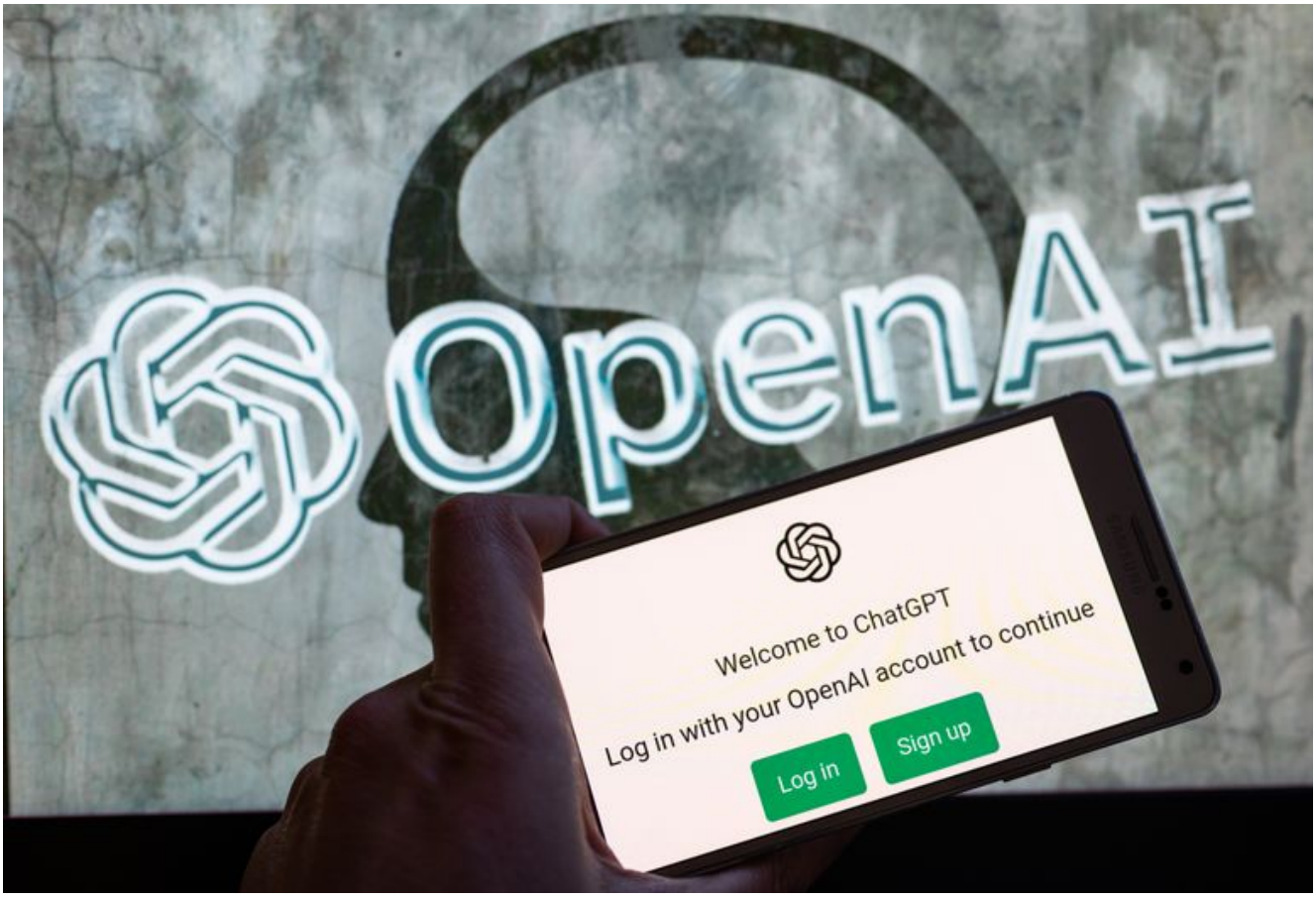
Dietro ChatGPT c’è un esercito di addestratori sottopagati

Per imparare a distinguere un semaforo, un **algoritmo di deep learning** deve **passare in rassegna centinaia di migliaia di immagini** in cui è segnalata la presenza di semafori finché non è in grado di **riconoscerle in autonomia**. Ma chi è che etichetta in primo luogo le **immagini utilizzate per l’addestramento**, indicando quali figure – semafori, gatti, persone, ponti e quant’altro – sono presenti al loro interno?

