GIUSTIZIA INSIEME

Diritto e innovazione" class="voce">

Le iniziative del Consiglio d'Europa in materia di intelligenza artificiale generativa nella giustizia: un aggiornamento

di Franco De Stefano

5 marzo 2024

Sommario: 1. Premessa. - 2. La nota informativa del Gruppo di Lavoro sulla giustizia cibernetica e sull'intelligenza artificiale. - 2.1. Il funzionamento dell'intelligenza artificiale generativa. - 2.2. I rischi dell'impiego dell'intelligenza artificiale generativa. - 2.3. Come avvalersi dell'intelligenza artificiale generativa. - 2.4. Quando non usare l'intelligenza artificiale generativa. - 3. Cenni al seminario del 20 febbraio 2024.

1. Premessa.

In materia di intelligenza artificiale e giustizia il Consiglio d'Europa ha costituito, nell'ambito dell'*European Cyberjustice Network* e quale articolazione della CEPEJ, cioè la Commissione europea per l'efficienza della giustizia, il Gruppo di lavoro sulla giustizia cibernetica e sull'intelligenza artificiale [*"CEPEJ Working group on Cyberjustice and Artificial Intelligence* (CEPEJ-GT-CYBERJUST)].

Tra le iniziative permanenti, si segnala il "*Resource Centre on Cyberjustice and AI*", che costituisce un sito pubblicamente accessibile per informazioni affidabili sui sistemi di intelligenza artificiale

ed altri strumenti chiave in tema di giustizia cibernetica applicati alla digitalizzazione del sistema giudiziario; e che consente di ottenere uno sguardo d'insieme di tali strumenti, costituendo il punto di partenza per gli approfondimenti sui loro rischi e benefici per i professionisti e gli utenti finali, in linea con la Carta etica europea sull'uso dell'intelligenza artificiale nei sistemi giudiziari e nel loro ambiente (risalente, ormai, già al 2018, reperibile al sito **https://rm.coe.int/charte-ethique-fr-pour-publication-4-decembre-2018/16808f699b**).

Il gruppo di lavoro ha rilasciato, il 12 febbraio scorso, una nota informativa sull'Uso dell'Intelligenza artificiale generativa da parte dei professionisti del diritto in ambito lavorativo (avallata dal Comitato consultivo sull'Intelligenza artificiale - *Artificial Intelligence Advisory Board* - AIAB della CEPEJ; e reperibile all'indirizzo **https://www.coe.int/en/web/cepej/resource-centre-on-cyberjustice-and-ai**); ed ha organizzato, nel pomeriggio del 20 febbraio, un webinar "*European Cyberjustice Network (ECN) Webinar #7/2024 Generative Artificial intelligence (AI) in the field of Justice*"

2. La nota informativa del Gruppo di Lavoro sulla giustizia cibernetica e sull'intelligenza artificiale.

La nota si apre con un paragrafo introduttivo, nel quale fornisce una definizione di Intelligenza artificiale generativa (d'ora in avanti, IAG): si tratta di programmi che comunicano in linguaggio naturale, capaci di fornire risposte a domande relativamente complesse e di creare contenuti (somministrare un testo, un'immagine o un suono) a seguito della formulazione di una domanda o di specifiche istruzioni ("*prompt*"): strumenti che includono OpenAI®, Copilot®, Gemini® e Bard®, tutti in rapida evoluzione. Lo scopo dichiarato della nota è di fornire alcune riflessioni su quanto i giudici e gli altri professionisti del settore della giustizia possono aspettarsi dall'uso degli strumenti di intelligenza artificiale generativa in un contesto giudiziario.

2.1. Il funzionamento dell'intelligenza artificiale generativa.

La nota prosegue con l'esame del modo di funzionamento dell'IAG: questa apprende regole e caratteristiche da ampie raccolte di dati ed è basata sulla comprensione statistica del linguaggio; il suo scopo è definire, con la maggiore possibile affidabilità, la parola più affine, senza conoscerne il significato. Ad esempio, quando il sistema scrive che J.F. Kennedy fu presidente degli Stati Uniti, non è perché si sta basando su di una base di conoscenza con un legame diretto tra i due frammenti di informazione, ma perché, nei dati di addestramento (training data) forniti, ha riscontrato una rilevante frequenza statistica dell'associazione tra Kennedy e "presidente degli Stati Uniti": in tal modo, il programma ha dedotto che questa associazione

doveva essere rilevante. Per la maggior parte, i dati di addestramento sono le informazioni fornite da altri utenti alla macchina attraverso i *prompt*.

