16.Aprile

Quando l'intelligenza va in "pausa"

La pausa, anche essa fa parte della musica.
Stefan Zweig

Una lettera aperta che chiede una pausa sullo sviluppo di sistemi avanzati di intelligenza artificiale (AI) ha diviso i ricercatori. Attirando le firme di artisti del calibro del CEO di Tesla Elon Musk e del co-fondatore di Apple Steve Wozniak, **la lettera (*allegato 1)*** pubblicata all'inizio della scorsa settimana, sostiene una moratoria di 6 mesi per dare alle società di intelligenza artificiale e alle autorità di regolamentazione il tempo di formulare salvaguardie per proteggere la società da potenziali rischi di la tecnologia.

L'intelligenza artificiale ha galoppato dal lancio lo scorso anno del generatore di immagini DALL-E 2, della società OpenAI supportata da Microsoft. Da allora la società ha rilasciato ChatGPT e GPT-4, due chatbot che generano testo, ottenendo un successo frenetico. La capacità di questi cosiddetti modelli "generativi" di imitare i risultati umani, combinata con la velocità di adozione (secondo quanto riferito, ChatGPT ha raggiunto più di 100 milioni di utenti entro gennaio

e le principali aziende tecnologiche stanno correndo per incorporare l'IA generativa nei loro prodotti) hanno catturato molti di sorpresa.

"Penso che le intuizioni di molte persone sull'impatto della tecnologia non siano ben calibrate rispetto al ritmo e alla portata di [questi] modelli di intelligenza artificiale", afferma il firmatario della lettera **Michael Osborne,** ricercatore di machine learning e co-fondatore della società di intelligenza artificiale Mind Foundry.

mind foundry

È preoccupato per l'impatto sociale dei nuovi strumenti, compreso il loro potenziale di licenziare le persone e diffondere la disinformazione. "*Sento che una pausa di 6 mesi darebbe alle autorità* di regolamentazione abbastanza tempo per mettersi al passo con il rapido ritmo dei progressi", afferma.

La lettera, pubblicata da un'organizzazione senza scopo di lucro chiamata *Future of Life Initiative*, irrita alcuni ricercatori invocando danni lontani e speculativi. Chiede: “Dovremmo sviluppare menti non umane che alla fine potrebbero essere più numerose, superate in astuzia, obsolete e sostituirci? Dovremmo rischiare di perdere il controllo della nostra civiltà?

Sandra Wachter, esperta di regolamentazione tecnologica presso l'Università di Oxford,

afferma che ci sono molti danni noti che devono essere affrontati oggi. Wachter, che non ha firmato la lettera, afferma che l'attenzione dovrebbe essere concentrata su come i sistemi di intelligenza artificiale possono essere motori di disinformazione, convincendo le persone di informazioni errate e potenzialmente diffamatorie; come perpetuano il pregiudizio sistemico nelle informazioni che trasmettono alle persone; e come fanno affidamento sul lavoro invisibile dei lavoratori, che spesso lavorano duramente in condizioni precarie, per etichettare i dati e addestrare i sistemi.

La privacy è un'altra preoccupazione emergente, poiché i critici temono che ai sistemi possa essere richiesto di riprodurre esattamente le informazioni di identificazione personale dai loro set di formazione.

L'autorità italiana per la protezione dei dati personali ha bandito **ChatGPT**

il 31 marzo per timore che i dati personali degli italiani vengano utilizzati per addestrare i modelli di OpenAI. (Un post sul blog di OpenAI afferma : "*Lavoriamo per rimuovere le informazioni personali dal set di dati di addestramento ove possibile, perfezionare i modelli per rifiutare le richieste di informazioni personali di privati e rispondere alle richieste degli individui di eliminare le proprie informazioni personali dai nostri sistemi. ”*

Alcuni esperti di tecnologia avvertono di minacce alla sicurezza più profonde. Gli assistenti digitali pianificati basati su **ChatGPT** in grado di interfacciarsi con il Web e leggere e scrivere e-mail potrebbero offrire nuove opportunità agli hacker, afferma

Florian Tramèr, informatico presso l'ETH di Zurigo.

