecnica combina le reti neurali con modelli probabilistici per rappresentare l'incertezza e generare contenuti con maggiore varietà e controllo,
- **AI neuro-simbolica**: questo metodo integra le reti neurali con sistemi di ragionamento simbolico per creare modelli generativi più spiegabili e capaci di ragionamento astratto.

Nonostante i significativi progressi effettuati, l'AML deve ancora affrontare diverse sfide:

- **Scalabilità**: è fondamentale sviluppare algoritmi in grado di gestire dataset sempre più grandi e complessi,
- **Robustezza**: è necessario creare modelli di AI resistenti a errori e perturbazioni nei dati,
- **Etica e responsabilità**: è essenziale garantire che l'apprendimento automatico avanzato sia utilizzato in

modo responsabile ed etico, evitando bias e discriminazioni.

L'ML avanzato apre nuove opportunità per l'AI generativa in diversi ambiti:

- **Creazione di contenuti realistici**: generare immagini, video, audio e testo di alta qualità, indistinguibili da quelli reali,
- **Scoperta scientifica**: generare nuove molecole, materiali e prodotti con proprietà desiderate,
- **Arte e creatività**: assistere artisti e creativi nella generazione di nuove opere d'arte, musica e design,
- **Personalizzazione dell'esperienza utente**: creare esperienze personalizzate in base alle preferenze individuali degli utenti.

L'ML avanzato sta trasformando l'AI generativa, consentendo la creazione di modelli sempre più potenti e versatili. Questa sinergia apre nuove frontiere in diversi ambiti, ma richiede anche un approccio responsabile per garantire un uso etico e benefico di questa tecnologia.

Embodied AI: un corpo per l'AI Generativa

L'Intelligenza Artificiale Generativa (Generative AI) consente di generare testi, immagini e codice con sorprendente creatività. Tuttavia, questi modelli operano in un mondo astratto, privi di un corpo fisico che permetta loro di interagire direttamente con l'ambiente. È qui che entra in gioco l'Embodied AI, un campo di ricerca che esplora l'integrazione dell'AI in corpi robotici, aprendo nuove e affascinanti prospettive per l'AI generativa.

L'Intelligenza Artificiale Embodied (Embodied AI) rappresenta un cambio di paradigma nello studio dell'AI, spostando il focus dalla pura elaborazione di informazioni all'interazione fisica con il mondo. Invece di limitarsi ad elaborare informazioni in modo digitale, l'Embodied AI si concentra sull'integrazione dell'intelligenza in corpi fisici, come robot o agenti virtuali, che interagiscono con l'ambiente circostante.[96].

Tradizionalmente, l'AI si è concentrata sullo sviluppo di algoritmi e modelli in grado di elaborare informazioni e risolvere problemi in modo astratto, senza una connessione diretta con il mondo fisico. L'Embodied AI, al contrario, enfatizza l'importanza dell'incarnazione dell'intelligenza in corpi fisici. Questo approccio si basa

96 *Pfeifer, R., & Bongard, J. (2006). How the body shapes the way we think: A new view of intelligence. MIT press.*

sull'idea che la cognizione sia strettamente legata all'esperienza fisica e all'interazione sensomotoria.[97]

L'Embodied AI si fonda su una serie di teorie e principi:

- **Cognizione incorporata**: l'Embodied AI si basa sulla teoria della cognizione incorporata, che sostiene che la cognizione è strettamente legata all'esperienza fisica e all'interazione con l'ambiente. Secondo questa teoria, il corpo non è un semplice strumento dell'intelligenza, ma un elemento costitutivo della cognizione stessa,
- **Interazione sensomotoria**: l'Embodied AI enfatizza l'importanza dell'interazione sensomotoria, ovvero la capacità di percepire l'ambiente attraverso i sensori e di agire su di esso attraverso gli attuatori. Questa interazione dinamica con il mondo consente all'Embodied AI di apprendere dall'esperienza e di adattarsi a nuove situazioni,
- **Apprendimento situato**: l'Embodied AI si concentra sull'apprendimento situato, ovvero l'apprendimento che avviene in un contesto specifico e che è guidato dall'interazione con l'ambiente. Questo tipo di apprendimento consente all'AI Embodied di acquisire conoscenze e competenze rilevanti per il suo ambiente e i suoi obiettivi.

Vediamo quali possono essere le implicazioni dell'Embodied AI:

- **Robotica cognitiva**: la robotica cognitiva si concentra sullo sviluppo di robot autonomi in grado di interagire con l'ambiente, apprendere dall'esperienza e risolvere problemi in modo flessibile e adattabile,[98]
- **Agenti virtuali embodied**: gli agenti virtuali embodied sono programmi che simulano il comportamento di agenti intelligenti in ambienti virtuali. Questi agenti possono essere utilizzati per studiare l'intelligenza artificiale, la cognizione umana e l'interazione sociale,
- **Interfacce cervello-computer**: le interfacce cervello-computer consentono di collegare il cervello umano a dispositivi esterni, come protesi robotiche o computer. Questa tecnologia può essere utilizzata per sviluppare nuove forme di Embodied AI che integrano l'intelligenza umana con quella artificiale.

---

97 Vernon, D., Metta, G., & Sandini, G. (2007). A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents. IEEE Transactions on Evolutionary Computation, 11(2), 151-180.

98 Ziemke, T. (2001). Are robots embodied?. In Embodied artificial intelligence (pp. 75-96). Springer, Berlin, Heidelberg.

I principali vantaggi connessi allo sviluppo dell'Embodied AI possono essere sintetizzati in:

- **Comprensione della cognizione**: l'Embodied AI può fornire nuove intuizioni sul funzionamento della cognizione umana, studiando come l'intelligenza emerge dall'interazione tra corpo, cervello e ambiente,
- **Sviluppo di robot autonomi**: l'Embodied AI può consentire lo sviluppo di robot autonomi più intelligenti e adattabili, in grado di operare in ambienti complessi e di collaborare con gli esseri umani,
- **Nuove interfacce uomo-macchina**: l'Embodied AI può portare allo sviluppo di nuove interfacce uomo-macchina più intuitive e naturali, che consentono una migliore interazione tra esseri umani e computer,
- **Modellazione di sistemi complessi**: l'Embodied AI può essere utilizzata per modellare e simulare sistemi complessi, come ecosistemi, società e organismi biologici, fornendo nuove prospettive di ricerca.

