NOVA 24 Il Sole 24 Ore 04 FEBBRAIO 2024

DIGITALE REGOLE

Intelligenza artificiale, urgente uno standard sulle etichette

Rocco Panetta Vincenzo Tiani

Il centro di ricerca del Parlamento europeo ha recentemente pubblicato un report sull'uso dei watermark per l'AI, tema che si allaccia alle questioni aperte del copyright e della disinformazione, molto attuale visto che nel 2024 il 40% della popolazione andrà al voto. Non riuscire a capire che si sta interagendo con un testo, un messaggio vocale, un video o una fotografia prodotti dall'Ai, potrà avere un peso considerevole alle prossime elezioni.

Il watermark si usa da decenni per le fotografie, a volte in modo visibile all'utente (si pensi a quello di Getty Images), altre in modo visibile solo alla macchina attraverso i metadati, per tracciare una eventuale violazione del copyright. Ma con l'arrivo dell'Ai tutto è cambiato.

Come spiega il report del Parlamento europeo, il watermarking nell'Ai si divide in due fasi, la prima che prevede la creazione durante la fase di addestramento del modello, insegnando al modello a incorporare un identificatore specifico nel contenuto generato. La seconda prevede che sia identificabile e permetta il tracciamento a ritroso verso la fonte.

L'uso di questi sistemi sarebbe salvifico visto che, finora, solo l'ammissione degli autori stessi di un'opera creativa ha permesso di sapere se fosse stata impiegata l'Ai generativa. Ma se questi sembrano dettagli solo per avvocati specializzati in diritto d'autore, gli effetti per il grande pubblico diventano ben tangibili nel caso in cui questi materiali, specialmente audio e video, sono utilizzati per creare deep fake non dichiarati, con possibili effetti nefasti sul dibattito pubblico, come accaduto di recente negli Stati Uniti con la finta dichiarazione del presidente Biden.

Sul punto, Google, Microsoft e Meta sono già al lavoro per trovare una soluzione. Ma non è facile come sembra. Difficilissimi da creare per i testi, anche per le immagini i watermark tuttora usati sono facilmente alterabili e rimovibili, inoltre non esiste uno standard comune e il rischio è quello di fare la fine delle smarthome: ogni brand ha creato il suo ecosistema, ma solo dopo anni hanno iniziato a lavorare insieme condividendo uno standard comune che ha permesso ai device di dialogare tra loro. Quello che urge, dunque, è lavorare a uno standard comune a livello

globale, che tenga conto, soprattutto per quanto riguarda i testi, anche delle differenze linguistiche, per non avere un effetto deterrente sulle lingue diverse dall'inglese. Secondo il report del Parlamento, la Commissione ha chiesto alla European Standardisation Organisation di preparare una serie di standard, inclusi quelli sulla trasparenza verso gli utenti, entro gennaio 2025.

La Cina sul punto sembra avere le idee già molto chiare. «I contenuti generati sono soggetti a un "watermark esplicito", ossia un testo immediato che indichi "generato dall'Ia". Le immagini, i video e gli audio generati dall'IA sono soggetti a un "watermark implicito", ossia a un'etichettatura tecnica che include almeno il nome del fornitore del servizio, impercettibile all'uomo ma tecnicamente rilevabile tramite un'interfaccia». Usa e G7, invece, sono ancora fermi a livello di principi e progetti. L'Ai Act europeo, dal canto suo, prevede un obbligo di trasparenza per i deep fake, per i contenuti generati da AI generativa e per quelli allenati su contenuti coperti da copyright, per facilitare la tracciabilità.

Secondo alcuni esperti però, come accade da anni con i cookie banner, gli utenti non baderanno a questi watermark e, pertanto, la cosa migliore sarebbe quella di sviluppare un linguaggio "tipico dell'Ai", per favorire una netta differenziazione, facilmente riconoscibile dagli utenti durante le loro interazioni. Da ultimo, non bisogna dimenticare che il watermark dovrebbe essere il bollino finale per il consumatore, ma a monte dovranno essere stati rispettati gli obblighi di trasparenza, audit, impatto sui diritti umani, e, non meno importante, sarà necessaria una campagna informativa per il grande pubblico, perché gli utenti capiscano potenzialità e pericoli di questi strumenti.