Lo scontro tra i giornali e OpenAi: blocchi e trattative per l'accesso ai contenuti

di Maria Giusti - pubblicato su "www.irpa.eu" - Osservatorio sullo Stato digitale, 13 dicembre 2023

Importanti gruppi media hanno installato blocchi per impedire che i loro contenuti digitali vengano utilizzati liberamente da OpenAI per addestrare ChatGPT. Il risultato è che attualmente OpenAi non può istruire i suoi modelli accedendo a molte pagine web che contengono informazioni autorevoli e di alta qualità. Perché possa farlo, servirebbe che raggiunga delle intese con gli editori, che richiedono un equo compenso per l'utilizzo dei loro contenuti.

OpenAI, la società creatrice del più famoso e discusso software di intelligenza artificiale relazionale, ChatGPT (ne abbiamo parlato qui, qui, qui, qui e qui), ha dato il via a "GPTBot" durante l'estate 2023. Si tratta di un web crawler, ovvero di un software incaricato di scaricare e indicizzare contenuti dalle pagine web, raccogliendo e organizzando i dati presenti su internet. In un post sul suo blog, OpenAI ha annunciato che lo strumento sarà utilizzato per addestrare i futuri modelli del suo agente conversazionale, filtrando però le fonti soggette a restrizioni di paywall, quelle che violano le politiche di OpenAI e quelle che raccolgono principalmente informazioni di identificazione personale. "Consentire a GPTBot di accedere al tuo sito può aiutare i modelli di intelligenza artificiale a diventare più accurati e a migliorare le loro capacità generali e la sicurezza", si legge nel post. Questo include comunque istruzioni per gli amministratori del web su come impedire al crawler di accedere ai loro siti.

Trascorse due settimane dal lancio di GPTBot, circa il dieci per cento delle pagine web più influenti del mondo ha deciso di bloccare il crawler. Ad oggi, gli operatori che hanno fatto questa scelta sono oltre 530 (si veda l'articolo dell'8 novembre de Il Sole 24 ore "Giornali chiedono a ChatGPT di pagare gli articoli utilizzati"). Tra loro, Amazon, Lonely Planet, il sito di annunci di lavoro Indeed, il sito di domande e risposte Quora e Dictionary.com. Anche numerose testate giornalistiche hanno scelto di impedire ad OpenAI di accedere liberamente ai loro archivi digitali per addestrare i suoi modelli. A importanti gruppi media di lingua inglese (il New York Times, la CNN, la Reuters e il Chicago Tribune, tra gli altri) si sono presto aggiunti diversi operatori francesi, tra cui France 24, TF1 e Radio France. Il risultato è che ad OpenAI è precluso di addestrare i suoi modelli "scandagliando" molte delle pagine web che contengono informazioni provenienti da importanti media internazionali.

La ragione principale dietro la scelta dei giornali è impedire che OpenAi sfrutti a proprio vantaggio i loro contenuti in mancanza di licenze e del riconoscimento di compensi. Se i dati sono la linfa vitale dei servizi di intelligenza artificiale, chi li ha generati investendo a tal fine importanti risorse ha diritto ad una giusta remunerazione. Permettere che OpenaAi si nutra gratuitamente dei contenuti dei giornali non sarebbe possibile anche perché ChatGPT, una volta avuto accesso ai loro archivi, potrebbe divenire un concorrente diretto. Il rischio è che i lettori si limiterebbero ad interrogare l'invenzione di OpenAi senza visitare più le pagine dei giornali.

Una definizione chiara del problema si trova nel Libro bianco "How the Pervasive Copying of Expressive Works to Train and Fuel Generative Artificial Intelligence Systems Is Copyright Infringement And Not a Fair Use". Il documento è stato scritto dalla News Media Alliance, che rappresenta oltre 2.2000 media company in Usa e in Canada, ed inviato al Copyright Office degli Stati Uniti.