I modelli di Generative AI attuali, pur sofisticati, mancano di una comprensione profonda del mondo fisico. L'Embodied AI, invece, permette all'AI di "vivere" il mondo attraverso un corpo robotico, acquisendo informazioni sensoriali e interagendo con l'ambiente. Questa esperienza incarnata può arricchire l'AI generativa in diversi modi:

- **Comprensione contestuale:** l'interazione con l'ambiente fornisce un contesto cruciale per la generazione di contenuti. Un robot che "vede" una mela può generare descrizioni più accurate e creative rispetto a un modello che ha appreso solo da immagini,
- **Apprendimento grounded:** l'Embodied AI permette di "radicare" la conoscenza in esperienze concrete, creando rappresentazioni più robuste e significative,
- **Creatività situata:** la creatività non è un processo astratto, ma è influenzata dal contesto e dalle interazioni. Un robot che "vive" in un ambiente può sviluppare una creatività più ricca e sfaccettata.

L'Embodied AI può spingere l'AI generativa oltre la semplice produzione di contenuti digitali, consentendo la creazione di opere fisiche e l'interazione con il mondo reale:

- **Arte robotica:** Robot dotati di AI generativa potrebbero dipingere, scolpire o comporre musica, esprimendo la loro creatività in forme nuove,
- **Design generativo:** L'Embodied AI può guidare la progettazione e la fabbricazione di oggetti fisici, adat-

tandoli alle esigenze dell'ambiente e degli utenti,

- **Interazione uomo-robot:** Robot con capacità generative potrebbero comunicare e collaborare con gli esseri umani in modi più naturali e intuitivi.

L'integrazione di Embodied AI e Generative AI presenta sfide importanti:

- **Complessità:** sviluppare robot con capacità sensoriali, motorie e cognitive avanzate è un compito complesso,
- **Generalizzazione:** i modelli generativi devono essere in grado di generalizzare a diversi ambienti e situazioni,
- **Etica:** è fondamentale garantire che l'Embodied AI sia sviluppata e utilizzata in modo responsabile ed etico.

Nonostante le sfide, l'Embodied AI offre un potenziale enorme per l'AI generativa. In futuro, potremmo vedere robot che non solo generano contenuti creativi, ma che interagiscono con il mondo in modi significativi, contribuendo a risolvere problemi e a migliorare la nostra vita.

In conclusione, l'Embodied AI rappresenta una nuova frontiera per l'AI generativa, offrendo la possibilità di creare sistemi di AI più intelligenti, creativi e integrati con il mondo fisico.

Generalized Artificial Intelligence

L'attuale generazione di modelli di AI generativa presenta delle limitazioni, tra cui la difficoltà nel comprendere il contesto, la mancanza di senso comune e la tendenza a generare output incoerenti o irrealistici. L'avvento dell'Intelligenza Artificiale Generale (AGI), ovvero un'AI con capacità cognitive simili a quelle umane, potrebbe rivoluzionare l'AI generativa, aprendo nuove frontiere di creatività e innovazione.

L'Intelligenza Artificiale Generalizzata (AGI), spesso definita anche come "IA forte" o "IA completa", rappresenta un obiettivo ambizioso nel campo dell'intelligenza artificiale. A differenza dell'AI ristretta, che eccelle in compiti specifici, l'AGI mira a replicare l'intelligenza umana nella sua totalità, creando macchine in grado di apprendere, ragionare, risolvere problemi e interagire con il mondo in modo flessibile e adattabile, proprio

come gli esseri umani.[99] [100]

L'AGI si distingue dall'AI ristretta per alcune caratteristiche fondamentali:

- **Generalità**: la capacità di apprendere e svolgere un'ampia gamma di compiti, anziché essere limitata a un dominio specifico,
- **Adattabilità**: la possibilità di trasferire conoscenze e competenze da un dominio all'altro e di adattarsi a nuove situazioni e ambienti,
- **Astrazione**: l'opportunità di ragionare a livelli di astrazione elevati, di formare concetti e di generalizzare da esempi specifici,
- **Coscienza e autoconsapevolezza**: sebbene ancora oggetto di dibattito, alcuni ricercatori ritengono che l'AGI possa sviluppare una forma di coscienza e autoconsapevolezza, simile a quella umana.

Al contempo, lo sviluppo concreto di questa modellazione presuppone la risoluzione di una serie di criticità:

- **Complessità del cervello umano**: il cervello umano è un sistema incredibilmente complesso, con miliardi di neuroni interconnessi. Replicarne il funzionamento richiede una comprensione profonda dei meccanismi neurali e cognitivi,
- **Mancanza di una teoria unificata dell'intelligenza**: ancora non esiste una teoria completa e condivisa su cosa sia l'intelligenza e come funzioni,
- **Rappresentazione della conoscenza**: la rappresentazione della conoscenza del mondo, in modo che possa essere utilizzata da un sistema di AGI per ragionare e risolvere problemi,
- **Apprendimento continuo**: come sviluppare sistemi di AGI in grado di apprendere in modo continuo e autonomo, adattandosi a nuove informazioni e situazioni,
- **Motivazione e obiettivi**: come dotare i sistemi di AGI di motivazioni e obiettivi intrinseci, in modo che possano agire in modo autonomo e finalizzato.

---

99 Goertzel, B. (2014). Artificial General Intelligence: Concept, State of the Art, and Future Prospects. Journal of Artificial General Intelligence, 5(1), 1-48.

100 Marcus, G. (2018). Deep Learning: A Critical Appraisal. arXiv preprint arXiv:1801.00631.

Per affrontare queste sfide si utilizzano i seguenti approcci scientifici in parte esaminati:

- **AI Simbolica**: si basa sulla manipolazione di simboli e sulla logica formale per rappresentare la conoscenza e il ragionamento,
- **Reti neurali artificiali**: simulano il funzionamento del cervello umano attraverso reti di neuroni artificiali interconnessi,
- **Approcci evolutivi**: utilizzano algoritmi genetici per far evolvere sistemi di AI con capacità cognitive sempre più avanzate,
- **AI Ibrida**: combina diverse tecniche di AI per sfruttare i punti di forza di ciascun approccio,
- **Cognizione incorportata**: studia l'intelligenza artificiale in relazione all'interazione con un corpo fisico e un ambiente.

