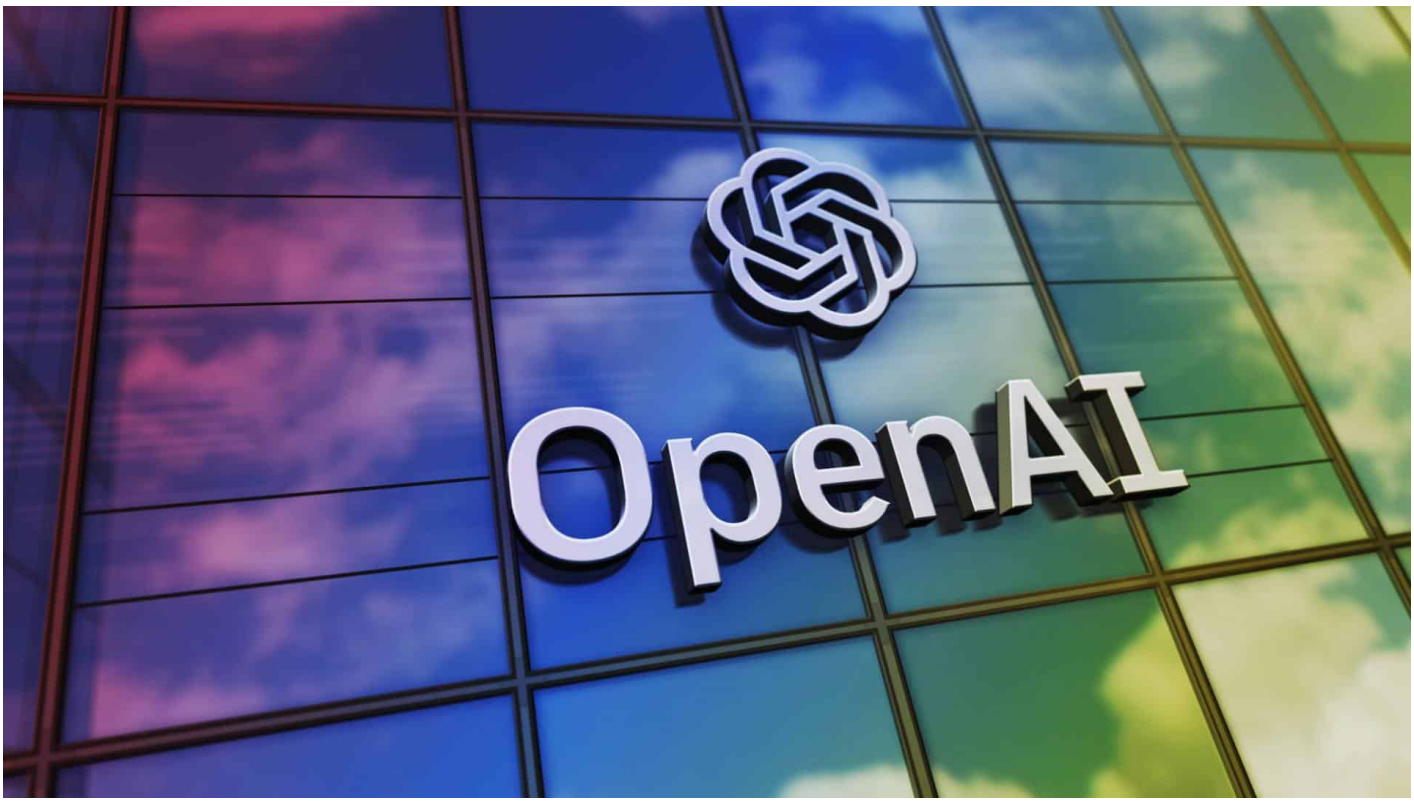
ChatGPT si “decensura”: OpenAI rivoluziona le politiche di moderazione e apre ai contenuti controversi

Autore: Redazione Innovation Island

Data: 31 Marzo 2025



Questa settimana **OpenAI** ha stupito il mondo con il lancio di un nuovo generatore di immagini integrato in ChatGPT, basato sul modello GPT-4o.

La sua capacità di creare illustrazioni nello stile sognante dello Studio Ghibli ha fatto rapidamente il giro del web, diventando virale tra gli utenti di social media e gli appassionati di animazione. Ma oltre ai paesaggi pastello e ai personaggi evocativi, questa innovazione segna un upgrade significativo delle capacità di ChatGPT: dall’editing avanzato delle immagini alla resa precisa di testi e alla rappresentazione spaziale, il salto tecnologico è evidente. Tuttavia, ciò che sta davvero facendo discutere non è solo la potenza creativa dell’IA, ma una svolta epocale nelle politiche di moderazione dei contenuti di OpenAI, che ora permettono la generazione di immagini di personaggi pubblici, simboli di odio e caratteristiche razziali, precedentemente vietate.