IAG sembra offrire buoni risultati in un contesto chiaramente delimitato, come la traduzione di testi o la generazione di testi coerenti (ma non necessariamente veri) o di immagini o suoni, di sommari automatici di testi, di analisi semantiche e di rilevamento di opinioni, il "*text mining*" (tecnica che utilizza l'elaborazione del linguaggio naturale per trasformare il testo libero, non strutturato, di documenti/database in dati strutturati e normalizzati) e l'accesso ai contenuti.

2.2. I rischi dell'impiego dell'intelligenza artificiale generativa.

Quanto ai rischi, la nota del Gruppo di lavoro lucidamente li cataloga separatamente.

In primo luogo, c'è il rischio di potenziale produzione di informazioni fattualmente inaccurate (risposte false, "allucinazioni" e pregiudizi). Le risposte sbagliate possono derivare, prima di tutto, da dati di addestramento insufficienti o sbagliati; da dati falsi originano false risposte. Per allucinazioni si intendono le risposte semplicemente inventate: se non è rinvenuta nessuna risposta, l'algoritmo "inventa" una risposta plausibile o probabile, talvolta per l'elaborazione di una falsa correlazione tra i dati. Fondamentalmente, tutti i sistemi di IAG sono profondamente influenzati dai dati su cui sono stati addestrati. Per questo, essi non sono mai neutrali, ma, al contrario, incorporano tutti i pregiudizi, le inesattezze, le lacune o gli sbagli contenuti nelle basi di dati di addestramento o i pregiudizi culturali di quelli che hanno progettato il sistema e guidato il suo addestramento, validando alcune delle sue risposte. Possono esserci perfino casi in cui un pregiudizio può essere intenzionalmente stato immesso nell'algoritmo. L'opacità di programmazione dell'algoritmo e dei collegamenti dei sottostanti dati porta ad una ulteriore incomprensibilità e quindi a difficoltà nel riscontro di verità delle risposte fornite.

In secondo luogo, c'è il rischio di rivelazione di dati sensibili o riservati.

Le informazioni immesse sono trasmesse al fornitore del sistema e potenzialmente usate come dati di addestramento per gli utenti futuri e per generare futuri risultati: questo può portare ad una violazione della protezione dei dati personali o una non intenzionale rivelazione di dati riservati o altrimenti sensibili. Non è, per lo più, garantita la protezione dei dati trasmessi attraverso i sistemi di IAG; ne deriva che le conversazioni ed i relativi dati sono registrati nei server delle compagnie, spesso non europee, tanto da poter essere rivenduti (o esposti a razzie informatiche, a seconda del livello di sicurezza di questi server).

In terzo luogo, si segnala il pericolo di una perdita dei riferimenti dei dati forniti e di una possibile violazione del diritto di autore o di proprietà intellettuale. C'è scarsa trasparenza sull'origine delle informazioni adoperate per popolare le basi di dati e per l'addestramento. La maggior parte dei sistemi non può elencare e accreditare i testi usati per creare i risultati: e questo può non soltanto causare difficoltà nella verificazione di questi, ma anche integrare violazioni dei diritti d'autore. Mentre le cornici normative differiscono da paese a paese, tuttavia esse si applicano anche all'uso di IA, tanto che il contenuto realizzato potrebbe essere qualificato plagio.

Ancora, è limitata la capacità di fornire la stessa risposta ad una domanda uguale. La maggior parte dei sistemi di IAG contengono un grado di casualità che permette loro di proporre differenti risposte alla stessa domanda: le risposte possono differire, a seconda del momento in cui sono formulate o delle sfumature nella formulazione delle rispettive domande; pertanto, non si può sempre garantire lo stesso livello di qualità nelle risposte.

Inoltre, il risultato dei sistemi di IAG non è in alcun modo unico e può essere identico o simile a quello generato per un altro utente, sicché la sua fonte non dovrebbe mai essere tenuta nascosta. Inoltre, specialmente nel campo giudiziario, è essenziale essere trasparenti sull'uso di IA: la relazione con la parte è basata sulla fiducia.

Infine, mentre è variabile la stabilità e l'affidabilità dei modelli di IAG quanto a tempi di risposta e disponibilità dei servizi offerti (ciò che andrebbe quindi tenuto in debito conto nei processi per i quali il tempo è un fattore determinante), la relazione tra l'uomo e la macchina è di per sé pregiudicata dalle nostre capacità cognitive: anzi, tale relazione tende ad esaltare questi pregiudizi, poiché l'interazione con la macchina aumenta la percezione, da parte di questa, del fatto che quella sia "umana". Lo scambio non è mai neutrale.