Ecco alcuni modi in cui l'AGI potrebbe potenziare l'AI generativa:

1. **Comprensione del Contesto e del Senso Comune:** l'AGI, con la sua capacità di comprendere il mondo in modo olistico e di applicare il senso comune, potrebbe consentire all'AI generativa di:
    - **Generare contenuti più coerenti e pertinenti al contesto:** potrebbe aiutare l'AI generativa a interpretare le sfumature del linguaggio, a comprendere le relazioni tra i concetti e a generare output che siano logicamente validi e coerenti con il contesto,
    - **Creare contenuti più originali e creativi:** potrebbe fornire all'AI generativa la capacità di combinare idee in modi nuovi e inaspettati, di esplorare diverse prospettive e di generare contenuti veramente innovativi,
    - **Adattare i contenuti a specifici pubblici e obiettivi:** potrebbe consentire all'AI generativa di comprendere le esigenze e le preferenze degli utenti, generando contenuti personalizzati e mirati.
2. **Apprendimento Continuo e Adattamento:** l'AGI, con la sua capacità di apprendere continuamente e di adattarsi a nuove situazioni, potrebbe consentire all'AI generativa di:
    - **Migliorare le proprie capacità nel tempo:** potrebbe aiutare l'AI generativa ad apprendere dai propri errori, a perfezionare le proprie tecniche e a generare contenuti sempre più sofisticati,

- **Generalizzare a nuovi domini e compiti:** potrebbe consentire all'AI generativa di applicare le proprie conoscenze e competenze a nuovi ambiti, generando contenuti in diverse forme e stili,
- **Collaborare con gli esseri umani:** potrebbe facilitare la collaborazione tra umani e AI generativa, consentendo agli utenti di guidare e influenzare il processo creativo.

3. **Ragionamento e Problem Solving:** L'AGI, con la sua capacità di ragionare, risolvere problemi e prendere decisioni, potrebbe consentire all'AI generativa di:
   - **Generare contenuti più complessi e articolati:** potrebbe aiutare l'AI generativa a creare storie, articoli, programmi e altri contenuti che richiedono una struttura logica e una pianificazione complessa,
   - **Risolvere problemi creativi:** potrebbe consentire all'AI generativa di superare blocchi creativi, esplorare soluzioni alternative e trovare nuove idee,
   - **Automatizzare compiti complessi:** potrebbe consentire all'AI generativa di automatizzare la creazione di contenuti in diversi ambiti, liberando tempo e risorse per gli esseri umani.

L'AGI ha il potenziale per rivoluzionare l'AI generativa, trasformandola da uno strumento per la creazione di contenuti semplici a un partner creativo in grado di collaborare con gli esseri umani per raggiungere nuovi livelli di innovazione. L'integrazione dell'AGI nell'AI generativa potrebbe portare a una nuova era di creatività, in cui le macchine e gli esseri umani lavorano insieme per esplorare le infinite possibilità dell'immaginazione.

Lo sviluppo dell'AGI solleva importanti questioni etiche:

- **Controllo e sicurezza**: come garantire che i sistemi di AGI siano sicuri e controllati, evitando comportamenti dannosi o imprevisti,
- **Impatto sociale**: come gestire l'impatto dell'AGI sul lavoro, l'economia e la società nel suo complesso,
- **Diritti e responsabilità**: quali diritti e responsabilità dovrebbero essere attribuiti ai sistemi di AGI.

In tal senso, la ricerca sull'AGI richiede un approccio multidisciplinare, che integri conoscenze provenienti da diversi campi, come l'informatica, la neuroscienza, la psicologia e la filosofia. È fondamentale affrontare le sfide scientifiche ed etiche associate all'AGI per garantire che questa tecnologia sia sviluppata e utilizzata in

modo responsabile e benefico per l'umanità.[101]

Artificial Intelligence Agent

Questo paragrafo esplora il concetto di *"agente intelligente"* nell'ambito dell'Intelligenza Artificiale generativa. In particolare, si analizzano le caratteristiche distintive, le architetture e le tipologie, evidenziando il loro ruolo cruciale nell'evoluzione dell'AI verso sistemi autonomi, adattabili e interattivi. Vengono inoltre discusse le sfide aperte e le prospettive future della ricerca sugli agenti intelligenti, con particolare attenzione alle implicazioni etiche e sociali.[102]

L'agente intelligente rappresenta un paradigma fondamentale nell'AI, incarnando l'idea di un sistema informatico in grado di percepire l'ambiente, elaborare le informazioni e agire autonomamente per raggiungere obiettivi predefiniti.[103] A differenza dei sistemi di AI passivi, gli agenti intelligenti interagiscono in modo proattivo e adattivo con il mondo circostante, aprendo nuove possibilità per la creazione di sistemi autonomi e intelligenti.

Un agente intelligente si distingue per le seguenti caratteristiche[104]:

- **Autonomia**: sono oggetti complessi che operano senza l'intervento umano diretto, prendendo decisioni in base alle proprie percezioni e conoscenze,
- **Percezione**: sono in grado di acquisire informazioni sull'ambiente tramite sensori (fisici o virtuali), che possono includere telecamere, microfoni, sensori di temperatura, o anche flussi di dati da internet,
- **Azione**: eseguono azioni che modificano l'ambiente, tramite attuatori come motori, altoparlanti, o la capacità di inviare comandi ad altri sistemi,
- **Obiettivo**: agiscono per raggiungere uno o più obiettivi specifici, che possono essere definiti a priori o appresi durante l'interazione con l'ambiente,

101 *Russell, S. J., & Norvig, P. (2016). Artificial intelligence: a modern approach. Pearson Education Limited.*
102 *Russell, S. J., & Norvig, P. (2010). Artificial intelligence: a modern approach. Prentice Hall*
103 *Wooldridge, M. (2009). An introduction to multiagent systems. John Wiley & Sons*
104 *Weiss, G. (Ed.). (2013). Multiagent systems: a modern approach to distributed artificial intelligence. MIT press.*

- **Apprendimento**: possono migliorare le proprie prestazioni nel tempo attraverso l'apprendimento, adattando il proprio comportamento in base alle esperienze passate.

Un agente intelligente è tipicamente composto dai seguenti componenti:

Sensori: oggetti in grado di acquisire le informazioni dall'ambiente (es. telecamere, microfoni, sensori di temperatura),

- **Attuatori**: oggetti destinati a eseguire azioni sull'ambiente (es. motori, altoparlanti, display),
- **Unità di elaborazione**: il componente dedicato all'elaborazione delle informazioni sensoriali e alla decisione sulle azioni da intraprendere,
- **Base di conoscenza**: il nucleo in grado di contenere le informazioni sul mondo e sugli obiettivi dell'agente.