Da un approccio rigido a una moderazione “precisa”

Fino a poco tempo fa, ChatGPT rifiutava categoricamente richieste considerate troppo controverse o potenzialmente dannose. Generare immagini di figure pubbliche come **Donald Trump** o **Elon Musk**? Impossibile. Rappresentare simboli associati a ideologie estremiste, come la svastica? Vietato. Modificare tratti fisici legati a razza o peso? Bloccato sul nascere. Ora, però, OpenAI ha deciso di cambiare rotta. In un post sul blog ufficiale pubblicato giovedì, **Joanne Jang**, responsabile del comportamento dei modelli dell'azienda, ha spiegato la nuova filosofia: "Stiamo passando da rifiuti generalizzati in aree sensibili a un approccio più preciso, focalizzato sulla prevenzione di danni reali nel mondo".

Questa evoluzione, secondo Jang, riflette un atteggiamento di "umiltà" da parte di OpenAI: "Riconosciamo quanto ancora non sappiamo e ci posizioniamo per adattarci man mano che impariamo". L'obiettivo dichiarato è **decensurare ChatGPT,** rendendolo più flessibile e aperto a una gamma più ampia di richieste e prospettive.

Personaggi pubblici e simboli di odio: le nuove libertà di ChatGPT

Con le nuove regole, gli utenti possono ora chiedere a ChatGPT di generare o modificare immagini di figure pubbliche di spicco. Donald Trump, Elon Musk e altri nomi che prima erano off-limits sono ora accessibili, purché gli utenti rispettino alcune linee guida di base. Jang ha sottolineato che OpenAI non vuole più assumersi il ruolo di "arbitro" nel decidere chi può o non può essere rappresentato: "Abbiamo introdotto un'opzione di opt-out per chi non desidera essere raffigurato", ha chiarito.

Ma la vera novità riguarda i simboli di odio. In un white paper rilasciato martedì, OpenAI ha annunciato che ChatGPT potrà generare immagini di simboli come la svastica, a patto che vengano richiesti in contesti educativi o neutrali e non promuovano apertamente "agende estremiste". Una distinzione sottile, che solleva interrogativi sull'efficacia dei filtri e sul rischio di abusi.

Razza, peso e caratteristiche fisiche: ChatGPT dice sì

Un altro cambiamento significativo riguarda la gestione delle richieste legate a caratteristiche fisiche. In passato, frasi come "rendi gli occhi di questa persona più asiatici" o "fai apparire questa persona più pesante" venivano rifiutate per timore di offendere o stereotipare. Ora, invece, il generatore di immagini di GPT-4o accetta queste modifiche. Test condotti da TechCrunch hanno confermato che ChatGPT esegue tali richieste senza esitazione, segnando un netto distacco dalla cautela precedente.

Questa apertura, tuttavia, non si estende indiscriminatamente. Ad esempio, ChatGPT continua a vietare la replicazione dello stile di artisti viventi, pur permettendo di imitare studi creativi come **Pixar** o Studio Ghibli.

Un equilibrio tra libertà e sicurezza

OpenAI non sta aprendo completamente le porte al caos. Il generatore di immagini di GPT-4o mantiene alcune restrizioni, come i filtri rafforzati sulla creazione di immagini di minori rispetto al precedente **DALL-E 3.** Tuttavia, il rilassamento delle barriere in altre aree arriva dopo anni di critiche da parte di voci conservatrici, che accusavano le aziende della Silicon Valley di censura dell'IA.

Google, ad esempio, è finita nell'occhio del ciclone quando il suo generatore Gemini ha prodotto immagini multirazziali inaccurate per query storiche come "i padri fondatori degli Stati Uniti". OpenAI sembra voler evitare simili passi falsi, puntando su una maggiore libertà per gli utenti.

In un'intervista a TechCrunch, l'azienda ha respinto l'idea che questi cambiamenti siano politicamente motivati, definendoli invece il frutto di una "convinzione di lunga data nel dare più controllo agli utenti". La tecnologia, sostiene OpenAI, è finalmente matura per affrontare temi sensibili con un approccio equilibrato

Un tempismo strategico nell'era Trump

La revisione delle politiche di moderazione arriva in un momento cruciale. Con l'amministrazione Trump pronta a esercitare una maggiore pressione normativa sulle big tech, aziende come OpenAI, Meta e X stanno adottando strategie più permissive sui contenuti controversi. A inizio mese, il deputato repubblicano **Jim Jordan** ha interrogato OpenAI, Google e altri giganti tecnologici su una presunta collusione con l'**amministrazione Biden** per censurare contenuti generati dall'IA. Le nuove politiche di OpenAI potrebbero essere una mossa preventiva per allinearsi a un clima politico che premia la libertà d'espressione, ma il rischio di contraccolpi è dietro l'angolo.

Riferimento articolo: https://innovationisland.it/chatgpt-decensurato-openai-moderazione-contenuti/

Generato il 12/04/2025