Benvenuti nel mondo dei *data labeler*, gli **etichettatori di dati**: lavoratori umani al livello base della progettazione di software di deep learning e che operano all’interno di quelli che spesso vengono definiti gli **“scantinati dell’intelligenza artificiale”**. Scantinati che possono avere l’aspetto di fabbriche specializzate nell’etichettatura dei dati (spesso situate in nazioni in via di sviluppo), ma anche essere **piattaforme che assoldano lavoratori da remoto** o per cui, inconsapevolmente, lavoriamo gratuitamente anche noi (com’è il caso dei Captcha Code).

Il ruolo di questi operai del deep learning non è solo di **addestrare le intelligenze artificiali** a distinguere determinati elementi di ogni tipo, ma anche di insegnare loro quali forme di linguaggio, immagini e situazioni vanno a tutti i costi evitate. È questo il caso di **Sama**, una società con

sede a San Francisco ma i cui lavoratori operano in uffici situati in **Kenya, Uganda e India**. E che ha svolto un ruolo cruciale nell'addestrare **ChatGPT** – lo strabiliante sistema di intelligenza artificiale di **OpenAI** in grado di creare testi di ogni tipo – a evitare di produrre contenuti inappropriati.



Il repulisti dei dati

Facciamo un passo indietro. Quando, nel 2020, **OpenAI presentò al mondo GPT-3** (il *large language model* su cui è basato ChatGPT) in breve si capì che questa intelligenza artificiale presentava lo stesso problema già riscontrato in altri sistemi simili: poteva facilmente essere spinta a **creare testi razzisti, sessisti, omofobi, violenti** e quant'altro. In alcuni casi, i testi potevano contenere elementi di *hate speech* senza nemmeno che fosse stata spronata a produrli.

È quasi inevitabile che avvenga qualcosa del genere. D'altra parte, come tutti gli altri sistemi di deep learning, GPT-3 non fa che **rielaborare e ricombinare il materiale di**

partenza con cui è stato addestrato. Quando si viene addestrati con centinaia di gigabytes di contenuti testuali reperiti online, è quasi sicuro che al loro interno ci sia **una certa percentuale di contenuti odiosi**: un grosso limite alla diffusione commerciale.

E così, come scrive *Time* nella sua inchiesta con cui ha svelato il ruolo di queste attività, per risolvere il problema in vista del **lancio di ChatGPT**, OpenAI *"ha strappato una pagina dal libro di società come Facebook, che avevano già mostrato come fosse possibile creare delle intelligenze artificiali in grado di riconoscere casi di linguaggio tossico e aiutare così a rimuoverli dalle loro piattaforme"*. Il metodo è lo stesso con cui si insegna a un algoritmo di deep learning a riconoscere un semaforo: è necessario dargli in pasto una tale quantità di **esempi testuali di violenza, molestie sessuali, bullismo**, ecc. da permettergli di imparare a riconoscerli in autonomia.

Etichettatori umani

Come già visto nell'esempio dei semafori, anche per insegnare agli algoritmi a riconoscere contenuti violenti di ogni tipo è necessario che ci sia in primo luogo **qualcuno che analizza questi contenuti** e li etichetta come tali. Ed è qui che entra in gioco Sama: società che ha tra i suoi clienti anche **Meta, Microsoft e Google** e che, nel 2021, ha stretto un accordo commerciale con OpenAI al fine di etichettare il materiale necessario a creare **un "detector" di contenuti tossici**, che sarebbe poi stato integrato in ChatGPT.