In letteratura esistono diverse tipologie di agenti intelligenti, classificate in base alle loro capacità e al loro grado di sofisticazione:

- **Agenti reattivi semplici**: agenti che reagiscono direttamente agli stimoli ambientali senza memoria delle esperienze passate,
- **Agenti reattivi basati su modello**: agenti che mantengono un modello interno dell'ambiente e delle sue dinamiche,
- **Agenti basati su obiettivi**: agenti che agiscono per raggiungere obiettivi specifici, pianificando le azioni in base al modello dell'ambiente,
- **Agenti basati su utilità**: agenti che valutano le azioni in base a una funzione di utilità che misura il grado di soddisfazione degli obiettivi,
- **Agenti che apprendono**: agenti che migliorano le proprie prestazioni nel tempo attraverso l'apprendimento automatico.

Gli Agenti di AI, con la loro capacità di agire autonomamente e di interagire con l'ambiente, possono sbloccare nuove possibilità e applicazioni per l'AI generativa, aprendo la strada a una nuova era di creatività e innovazione.

Cosa può fare un Agente di AI per l'AI Generativa?

1. **Automatizzare il processo creativo**:
    - Gli agenti di AI possono automatizzare compiti ripetitivi e dispendiosi in termini di tempo, come la raccolta di dati, la pre-elaborazione e l'ottimizzazione dei parametri, consentendo ai creatori di concentrarsi sugli aspetti più creativi del processo,
    - Possono generare variazioni di un'opera, esplorare diverse opzioni creative e suggerire nuove idee, stimolando l'ispirazione e ampliando le possibilità artistiche,
2. **Personalizzare l'esperienza creativa**:
    - Gli agenti di AI possono apprendere le preferenze e lo stile di un creatore, adattando i modelli generativi alle sue esigenze specifiche,
    - Possono fornire feedback personalizzati e suggerimenti in tempo reale, guidando il processo creativo e migliorando la qualità dei risultati,
3. **Facilitare la collaborazione creativa:**
    - Gli agenti di AI possono fungere da intermediari tra diversi creatori, facilitando la comunicazione e la condivisione di idee,
    - Possono combinare i contributi di diversi artisti, generando opere collaborative innovative e stimolando nuove forme di espressione artistica,
4. **Espandere le applicazioni dell'AI generativa:**
    - Gli agenti di AI possono integrare l'AI generativa in nuovi contesti e applicazioni, come la creazione di ambienti virtuali interattivi, la progettazione di prodotti personalizzati e lo sviluppo di esperienze educative coinvolgenti,
    - Possono agire come "creativi artificiali", generando contenuti originali e innovativi in modo autonomo, aprendo nuove frontiere nell'arte e nell'intrattenimento,

Ecco alcuni esempi concreti di utilizzo degli agenti per la generazione di contenuti:

- Un agente di AI potrebbe assistere un musicista nella composizione di una canzone, suggerendo melo-

die, armonie e ritmi in base al suo stile e alle sue preferenze,

- Un agente di AI potrebbe collaborare con un designer per creare un abito su misura, generando diverse opzioni di design e adattandole alle misure e ai gusti del cliente,
- Un agente di AI potrebbe sviluppare un videogioco interattivo, generando ambienti, personaggi e storie in tempo reale, in base alle azioni del giocatore.

Lo stato dell'arte presuppone cha la ricerca debba ancora affrontare diverse sfide:

- **Scalabilità**: la capacità di sviluppare agenti in grado di operare in ambienti complessi e dinamici, con un elevato numero di variabili e interazioni,
- **Robustezza**: garantire la capacità degli agenti di gestire situazioni impreviste e incerte, adattando il proprio comportamento in modo flessibile,
- **Interazione uomo-agente**: progettare agenti in grado di interagire in modo naturale e intuitivo con gli esseri umani, comprendendo il linguaggio naturale e le emozioni,
- **Etica** e **sicurezza**: assicurare che gli agenti intelligenti siano utilizzati in modo responsabile ed etico, senza rischi per la sicurezza umana e nel rispetto dei valori sociali.

L'integrazione tra agenti di AI e AI generativa rappresenta un passo significativo verso lo sviluppo di sistemi più creativi, autonomi e collaborativi. Gli agenti possono amplificare il potenziale dell'AI generativa, automatizzando, personalizzando ed espandendo le sue applicazioni, aprendo la strada a una nuova era di innovazione e creatività.

AI Neuro-Simbolica

L'intelligenza artificiale ha compiuto progressi significativi negli ultimi anni, ma le due principali correnti di pensiero - il connessionismo e il simbolismo – che sono alla base delle moderne modellazioni di AI presentano limiti intrinseci. Le reti neurali, pur eccellendo nell'apprendimento di pattern da grandi quantità di dati, mancano di trasparenza e capacità di ragionamento simbolico. I sistemi di AI simbolica, d'altra parte, sono trasparenti e capaci di ragionamento logico, ma hanno difficoltà ad apprendere da grandi quantità di dati e a gestire l'incertezza.

L'AI neuro-simbolica si propone di superare questi limiti integrando i punti di forza di entrambi gli approcci. L'obiettivo è quello di creare sistemi di AI che siano in grado di:

- Apprendere da grandi quantità di dati,
- Ragionare in modo logico e simbolico,
- Spiegare il proprio processo decisionale,
- Generalizzare a nuove situazioni,
- Gestire l'incertezza.

L'AI neuro-simbolica si basa sull'idea che la cognizione umana combini aspetti sia connessionisti che simbolisti. Il cervello umano è un sistema complesso che elabora informazioni sia a livello sub-simbolico (reti neurali) che a livello simbolico (linguaggio e pensiero astratto). L'AI neuro-simbolica mira a emulare questa dualità, integrando reti neurali e sistemi di ragionamento simbolico in un'unica architettura.

