# Sotto il velo dell'Intelligenza Artificiale Generativa: le nuove Linee Guida emanate da *Open AI* per frenare i *deep fake*

24/01/2024 PRIVACY E CYBERSECURITY, PROSPETTIVE, SOCIETÀ.

Jacopo Piemonte
Federico Aluigi

Il 30 novembre 2022, giorno del lancio sul mercato di ChatGTP, è stata senza dubbio una data fondamentale per quanto riguarda l'utilizzo dell'intelligenza artificiale (IA). Tale strumento ha ravvivato il dibattito relativo a questa tecnologia e dimostrato che la stessa potrebbe avere effetti dirompenti nelle nostre vite.

Più particolarmente, l'IA Generativa (*"Generative AI"*) rappresenta una frontiera sempre più critica e complessa nel vasto campo dell'IA. Infatti, a differenza dei modelli convenzionali, che si limitano alla realizzazione di compiti specifici, i sistemi generativi sono progettati per creare autonomamente nuovi contenuti, come testi, immagini e addirittura intere opere d'arte e video. Al cuore di questa innovazione si trova la tecnologia dei modelli generativi come, appunto, ChatGTP (*Generative Pre-trained Transformer*), che ha dimostrato una straordinaria capacità di comprendere e produrre il linguaggio umano in modo coerente.

**Una tecnologia complessa al centro dell'attenzione**

Il funzionamento delle intelligenze artificiali generative è basato sulle reti neurali complesse, intese come modelli computazionali addestrati su enormi quantità di dati. Questi modelli, alimentati da un vasto *corpus* di informazioni, sono in grado di apprendere schemi, stili e strutture linguistiche e visive, consentendo loro di generare contenuti in modo realistico. Questa capacità di creare da sé nuovi contenuti offre un potenziale immenso in svariati settori, dalla creatività artistica alla scrittura automatica di testi complessi.

Tuttavia, con detta innovazione emergono anche importanti questioni

etiche e rischi associati. Uno dei principali aspetti critici riguarda la possibilità di generare contenuti falsi o manipolati in modo convincente, con implicazioni in ambito giornalistico, politico e sociale. La creazione dei *c.d. deepfake*, come dimostrano le manipolazioni di video nel contesto delle elezioni in Taiwan appena concluse[1], è un chiaro esempio di come queste tecnologie possano essere utilizzate per alterare la percezione della realtà: nel caso di specie, era stato diffuso *via social* un falso video del deputato statunitense Rob Wittman che si impegnava a sostenere militarmente Taiwan in caso di affermazione del *Democratic Progressive Party*.

A ciò, si aggiungano fenomeni come la potenziale perdita di controllo sulla produzione di contenuti dannosi o illegali, nonché la perpetuazione di pregiudizi e discriminazioni potenzialmente presenti nei dati di addestramento degli algoritmi. È fondamentale, di conseguenza, affrontare questi rischi sviluppando meccanismi di regolamentazione e controllo che garantiscano un utilizzo etico e responsabile di tali tecnologie.

**A che punto siamo con la regolamentazione di tale fenomeno?**

Su queste premesse, occorre notare come il legislatore europeo sia stato colto "di sorpresa" dall'arrivo di ChatGTP e dei suoi epigoni. La prima proposta di Regolamento sull'Intelligenza Artificiale (c.d. *AI ACT*) risale infatti all'aprile 2021 e non prendeva in considerazione tali tecnologie (le quali sono salite agli onori della cronaca soltanto alla fine del 2022). Di fronte a questa "lacuna", gli artefici dell'AI Act hanno prontamente adottato misure correttive, e nel corso del perfezionamento normativo, l'Unione Europea ha tentato di sviluppare un approccio efficace per la regolamentazione delle intelligenze artificiali generative.

A riguardo, il 19 novembre 2023, Italia, Francia e Germania avevano diffuso un documento informale delineando un possibile quadro regolamentare per i modelli di fondazione in quest'ambito. Questo documento proponeva un approccio *soft*, suggerendo la possibilità di adottare una forma di autoregolamentazione obbligatoria attraverso l'implementazione di codici di condotta[2].