*"Per ottenere queste etichette, OpenAI ha inviato **decine di migliaia di campioni di testo a Sama** a partire dal novembre 2021 –* prosegue *Time –. Buona parte di questi contenuti sembrano essere stati prelevati dai **più oscuri recondi della rete**. Alcuni descrivono con dettagli espliciti abusi sessuali su bambini, bestialità, omicidi, suicidio, tortura, autolesionismo e incesto"*.

Tutto ciò, inevitabilmente, significa che il lavoro degli etichettatori assoldati da Sama consisteva nel **leggere tutto il giorno i più terribili contenuti** partoriti dalla mente umana, per poi etichettarli in base alle loro caratteristiche: *"Un passo necessario per minimizzare la quantità di contenuti violenti e sessuali inclusi nei dati di addestramento e per **creare strumenti in grado di individuare contenuti nocivi**"*, ha spiegato un portavoce di OpenAI.

Sotto pressione

Come già avvenuto nel caso dei **moderatori di social network**, questo tipo di lavoro è però estremamente pesante: un lavoratore di Sama con il compito di leggere ed etichettare testi per OpenAI ha spiegato al *Time* di aver sofferto di **pensieri ossessivi** dopo aver letto la descrizione di un uomo che faceva sesso con un cane in presenza di un bambino. ***"È stata una tortura*** – ha spiegato il moderatore - . *Leggi una quantità di materiale del genere per tutta la settimana. Ora che arriva il venerdì, continuare a pensare a quelle immagini causa seri disturbi"*.

Il lavoro era stato organizzato dividendo i lavoratori in tre squadre, ciascuna delle quali si focalizzava su **violenze sessuali, hate speech o violenza generica**. Tre impiegati hanno spiegato al *Time* come *"ci si aspettava che fossero letti tra 150 e 250 estratti testuali in turni di nove ore. Questi estratti potevano andare dalle 100 a ben oltre mille parole l'uno"*. Tutti gli impiegati sentiti dal *Time* hanno spiegato di

essere stati ***"psicologicamente spaventati"*** dal loro lavoro (teoricamente i lavoratori avevano a disposizione delle sessioni psicologiche individuali, ma hanno spiegato di aver potuto usufruire, nonostante le loro richieste, soltanto di quelle di gruppo).

E **quanto si viene pagati per un lavoro di questo tipo**, che tra le altre cose richiede anche di assumersi la responsabilità di comprendere il contesto di certi testi o affermazioni, di **interpretare le inevitabili ambiguità**, di distinguere la satira e molto altro ancora? I lavoratori di Sama impiegati in Kenya ricevevano, a seconda del grado di anzianità, tra **1,3 e 2 dollari all'ora**; in una nazione in cui il salario minimo si aggira attorno a 1,5 dollari.

Per un breve periodo, gli etichettatori di Sama hanno anche dovuto lavorare, oltre che sui testi, anche su **immagini di violenza, stupri, uccisioni** e altri contenuti di questo tipo, probabilmente al fine di addestrare il sistema di deep learning – sempre di OpenAI – **Dall-E 2** (che genera immagini). Un'incomprensione con OpenAI relativa alla necessità di **raccogliere e visionare materiale illegale** ha però portato Sama a decidere di interrompere il contratto già nel febbraio 2022, otto mesi prima della scadenza.

Come detto, non è la prima volta che si viene a sapere delle **condizioni estremamente difficili** – sotto vari punti di vista – a cui sono sottoposti i moderatori dei social network da una parte e gli etichettatori di dati dall'altra. In questa occasione, inoltre, gli elementi negativi di entrambe le professioni si sono combinati, peggiorando ulteriormente la situazione. E contribuendo a svelare ciò che, spesso, si cela **dietro gli ultimi scintillanti algoritmi di intelligenza artificiale**.