Vediamo in dettaglio le motivazioni che fanno privilegiare l'AI Neuro-Simbolica:

- **Superare i limiti del connessionismo**: le reti neurali, pur dimostrando ottime performance nell'apprendimento di pattern da grandi quantità di dati, presentano dei limiti in termini di interpretabilità, ragionamento simbolico e capacità di generalizzazione,
- **Superare i limiti del simbolismo**: i sistemi di AI simbolica, pur eccellendo nel ragionamento logico e nella manipolazione di simboli, hanno difficoltà ad apprendere da grandi quantità di dati e a gestire l'incertezza,
- **Creare una AI più robusta e versatile**: l'AI neuro-simbolica mira a creare sistemi di AI più robusti, flessibili e adattabili, in grado di apprendere da grandi quantità di dati, di ragionare in modo logico e di gestire l'incertezza,
- **Sviluppare una AI più simile a quella umana**: la cognizione umana sembra combinare aspetti sia connessionisti che simbolisti. L'AI neuro-simbolica potrebbe quindi portare a sistemi di AI più simili a quella umana, in termini di flessibilità, creatività e capacità di comprensione.

Per raggiungere i propri obiettivi, l'AI neuro-simbolica si pone questi traguardi:

- **Integrazione di reti neurali e sistemi simbolici**: l'AI neuro-simbolica combina reti neurali artificiali con sistemi di ragionamento simbolico, come la logica del primo ordine o le reti semantiche,

- **Apprendimento e ragionamento**: i sistemi neuro-simbolici sono in grado sia di apprendere da grandi quantità di dati che di ragionare su tali dati utilizzando la conoscenza simbolica,
- **Interpretabilità e trasparenza**: l'AI neuro-simbolica mira a creare sistemi di AI più interpretabili e trasparenti, in cui il processo decisionale è comprensibile agli esseri umani.

Esistono diverse architetture e tecniche per modellare l'AI neuro-simbolica, tra cui:

- **Reti neurali con memoria esterna**: sono reti neurali che utilizzano una memoria esterna per memorizzare e recuperare informazioni simboliche,
- **Sistemi di ragionamento basati su reti neurali**: sono sistemi che utilizzano reti neurali per implementare algoritmi di ragionamento simbolico,
- **Apprendimento di rappresentazioni simboliche da dati**: sono modelli che sfruttano le reti neurali per apprendere rappresentazioni simboliche da dati grezzi,
- **Integrazione di conoscenza simbolica in reti neurali**: sono modelli che incorporano la conoscenza simbolica nella struttura o nei pesi di una rete neurale,
- **Sistemi ibridi che combinano reti neurali e sistemi esperti**: sono sistemi che integrano la capacità di apprendimento delle reti neurali con la conoscenza simbolica dei sistemi esperti.

L'AI neuro-simbolica offre un potenziale significativo per migliorare e ampliare le capacità dell'AI generativa. Ecco alcuni dei suoi contributi chiave:

1. **Migliorare la coerenza e la plausibilità dei contenuti generati**:
    - **Ragionamento e conoscenza del mondo**: i sistemi neuro-simbolici possono integrare la conoscenza del mondo e le regole logiche nel processo generativo, garantendo che i contenuti generati siano coerenti con la realtà e le leggi della fisica. Ad esempio, un sistema neuro-simbolico per la generazione di immagini potrebbe evitare di generare oggetti impossibili o scene surreali,
    - **Coerenza narrativa**: Nell'ambito della generazione di testo, l'AI neuro-simbolica può aiutare a mantenere la coerenza narrativa, evitando contraddizioni e incongruenze nella trama o nei personaggi,
2. **Aumentare il controllo e la direzionalità del processo generativo**:

- **Specificare vincoli e obiettivi**: l'AI simbolica consente di specificare vincoli e obiettivi in modo esplicito, guidando il processo generativo verso risultati desiderati. Ad esempio, si potrebbe specificare che un'immagine generata deve contenere determinati oggetti o rispettare uno stile artistico particolare,
- **Generazione condizionale**: l'AI neuro-simbolica può generare contenuti condizionati a informazioni specifiche, come descrizioni testuali, attributi o esempi,

3. **Abilitare la creatività e l'immaginazione**:
    - **Ragionamento analogico**: i sistemi neuro-simbolici possono utilizzare il ragionamento analogico per generare nuove idee e concetti, combinando elementi esistenti in modi innovativi,
    - **Esplorazione di spazi concettuali**: l'AI simbolica può aiutare a esplorare spazi concettuali complessi, consentendo all'AI generativa di scoprire nuove possibilità e generare contenuti originali,
4. **Rendere l'AI generativa più spiegabile e affidabile**:
    - **Spiegare le decisioni**: l'AI neuro-simbolica può fornire spiegazioni sul processo generativo, rendendo più trasparente il funzionamento dell'AI generativa e aumentando la fiducia negli output,
    - **Verificare la correttezza**: i sistemi simbolici possono essere utilizzati per verificare la correttezza e la coerenza dei contenuti generati, identificando eventuali errori o incongruenze.

In sintesi, l'AI neuro-simbolica offre un insieme di strumenti e tecniche che possono migliorare significativamente le capacità dell'AI generativa, rendendola più potente, versatile e affidabile. L'integrazione di ragionamento simbolico e apprendimento automatico apre nuove frontiere per la creatività e l'innovazione, consentendo di generare contenuti di alta qualità e di risolvere problemi complessi in diversi ambiti.

Ecco alcuni esempi di applicazione nell'ambito dell'AI generativa:

- **Generazione di codice:** Sviluppare sistemi in grado di generare codice sorgente corretto e efficiente, a partire da specifiche in linguaggio naturale.
- **Creazione di contenuti artistici:** Generare opere d'arte originali e creative, combinando l'estetica con il significato simbolico.
- **Design di farmaci:** Accelerare la scoperta di nuovi farmaci, generando molecole con proprietà specifiche.

L'AI neuro-simbolica offre un approccio promettente per potenziare l'AI generativa, superando i limiti dei modelli basati esclusivamente su reti neurali. L'integrazione del ragionamento simbolico può portare a sistemi generativi più spiegabili, controllabili, coerenti e in grado di risolvere problemi complessi. Questa sinergia apre nuove frontiere per l'innovazione in diversi settori, dallo sviluppo di software alla creazione artistica, fino alla ricerca scientifica.

AI Federata

L'intelligenza artificiale (AI) federata è una soluzione innovativa che consente di addestrare i modelli di machine learning su dataset distribuiti, preservando la privacy e la sicurezza dei dati.