Successivamente, l'accordo finale tra il Parlamento Europeo e il Consiglio dell'8 dicembre 2023 ha introdotto modifiche significative, optando per una maggiore cautela nel regolare la materia[3]. In base a tale accordo, i sistemi di intelligenza artificiale con scopi generali, definiti come "*General Purpose AI*" (GPAI), inclusi anche i modelli di IA Generativa, saranno tenuti a conformarsi ai requisiti di trasparenza originariamente proposti dal Parlamento. Questi requisiti includono l'obbligo per gli sviluppatori di tali sistemi di redigere documentazione tecnica, rispettare la legislazione sul diritto d'autore dell'Unione e fornire dettagliati resoconti sui dati utilizzati per l'addestramento degli algoritmi. In particolare, peraltro, per i modelli GPAI ad alto impatto che presentano rischi sistemici sono previsti obblighi ancora più rigorosi. Se tali modelli soddisfano determinati criteri, gli sviluppatori devono infatti condurre valutazioni e analisi mirate a mitigare i rischi sistemici. Inoltre, sono tenuti a eseguire test specifici e a segnalare alla Commissione Europea eventuali incidenti gravi derivanti dall'uso di tali modelli.

**Le Linee Guida di Open AI contro le *deep fake***

Come noto, il processo di finalizzazione dell'AI Act sta raggiungendo una fase cruciale. Tuttavia, è evidente che il periodo richiesto per l'effettiva implementazione di questa nuova legislazione potrebbe essere considerevole. Sono previsti, infatti, diversi mesi transitori per consentire alle

[1] Per ulteriori informazioni si veda il seguente LINK.
[2] Per ulteriori informazioni, si veda il nostro precedente contributo al seguente LINK.
[3] Per ulteriori informazioni, si veda il nostro precedente contributo al seguente LINK.

società di conformarsi alle sue disposizioni dopo la pubblicazione dell'AI Act sulla Gazzetta Ufficiale (che si vocifera potrebbe arrivare nella prima metà del 2024). Nel frattempo, d'altro conto, l'evoluzione tecnologica non si arresta e si pone sempre più la domanda su come regolare questo settore in assenza di una legislazione unitaria e coerente.

In questo contesto, appaiono rilevanti le mosse prese da Open AI. Il colosso americano (che è stata ultimamente protagonista in negativo[4] sulle cronache) si è ora fatta parte attiva e ha pubblicato in data 16 gennaio 2024 delle Linee Guida[5] contro l'uso dell'IA nelle campagne di disinformazione ("Linee Guida di Open AI"), primariamente in ragione delle numerose tornate elettorali che si avvicenderanno nel 2024 in tutto il mondo[6].

In particolare, la Società, tramite il suo sito web, dichiara il proprio impegno verso la sicurezza nell'ottica delle imminenti elezioni politiche che si susseguiranno nel 2024 in numerosi Stati, attraverso la predisposizione di una serie di misure.

In primo luogo, viene posta un'enfasi particolare sulla prevenzione di possibili abusi[7]. Open AI ha assunto l'impegno di anticipare, nella maggiore misura possibile, minacce come i *deepfakes* ingannevoli, le operazioni di influenza su larga scala e l'impiego di *chatbot* per impersonare candidati nel corso delle elezioni. Prima del lancio di nuovi sistemi, Open AI dichiara dunque di condurre rigorosi test, coinvolgendo utenti e collaboratori esterni per raccogliere *feedback*, e di implementare misure di sicurezza finalizzate a ridurre il potenziale rischio. Nelle stesse Linee Guida di Open AI viene peraltro sottolineato come la stessa Società sia ancora nel processo di apprendimento delle modalità con cui gli individui possano abusare della tecnologia in esame. In considerazione di ciò, le misure provvisorie per evitare abusi che Open AI avrebbe implementato consistono in: i) un divieto di sviluppo di applicazioni da utilizzarsi in campagne politiche e attività di *lobbying*; ii) un divieto di sviluppo di *chatbot* che fingano di essere persone reali (ad esempio, candidati) o istituzioni; iii) un divieto di sviluppo di applicazioni che scoraggino la partecipazione ai processi democratici, ad esempio, rappresentando in modo distorto i processi di voto e le qualifiche (quando, dove o chi è idoneo a votare) o che scoraggino il voto sostenendone la futilità; iv) un sistema di report delle potenziali violazioni a disposizione degli utenti.

