Rischi, prevenzione e responsabilità derivanti dall'utilizzo dell'intelligenza artificiale generativa: nesso di causalità ed *explainability* nei *large language model* gestione del rischio nella consultazione delle banche dati giurisprudenziali

Cod.: P24002

Napoli, Castel Capuano, 16 gennaio 2024

Relatore: Avv. Roberto Arcella

Intelligenza artificiale

- Deterministica = logica condizionale semplice per eseguire azioni specificate *(If this, than that)*

- Generativa = la macchina è espressamente programmata per apprendere da grandi masse di dati ed è in grado di creare contenuti nuovi ed originali

Provvedimenti normativi a tutela dell'utilizzo delle nuove tecnologie

- Normative nazionali ed unionali a tutela dei dati personali
- Direttiva CE 374/1985 sulla responsabilità del produttore
- Codice del Consumo
- Risoluzione del Parlamento UE 16/02/2017
- Risoluzione del Parlamento UE 20/10/2020
- AI Act

One size fits all

One size fits all è l'approccio di sicurezza alternativo a quello *risk based:* nel primo vengono dettate regole uniformi indipendentemente dal livello di rischio (es. GDPR)

AI ACT ed approccio normativo *risk-based*

- **Rischio Inaccettabile**: include sistemi di IA che violano i diritti fondamentali o sono utilizzati in modi che sono considerati manipolativi o ingiusti (es. software di sorveglianza di massa o sistemi di IA che impiegano tecniche di *social scoring* da parte dei governi).
- **Rischio Elevato**: vi rientrano sistemi di IA utilizzati in ambiti critici come la sanità, i trasporti, la giustizia e taluni aspetti del governo. Tali sistemi devono soddisfare requisiti rigorosi in termini di trasparenza, *explainability*, sicurezza e supervisione.
- **Rischio Limitato**: per questa categoria è enfatizzato il requisito della trasparenza e vi rientrano, ad esempio i *chatbot*, per i quali gli utenti dovrebbero essere informati che stanno interagendo con un sistema di IA.
- **Rischio Minimo**: sono generalmente quelli che presentano un basso potenziale di danno o impatto sui diritti e le libertà individuali (es. IA utilizzate in giochi, applicazioni di arte generativa o per creare playlist musicali personalizzate, algoritmi impiegati per filtrare i contenuti indesiderati…).

Giustizia => rischio elevato

- Elevate garanzie di sicurezza
- Trasparenza e informazioni agli utenti
- Documentazione dettagliata
- Supervisione umana
- Misure di gestione del rischio
- Informazione agli utenti quando interagiscono con una macchina (tranne il caso di indagini penali)

ADAS (Advanced Driver-Assistance Systems

Adotta un approccio alla sicurezza che è prevalentemente risk-based, piuttosto che un approccio "*one size fits all*". Questo significa che i sistemi ADAS sono progettati per valutare e rispondere a vari livelli di rischio in situazioni di guida specifiche, piuttosto che adottare una soluzione standard indipendentemente dalla situazione.

Sei livelli di rischio basati sulla categorizzazione SAE da «nessuna automazione» ad «automazione completa»

Graduazione in base alla **quantità ed alla qualità dell'interazione umana**

ADAS (Advanced Driver-Assistance Systems

Automazione condizionata obbligatoria da luglio 2024

- **ISA (intelligent speed assistant)**
- **Scatola nera**
- **Alcolock**
- **Avviso di disattenzione e stanchezza conducente**
- **Segnalazione di arresto di emergenza**
- **Rilevamento in retromarcia e frenata automatica di emergenza**

Proposte di direttive UE

Proposta *Product Liability Directive*

- Aggiornamento della direttiva sulla responsabilità da prodotti difettosi
- Aggiorna le definizioni di difetto e danno per adeguarlo alle nuove realtà tecnologiche

Proposta *AI Liability Directive*

- Adeguamento delle norme sulla responsabilità civile extracontrattuale in relazione all'IA
- Responsabilità diretta (produttori e fornitori di IA)
- Responsabilità indiretta
- (utilizzatori di IA)

Il fattore comune delle tutele risiede nella trasparenza e nella explainability dei modelli di IA ....

“

Explainability

Prompt (input)

Output

... e nella valutazione dell'apporto del controllo umano sull'azione dell'IA

2

Gli LLM e la BDR

Large Language Model e loro applicazione nella Banca dati della Giurisprudenza di merito

La banca dati della giurisprudenza

- Adopera Chat-GPT3
- E’ stata addestrata ed adopera un dataset di circa 4.000.000 di provvedimenti
- Estrazione di lemmi
- Summarization
- Interrogazione con linguaggio naturale

Summarization è...

- **Comprensione del Testo**: Il modello deve prima comprendere il testo originale, identificando le idee principali, i temi chiave e le informazioni rilevanti.
- **Identificazione delle Informazioni Essenziali**: Successivamente, il LLM seleziona le parti più importanti del testo, distingue tra informazioni cruciali e dettagli secondari.
- **Riformulazione e Condensazione**: Il modello riformula le informazioni essenziali in un formato più breve, mantenendo l'essenza e la coerenza del testo originale, con la combinazione di diverse frasi o concetti in una formulazione più sintetica.
- **Mantenimento della Coerenza e della Logica**: Durante il processo di riassunto il modello mantiene una narrazione logica e coerente, assicurando che il riassunto sia comprensibile e fedele al testo originale.
- **Stile e Adattabilità**: lo stile del riassunto viene adattato al contesto o alle preferenze dell'utente, come ad esempio un riassunto formale per un contesto accademico o un riassunto più colloquiale per una conversazione informale.

Summarization non è ragionamento!

- i LLM non "ragionano" nel senso umano del termine; piuttosto, essi generano risposte basandosi su pattern statistici appresi durante il training, per il che non c'è un percorso logico o causale che può essere facilmente spiegato o seguito;

- i risultati (*output*) possono essere viziati, oltre che dai *bias* connaturati alla progettazione del modello ed a quelli relativi alla scelta del dataset, anche da vere e proprie “allucinazioni” nelle quali possono incorrere

La BDR implica un «rischio elevato»? Probabilmente NO

- I sistemi di IA utilizzati nella ricerca giurisprudenziale forniscono tipicamente supporto informativo e **non prendono decisioni autonome**. Non si tratta, in altri termini, di giustizia predittiva, mentre il ruolo dell'IA è principalmente quello di assistere gli operatori del diritto nell'individuazione di precedenti rilevanti;

- Nella ricerca giurisprudenziale, gli avvocati ed i giudici utilizzano gli strumenti basati su IA come **mero ausilio**;

- Da quanto detto ai punti che precedono, diventano centrali la figura non solo dell'operatore "fornitore" del servizio di back-end (la DGSIA), ma anche quella dell'operatore front-end (il magistrato, l'avvocato – se e quando sarà ammesso alla consultazione della BDR che formulano i *prompt*), con le consequenziali ricadute in termini di responsabilità;

- Mentre l'uso di IA in altre aree del diritto (come nella predizione delle decisioni o nel *profiling* degli imputati) può avere un impatto diretto sui diritti fondamentali delle persone, **la ricerca giurisprudenziale tramite IA si concentra principalmente sull'accesso e sull'analisi delle informazioni**.

Grazie per l'attenzione!

SSM

SCUOLA SUPERIORE DELLA MAGISTRATURA