L'apprendimento automatico tradizionale si basa sulla centralizzazione dei dati su un singolo server. Tuttavia, questa prassi solleva crescenti preoccupazioni in termini di privacy, sicurezza e sovranità dei dati. L'AI federata offre un'alternativa, consentendo di addestrare modelli su dati decentralizzati, residenti su dispositivi individuali (come smartphone o sensori) o server locali.

Nell'AI federata, un server centrale coordina l'addestramento di un modello globale, mentre i dati rimangono distribuiti tra i client. Il processo tipico prevede le seguenti fasi:

- **Inizializzazione**: il server centrale inizializza un modello globale e lo distribuisce ai client,
- **Addestramento locale**: ogni client addestra il modello sui propri dati locali,
- **Aggregazione**: i client inviano gli aggiornamenti del modello (ad esempio, i gradienti) al server centrale,
- **Aggiornamento del modello globale**: il server aggrega gli aggiornamenti ricevuti dai client e aggiorna il modello globale,
- **Iterazione**: le fasi 2-4 vengono ripetute fino a raggiungere la convergenza del modello.

Ovviamente, l'implementazione di un AI federata deve affrontare diverse sfide connesse alle tecnologie impiegate e alle implicazioni legali, quali:

- **Eterogeneità dei dati**: i dati possono variare in termini di qualità, formato e distribuzione tra i client. L'utilizzo di tecniche come l'apprendimento transfer learning e il federated multi-task learning potrebbero mitigare questo problema,

- **Comunicazione**: la comunicazione tra client e server può essere limitata dalla larghezza di banda e dalla latenza. Lo sfruttamento di tecniche di compressione dei dati e aggiornamento selettivo potrebbero ottimizzare la comunicazione,
- **Privacy**: garantire la privacy dei dati durante l'addestramento è fondamentale. A rigurado si potrebbero applicate soluzioni crittografiche, come la differential privacy e la homomorphic encryption, in grado di proteggere i dati sensibili.

Nell'ambito della GenAI si ricorda che l'addestramento dei modelli generativi richiede enormi quantità di dati, spesso sensibili e distribuiti tra diverse fonti. Qui entra in gioco l'AI federata, un paradigma di apprendimento automatico che consente di addestrare modelli su dataset decentralizzati, preservando la privacy e la sicurezza dei dati.

L'AI federata può potenziare l'AI generativa in questi ambiti:

1. **Accesso a dataset più ampi e diversificati**: l'AI federata consente di aggregare dati provenienti da diverse fonti, come dispositivi mobili, sensori IoT e server aziendali, senza la necessità di centralizzarli. Questo permette di addestrare modelli generativi su dataset più ampi e diversificati, migliorandone la capacità di generalizzazione e la creatività,
2. **Preservazione della privacy**: l'addestramento di modelli generativi su dati sensibili, come informazioni mediche o conversazioni personali, solleva preoccupazioni in termini di privacy. L'AI federata affronta questo problema consentendo di addestrare i modelli localmente sui dispositivi degli utenti, senza condividere i dati grezzi,
3. **Personalizzazione**: l'AI federata consente di addestrare modelli generativi personalizzati per ciascun utente o dispositivo, tenendo conto delle preferenze individuali e del contesto. Questo apre nuove opportunità per la creazione di contenuti personalizzati, come assistenti virtuali più intelligenti e sistemi di raccomandazione più accurati,
4. **Scalabilità**: l'addestramento di modelli generativi su larga scala può essere complesso e dispendioso. L'AI federata distribuisce il carico computazionale tra diversi dispositivi, rendendo l'addestramento più efficiente e scalabile.

Ecco alcuni esempi di applicazioni:

- **Generazione di testo**: addestrare modelli di linguaggio su conversazioni provenienti da diversi utenti, preser-

vando la privacy e personalizzando le risposte,

- **Creazione di musica**: comporre musica originale basata sulle preferenze musicali di diversi utenti, senza condividere le loro librerie musicali,
- **Sintesi di immagini**: generare immagini mediche sintetiche per l'addestramento di algoritmi di diagnosi, senza compromettere la privacy dei pazienti.

Vi sono sfide e direzioni future:

Eterogeneità dei dati: gestire la variabilità dei dati provenienti da diverse fonti.

Comunicazione: ottimizzare la comunicazione tra i dispositivi e il server centrale.

Sicurezza: proteggere gli aggiornamenti del modello da attacchi malevoli.

L'AI federata offre un potenziale enorme per l'AI generativa, consentendo di addestrare modelli più potenti, personalizzati e privacy-preserving. Questo connubio apre nuove frontiere per l'innovazione in diversi settori, dalla creazione di contenuti personalizzati allo sviluppo di nuove applicazioni nell'ambito della sanità e dell'IoT.

AI Quantistica

L'AI quantistica (IAQ) rappresenta un campo di ricerca emergente che esplora l'intersezione tra l'intelligenza artificiale e la meccanica quantistica. Sfruttando i principi della computazione quantistica, come la sovrapposizione e l'entanglement, l'IAQ mira a sviluppare algoritmi e modelli di AI più potenti, efficienti e capaci rispetto a quelli classici, aprendo nuove possibilità per la risoluzione di problemi complessi e la scoperta scientifica. Questo paragrafo esamina i fondamenti dell'IAQ, le sue potenziali applicazioni e le sfide che la ricerca deve affrontare.[105]

L'AI quantistica si basa sull'idea di utilizzare i computer quantistici per eseguire algoritmi di AI. I computer quantistici sfruttano i principi della meccanica quantistica per elaborare le informazioni in modo radicalmente diverso dai computer classici[106]. Questo consente di affrontare problemi complessi che sono intrattabili per i computer

105 *Biamonte, J. et al. (2017). Quantum machine learning. Nature, 549(7671), 195-202.*
106 *Schuld, M., & Petruccione, F. (2018). Supervised learning with quantum computers. Springer.*

classici, aprendo nuove possibilità per l'AI.[107]

I vantaggi che ciò implica sono:

- **Velocità di elaborazione**: i computer quantistici possono elaborare le informazioni a velocità esponenzialmente superiori rispetto ai computer classici, consentendo di addestrare modelli di AI più complessi e di risolvere problemi più velocemente,
- **Efficienza**: gli algoritmi quantistici possono essere più efficienti degli algoritmi classici per determinati compiti di AI, come l'ottimizzazione e la ricerca,
- **Capacità di rappresentazione**: i computer quantistici possono rappresentare e manipolare informazioni in modi che sono impossibili per i computer classici, consentendo di sviluppare nuovi modelli di AI con capacità superiori.