In secondo luogo, le Linee Guida prospettano una maggiore trasparenza sui contenuti generati dall'intelligenza artificiale[8]. In questo ambito l'attenzione è posta su un modello *text-to-image* sviluppato da Open AI, denominato DALL-E 3[9]. Tale applicazione utilizza metodologie di *deep learning* per generare immagini digitali partendo da descrizioni in linguaggio naturale, anche dette "*prompt*". A questo riguardo, le Linee Guida anticipano l'imminente implementazione delle credenziali digitali elaborate dalla *Coalition for Content Provenance and Authenticity*[10], che dovrebbero codificare e "restituire" all'utente i dettagli sull'origine delle immagini generate da DALL-E 3. Per quanto riguarda, d'altra parte, la

[4] Per ulteriori informazioni, si veda il nostro precedente contributo al seguente LINK.
[5] Per ulteriori informazioni, si veda il seguente LINK.
[6] Oltre alle elezioni europee, alcuni tra gli Stati che andranno al voto sono Stati Uniti, Russia, Messico, Brasile e India.
[7] Per ulteriori informazioni, si veda il paragrafo "*Preventing Abuse*" delle Linee Guida.
[8] Per ulteriori informazioni, si veda il paragrafo "*Transparency around AI-generated content*" delle Linee Guida.
[9] DALL-E 3 è un modello *text-to-image* sviluppato da Open AI utilizzando metodologie di *deep learning* per generare immagini digitali partendo da descrizioni in linguaggio naturale, anche dette "*prompt*".
[10] Per ulteriori informazioni, si veda il seguente LINK.

piattaforma ChatGPT, si evidenzia come quest'applicazione si stia integrando sempre di più con fonti esistenti di informazioni. Ad esempio, utilizzando lo strumento in parola, gli utenti inizieranno ad avere accesso a notizie in tempo reale a livello globale, con l'inclusione dei relativi *link* riferiti alle fonti. La trasparenza sulla provenienza delle informazioni e l'equilibrio nelle fonti delle notizie potranno così aiutare gli elettori a valutare meglio le informazioni e decidere autonomamente a cosa possono dare fiducia.

Infine, viene riportata una collaborazione in corso d'opera[11], negli Stati Uniti, con la *National Association of Secretaries of State* (NASS), l'organizzazione professionale *non-partisan* più antica del paese per funzionari pubblici. Nel contesto di questa *partnership*, ChatGPT è stato progettato per indirizzare gli utenti verso *CanIVote.org*, l'autorevole sito che fornisce informazioni sul voto negli Stati Uniti quando sorgono domande procedurali legate alle elezioni, ad esempio circa la localizzazione dei seggi elettorali o dei requisiti per l'ammissione al voto.

**Considerazioni conclusive**

In un momento cruciale per la società civile, l'impatto rivoluzionario dell'Intelligenza Artificiale Generativa è innegabile. Sorgono rischi significativi, soprattutto nella diffusione di *fake news*, rendendo imperativo un intervento regolamentare tempestivo per mitigare tali sfide. L'Europa ha risposto positivamente a questa chiamata, preparandosi ad implementare l'AI Act per regolare l'uso dell'Intelligenza Artificiale.

È degno di nota che i tempi di attuazione potrebbero protrarsi a lungo, soprattutto in rapporto al rapido sviluppo delle IA generative. In questo periodo di transizione, è possibile che gli operatori nel campo dell'intelligenza artificiale scelgano di autoregolamentarsi, promuovendo misure regolatorie e iniziative di *soft law* "ad interim" *ad hoc*, seguendo l'esempio di Open AI. In ogni caso, per affrontare immediatamente le sfide emergenti, è essenziale un dialogo continuo tra tutti gli *stakeholders*. Solo questo potrà infatti garantire che la regolamentazione si mantenga al passo con l'evoluzione accelerata della tecnologia generativa.

[11] Per ulteriori informazioni, si veda il paragrafo "*Improving access to authoritative voting information*" delle Linee Guida.


Jacopo Piemonte

**ASSOCIATE**

j.piemonte@dejalex.com

+39 02 72554.1

Via San Paolo 7
20121 – Milano

+32 (0)26455670

Chaussée de La Hulpe 187
1170 - Bruxelles



Federico Aluigi

**ASSOCIATE**

f.aluigi@dejalex.com

+32 (0)26455670

Chaussée de La Hulpe 187
1170 - Bruxelles



MILANO
Via San Paolo, 7 · 20121 Milano, Italia
T. +39 02 72554.1 · F. +39 02 72554.400
milan@dejalex.com

ROMA
Via Vincenzo Bellini, 24 · 00198 Roma, Italia
T. +39 06 809154.1 · F. +39 06 809154.44
rome@dejalex.com

BRUXELLES
Chaussée de La Hulpe 187 · 1170 Bruxelles, Belgique
T. +32 (0)26455670 · F. +32 (0)27420138
brussels@dejalex.com

MOSCOW
Potapovsky Lane, 5, build. 2, 4th floor, office 401/12/9 · 101000, Moscow, Russia
T. +7 495 792 54 92 · F. +7 495 792 54 93
moscow@dejalex.com