Secondo il libro Bianco, «L'Alleanza riconosce i potenziali benefici dell'intelligenza artificiale ed è ampiamente favorevole alle sue applicazioni e tecnologie. Gli interessi degli editori e degli sviluppatori di intelligenza artificiale generativa potrebbero allinearsi, per esempio con l'adozione di accordi che assicurino un compenso equo per le licenze di accesso a materiali formativi di alta qualità. Tuttavia, questa promessa di partnership non si è ancora concretizzata, se non in pochi casi. Al contrario, molti sviluppatori hanno scelto di "raschiare" i contenuti degli editori senza autorizzazione e di utilizzarli per l'addestramento dei loro modelli e per creare prodotti concorrenti. Mentre gli editori fanno gli investimenti e si assumono i rischi, gli sviluppatori raccolgono i frutti in termini di utenti, dati e risorse pubblicitarie […]".
Posizioni simili sono state espresse anche dai media francesi. Laurent Frisch, il direttore della strategia digitale di Radio France, ha per esempio evidenziato come "I nostri contenuti hanno un valore e uno scopo principale, che è quello di distribuirli al grande pubblico per adempiere al nostro compito di servizio pubblico. Non rientra nella nostra missione servire cibo gratuito agli algoritmi. Indicizzando siti le cui informazioni hanno un costo di produzione reale, i bot creano valore per sé stessi a costo zero".
Con il blocco di GPTBot, il settore ha quindi mirato ad ottenere un giusto compenso dai colossi dell'intelligenza artificiale. Secondo alcune fonti, le trattative tra OpenAI e il New York Times non sono andate però a buon fine: dopo settimane di discussioni, le due parti non hanno raggiunto un accordo di licenza che consenta alla società di accedere a titolo oneroso ai materiali della testata giornalistica. OpenAi avrebbe invece raggiunto un accordo con Associated Press sull'uso degli articoli dei suoi associati. Sembrerebbe quindi che, per quanto difficile, il raggiungimento di intese sia possibile. Resta fermo che l'addestramento degli strumenti di OpenAi richiede enormi quantità di dati, rendendo impraticabile il consenso e la remunerazione individuale e necessaria un'azione strutturale più che singole iniziative isolate (si veda l'articolo dell'8 novembre de Il Sole 24 ore già citato).
In ogni caso, il dibattito non si limita a una semplice transizione economica. Le preoccupazioni delle testate giornalistiche non riguarderebbero solo la questione della remunerazione. Bloccando il crawler, i media mirerebbero a scongiurare tanto lo sfruttamento gratuito dei loro contenuti quanto l'associazione di questi ultimi a fenomeni di disinformazione. Una volta utilizzati da tecnologie alimentate da grandi quantità di dati raccolti da internet e non necessariamente affidabili, i contenuti delle testate giornalistiche potrebbero finire per essere associati a notizie false o risultare alterati. Secondo Louis Dreyfus dal Gruppo Le Monde, "L'intelligenza artificiale rappresenta un rischio sistemico per molti gruppi mediatici. Comporta il pericolo che le informazioni prodotte vengano combinate ad altre informazioni di qualità inferiore o che siano distorte". Vincent Fleury, il direttore del digitale a France Médias Monde, afferma che non vogliono che i loro contenuti siano associati a risposte inesatte generate dalla chat di OpenAI. Per il Guardian, si tratta di evitare che gli strumenti di intelligenza artificiale possano distorcere o male interpretare le notizie raccolte dal sito web del giornale.
La questione dell'impiego di contenuti editoriali nel training di algoritmi di intelligenza artificiale è centrale nell'attuale panorama digitale, e servirà del tempo per comprendere come si assesterà. L'auspicio è che i media e OpenAI sapranno trovare degli accordi in grado di valorizzare le esigenze di entrambi. Si tratta di assicurare che l'avanzamento dell'intelligenza artificiale e, di conseguenza, il miglioramento dell'accuratezza delle risposte fornite da ChatGPT, avvenga nel pieno rispetto della protezione della proprietà intellettuale, e che agli editori sia riconosciuta un'equa remunerazione per il valore che creano.