L'IAQ si basa su diversi concetti chiave della meccanica quantistica:

- **Qubit**: a differenza dei bit classici, che possono essere 0 o 1, i qubit possono esistere in una sovrapposizione di entrambi gli stati contemporaneamente,
- **Entanglement**: due o più qubit possono essere entangled, il che significa che sono correlati in modo tale che la misurazione di uno influenza istantaneamente lo stato dell'altro, anche se sono separati da grandi distanze,
- **Sovrapposizione**: la capacità di un sistema quantistico di esistere in più stati contemporaneamente,
- **Algoritmi quantistici**: algoritmi progettati per essere eseguiti su computer quantistici, sfruttando le proprietà dei qubit e dell'entanglement per risolvere problemi in modo più efficiente rispetto agli algoritmi classici.

Le principali applicazioni dell'AI Quantistica sono sintetizzate in:

- **Apprendimento automatico quantistico**: sviluppo di algoritmi di apprendimento automatico che sfruttano i principi della meccanica quantistica per migliorare l'accuratezza, l'efficienza e la capacità di generalizzazione dei modelli di AI,
- **Ottimizzazione quantistica**: utilizzo di algoritmi quantistici per risolvere problemi di ottimizzazione comples-

107 Wiebe, N., Kapoor, A., & Svore, K. M. (2016). Quantum deep learning. Quantum Information & Computation, 16(7-8), 541-587.

si, come la selezione del portafoglio, la logistica e la progettazione di farmaci,

- **Elaborazione del linguaggio naturale quantistica**: sviluppo di modelli di AI quantistica per la comprensione e la generazione del linguaggio naturale, con potenziali applicazioni nella traduzione automatica, nella sintesi vocale e nell'analisi del sentiment,
- **Visione artificiale quantistica**: utilizzo di algoritmi quantistici per migliorare l'accuratezza e l'efficienza dei sistemi di visione artificiale, con applicazioni nel riconoscimento di immagini, nella guida autonoma e nella robotica.

L'AI quantistica è un campo di ricerca ancora giovane, con molte sfide da affrontare quali:

- **Sviluppo di hardware quantistico**: la costruzione di computer quantistici affidabili e scalabili è una sfida tecnologica significativa,
- **Sviluppo di algoritmi quantistici**: sono necessari nuovi algoritmi quantistici specificamente progettati per l'AI,
- **Integrazione con l'AI classica**: È importante sviluppare metodi per integrare l'AI quantistica con quella classica, sfruttando i punti di forza di entrambi gli approcci.

Nell'ambito della GenAI è noto che l'addestramento di modelli generativi complessi richiede risorse computazionali significative e può presentare limiti in termini di efficienza e scalabilità.

È qui che entra in gioco l'AI quantistica, promettendo di superare questi ostacoli e di spingere l'AI generativa verso nuovi orizzonti. Ma cosa può fare esattamente l'AI quantistica per l'AI generativa?

1. **Accelerazione dell'addestramento:** l'addestramento di modelli di AI generativa, come le GAN (Generative Adversarial Networks) e i VAE (Variational Autoencoders), comporta l'ottimizzazione di milioni, se non miliardi, di parametri. I computer quantistici, sfruttando i principi della meccanica quantistica come la sovrapposizione e l'entanglement, possono elaborare informazioni in modo esponenzialmente più veloce rispetto ai computer classici. Questo si traduce in una drastica riduzione dei tempi di addestramento, consentendo di sviluppare modelli più complessi e performanti,
2. **Miglioramento della qualità dei contenuti generati:** l'AI quantistica può migliorare la qualità dei contenuti generati in diversi modi:
    - **Generazione di distribuzioni di probabilità più complesse:** i computer quantistici possono rappresentare

e manipolare distribuzioni di probabilità più complesse rispetto ai computer classici, consentendo ai modelli generativi di creare contenuti più realistici e diversificati,

- **Ottimizzazione di funzioni di costo più sofisticate:** l'ottimizzazione quantistica può trovare soluzioni migliori per funzioni di costo complesse, utilizzate per valutare la qualità dei contenuti generati. Questo porta a modelli più accurati e capaci di generare contenuti di qualità superiore,

3. **Esplorazione di nuovi modelli generativi:** l'AI quantistica apre la strada a nuovi modelli generativi, impossibili da implementare su computer classici. Ad esempio, i modelli generativi basati su algoritmi quantistici potrebbero sfruttare fenomeni come il tunneling quantistico per esplorare lo spazio delle soluzioni in modo più efficiente, scoprendo nuove e creative forme di contenuto.

Ecco alcune applicazioni concrete:

- **Drug discovery:** accelerare la scoperta di nuovi farmaci generando molecole con proprietà specifiche,
- **Scienza dei materiali:** progettare nuovi materiali con caratteristiche desiderate, come resistenza, leggerezza e conducibilità,
- **Creazione di contenuti artistici:** generare opere d'arte originali e innovative, spingendo i confini della creatività,
- **Modellazione finanziaria:** sviluppare modelli più accurati per la previsione dei mercati finanziari e la gestione del rischio.

Nonostante il potenziale rivoluzionario, l'AI quantistica è ancora in una fase iniziale di sviluppo. La costruzione di computer quantistici su larga scala e lo sviluppo di algoritmi quantistici efficienti per l'AI generativa rappresentano sfide significative. Tuttavia, la ricerca in questo campo sta progredendo rapidamente e le prime applicazioni concrete iniziano ad emergere.

L'AI quantistica promette di amplificare il potere dell'AI generativa, aprendo nuove possibilità in diversi settori. L'accelerazione dell'addestramento, il miglioramento della qualità dei contenuti generati e l'esplorazione di nuovi modelli generativi sono solo alcune delle potenzialità di questo connubio dirompente. Sebbene le sfide siano ancora molte, l'AI quantistica è destinata a giocare un ruolo cruciale nell'evoluzione dell'AI generativa e nel suo impatto sul nostro futuro.

Conclusioni

L'ingegneria dell'intelligenza artificiale è una disciplina emergente focalizzata sullo sviluppo di strumenti, sistemi e processi per consentire l'applicazione dell'intelligenza artificiale in contesti del mondo reale.