2.3. Come avvalersi dell'intelligenza artificiale generativa.

La nota del Gruppo di lavoro prosegue con il suggerimento delle cautele con cui è possibile avvalersi di IAG:

> Assicurarsi che l'uso dello strumento sia autorizzato e appropriato per lo scopo desiderato;
> Tenere a mente che si tratta solo di uno strumento e cercare di capire come funziona, rimanendo consapevoli dei pregiudizi cognitivi umani;
> Preferire i sistemi addestrati su dati certificati e ufficiali, di cui sia conosciuta la lista, per limitare i rischi di pregiudizi, allucinazioni o violazioni del diritto d'autore;
> Dare allo strumento istruzioni chiare (prompt) su cosa si aspetta da quello; è attraverso la conversazione che la macchina otterrà le istruzioni di cui ha bisogno, perciò non esitare a

impegnarla sul punto, non come un motore di ricerca. È auspicabile chiedere un chiarimento o anche modificare o rifinire una domanda, ma è essenziale fornire adeguatamente le istruzioni: è bene dare alla macchina un contesto – geografico o temporale – o definire il compito – scrivere un sommario in un tot numero di parole – o indicare lo scopo al quale deve servire il risultato, il modo in cui questo deve essere prodotto ed il tono che il sistema dovrà adottare, chiedere uno specifico formato di presentazione, controllare che le istruzioni siano state appropriatamente comprese (chiedendo alla macchina di riformularle), fornire un esempio delle risposte attese per domande analoghe per mettere il programma in grado di imitarne la forma e lo stile;
Fornire solo dati non sensibili e informazioni già di pubblico dominio;
Controllare sempre la correttezza delle risposte, anche quando sono indicati precedenti (soprattutto, controllarne l'esistenza);
Essere trasparenti ed indicare se un'analisi o un contenuto sia generato da IAG;
Riformulare il testo generato se debba essere inserito in un documento ufficiale o legale;
Rimanere nel pieno controllo delle scelte e del processo decisionale, mantenendo un occhio critico sulle proposte formulate dalla macchina.

2.4. Quando non usare l'intelligenza artificiale generativa.

Il gruppo di lavoro conclude la sua nota informativa lanciando anche i suoi moniti, invitando a non fare uso di IAG:

Quando non si conoscono, non si comprendono o non si condividono i termini e le condizioni d'uso;
Quando tanto sia vietato dalle regole dell'organizzazione di appartenenza o sia ad esse contrario;
Quando non si è in grado di valutare il risultato quanto a correttezza dei fatti e ad assenza di pregiudizi;
Quando si è richiesti di fornire e quindi svelare dati personali, confidenziali, sensibili o altrimenti protetti dal diritto d'autore;
Quando sia necessario conoscere la fonte della risposta;
Quando si attenda che la risposta sia genuinamente personale.

3. Cenni al seminario del 20 febbraio 2024.

Il seminario, che ha visto la partecipazione di oltre una sessantina di professionisti della Giustizia, ha visto la presentazione di due esperienze nazionali: quella portoghese, di impiego di autentici *chatbot* per una guida pratica all'accesso alla giustizia, aperta al pubblico e con la premessa che non si tratta affatto di somministrazione di consigli legali, ma solo – appunto – di informazioni di orientamento per le procedure da seguire ed i relativi costi (**https://justica.gov.pt/en-gb/Servicos/Justice-Practical-Guide-Beta-Version**), per di più, al momento, limitata ad alcuni specifici settori, verosimilmente di più immediato interesse per la

generalità dei cittadini; quella spagnola, che consente ai professionisti (e, soprattutto, ai giudici) di gestire la mole di informazioni contenute nei testi per estrarne sommari e dati rilevanti ai fini della formazione degli atti successivi, idonei a presentare il contenuto e perfino, talvolta, a conseguentemente classificare l’atto (**https://www.mjusticia.gob.es/es/JusticiaEspana/ProyectosTransformacionJusticia/Documents/Transp**).

Alle relative presentazioni è seguito un interessante dibattito coi relatori in ordine ai limiti, alle potenzialità, all’affidabilità, alla struttura stessa, alla replicabilità dei sistemi. Nessuna specifica esperienza è stata addotta per la realtà italiana.