In contrasto con la corsa prevalente per sviluppare capacità e far progredire singoli strumenti, AI Engineering pone una serie diversa di domande: in che modo l'AI può aiutare gli esseri umani a raggiungere i risultati della missione? Quali sono i limiti dei sistemi di AI nella pratica odierna? Come possiamo garantire che gli standard etici siano rispettati quando i sistemi di AI vengono implementati?

L'aumento della disponibilità di potenza di calcolo e di enormi set di dati ha portato alla creazione di nuova AI, modelli e algoritmi che comprendono migliaia di variabili e sono in grado di prendere decisioni rapide e di impatto. Troppo spesso, però, queste capacità funzionano solo in ambienti controllati e sono difficili da replicare, verificare e convalidare nel mondo reale.

È urgente la necessità di una disciplina ingegneristica che guidi lo sviluppo e l'implementazione delle capacità di intelligenza artificiale. L'ingegneria dell'intelligenza artificiale mira a fornire un framework e strumenti per progettare in modo proattivo sistemi di intelligenza artificiale che funzionino in ambienti caratterizzati da elevati gradi di complessità, ambiguità e dinamismo. La disciplina dell'ingegneria dell'intelligenza artificiale mira a dotare i professionisti degli strumenti necessari per sviluppare sistemi nell'intero spettro enterprise-to-edge, per anticipare i requisiti in mutevoli ambienti e condizioni operative e per garantire che le esigenze umane siano tradotte in un'intelligenza artificiale comprensibile, etica e quindi affidabile.

In tal senso, è opportuno rimarcare che l'ingegneria dell'intelligenza artificiale è una disciplina che combina i principi dell'ingegneria dei sistemi, dell'ingegneria del software, dell'informatica e della progettazione incentrata sull'uomo per creare sistemi di intelligenza artificiale in base alle esigenze umane per i risultati della missione. Alla luce di queste considerazioni è opportuno guidare lo sviluppo dell'AI seguendo questi tre pilastri:

Human-centered AI

La chiave per l'implementazione di un AI responsabile si basa sulla comprensione approfondita delle perso-

ne che utilizzeranno questa tecnologia. Questo principio evidenzia come i sistemi di AI siano progettati per allinearsi con le caratteristiche degli esseri umani, i loro comportamenti e i loro valori.

I sistemi di intelligenza artificiale incentrati sull'uomo sono progettati per funzionare con e per le persone. Man mano che cresce l'esigenza di utilizzare sistemi di intelligenza artificiale, i principi di ingegneria incentrati sull'uomo saranno fondamentali per guidare lo sviluppo dell'AI verso un'implementazione efficace, in grado di ridurre al minimo le conseguenze indesiderate. Identifichiamo tre aree specifiche di attenzione per far progredire l'intelligenza artificiale incentrata sull'uomo:

- I progettisti e i sistemi devono comprendere il contesto di utilizzo e rilevare i cambiamenti nel tempo,
- Occorre sviluppare strumenti, processi e procedure per definire e facilitare la cooperazione tra uomo e macchina,
- Serve implementare metodi, meccanismi e ragionamenti per raggiungere la supervisione critica.

Scalable AI

Lo sviluppo di sistemi di AI efficaci richiede grandi investimenti in termini di tempo e denaro. Questo principio ribadisce il concetto per cui le infrastrutture, i dati e i modelli di AI possano essere riutilizzati in domini con problematiche e implementazioni variabili.

L'AI scalabile è la capacità degli algoritmi, dei dati, dei modelli e dell'infrastruttura di AI di operare alle dimensioni, alla velocità e alla complessità richieste per la missione. La scalabilità è un concetto critico in molte discipline ingegneristiche ed è fondamentale per realizzare la capacità operativa. Identifichiamo tre aree di attenzione per far progredire l'AI scalabile:

- Gestione scalabile di dati e modelli per superare la scarsità di dati e le sfide di raccolta,
- Scalabilità aziendale dello sviluppo e dell'implementazione dell'intelligenza artificiale,
- Algoritmi e infrastrutture scalabili.

Robust and Secure AI

I sistemi di AI robusti e sicuri sono sistemi di AI che operano in modo affidabile all'interno dei livelli di prestazioni previsti, anche quando si trovano di fronte a stati di incertezza o in presenza di pericolo o minacce. Questi sistemi hanno strutture e meccanismi di mitigazioni integrate per prevenire, evitare o fornire resilienza in caso di minacce. Identifichiamo tre aree specifiche di attenzione per promuovere l'AI robusta e sicura:

- Migliorare la robustezza dei componenti e dei sistemi di intelligenza artificiale,
- Progettare per le sfide di sicurezza nei moderni sistemi di intelligenza artificiale,
- Sviluppo di processi e strumenti per testare, valutare e analizzare i sistemi di intelligenza artificiale.

Il futuro dell'Artificial Intelligence non consisterà in un ampliamento, in termini quantitativi e qualitativi, delle capacità cognitive e computazionali di quella attuale, bensì evolverà verso sistemi intelligenti a supporto dell'uomo per lo svolgimento delle più disparate attività (intellettuali, produttive, ricerca, analisi, computazionali, ecc.). Assisteremo allo sviluppo di una disciplina dell'ingegneria dell'AI in grado di fornire la capacità di sviluppare, integrare ed evolvere soluzioni di AI, con un focus specifico sui temi della *"safety"*, della *"security"*, della *"robustness"*, della *"reliability"*, della *"resiliency"* e dell'etica. Questo approccio consentirà di creare e utilizzare prodotti e servizi dell'AI *"Safe, Secure and Trustworthy"*.

A riguardo, l'AI generativa offre enormi opportunità, ma è essenziale affrontare i rischi connessi al suo sviluppo e utilizzo. Affrontare i rischi dell'AI generativa richiede un approccio multidisciplinare che coinvolga ricercatori, sviluppatori, legislatori e la società civile. Un approccio responsabile e proattivo permetterà di sfruttare appieno il potenziale di questa tecnologia, minimizzando al contempo i suoi effetti negativi.

FORUM ICT SECURITY

22-23 OTTOBRE 2025

AUDITORIUM DELLA TECNICA, ROMA

Iscriviti alla newsletter di ICT Security Magazine
per conoscere l'agenda e partecipare alla
23ª Edizione del Forum ICT Security

ICT Security MAGAZINE

ISCRIVITI ALLA NEWSLETTER

per ricevere aggiornamenti sulle prossime iniziative. Seguici sui canali social: Linkedin, Facebook, Twitter