Gli hacker si affidano già a una tattica chiamata "prompt injection" per indurre i modelli di intelligenza artificiale a dire cose che non dovrebbero, come offrire consigli su come svolgere attività illegali. Alcuni metodi prevedono di chiedere allo strumento di interpretare il ruolo di un confidente malvagio o di agire come traduttore tra lingue diverse, il che può confondere il modello e spingerlo a ignorare le sue restrizioni di sicurezza.

Tramèr teme che la pratica possa evolversi in un modo per gli hacker di ingannare gli assistenti digitali attraverso una "iniezione di prompt indiretta", ad esempio inviando a qualcuno un invito del calendario con le istruzioni per l'assistente di esportare i dati del destinatario e inviarli all'hacker. "Questi modelli verranno sfruttati a destra ea sinistra per far trapelare le informazioni private delle persone o per distruggere i loro dati", afferma. Dice che le aziende di intelligenza artificiale devono iniziare ad avvertire gli utenti dei rischi per la sicurezza e la privacy e fare di più per affrontarli.

OpenAI sembra essere sempre più attento ai rischi per la sicurezza. Il presidente e co-fondatore di OpenAI **Greg Brockman**

ha twittato il mese scorso che la società sta "considerando l'avvio di un programma di taglie" per gli hacker che segnalano punti deboli nei suoi sistemi di intelligenza artificiale, riconoscendo che la posta in gioco "aumenterà *molto* nel tempo".

Tuttavia, molti dei problemi inerenti ai modelli IA odierni non hanno soluzioni facili. Un problema fastidioso è come rendere identificabili i contenuti generati dall'intelligenza artificiale. Alcuni ricercatori stanno lavorando al "watermarking", creando una firma digitale impercettibile nell'output dell'IA. Altri stanno cercando di escogitare mezzi per rilevare modelli che solo l'IA produce. Tuttavia, una recente ricerca ha rilevato che gli strumenti che riformulano leggermente il testo prodotto dall'intelligenza artificiale possono minare in modo significativo entrambi gli approcci. Man mano che l'intelligenza artificiale inizia a sembrare più umana, affermano gli autori,

Il so output diventerà solo più difficile da rilevare. Altre protezioni sfuggenti includono quelle per impedire ai sistemi di generare immagini violente o pornografiche. Tramèr afferma che la maggior parte dei ricercatori sta semplicemente applicando filtri a posteriori, insegnando all'intelligenza artificiale a evitare risultati "cattivi". Ritiene che questi problemi debbano essere risolti prima della formazione, a livello di dati. "Dobbiamo trovare modi migliori per curare i set di addestramento di questi modelli generativi per rimuovere del tutto i dati sensibili", afferma.

La pausa stessa sembra improbabile che accada. Il CEO di OpenAI **Sam Altman**

non ha firmato la lettera, dicendo al Wall Street Journal che l'azienda ha sempre preso sul serio la sicurezza e collabora regolarmente con l'industria sugli standard di sicurezza. Il co-fondatore di Microsoft Bill Gates ha dichiarato a Reuters che la pausa proposta non "risolverà le sfide" future.

Osborne ritiene che i governi dovranno intervenire. "*Non possiamo fare affidamento sui giganti della tecnologia per autoregolamentarsi",* afferma. L'amministrazione Biden ha proposto una "Carta dei diritti" dell'IA progettata per aiutare le aziende a sviluppare sistemi di intelligenza artificiale sicuri che proteggano i diritti dei cittadini statunitensi, ma i principi sono volontari e non vincolanti. La legge sull'IA dell'Unione europea, che dovrebbe entrare in vigore quest'anno, applicherà diversi livelli di regolamentazione a seconda del livello di rischio. Ad esempio, i sistemi di polizia che mirano a prevedere i crimini individuali sono considerati inaccettabilmente rischiosi
E sono pertanto vietati

Wachter afferma che una pausa di 6 mesi sembra arbitraria e che è cauta nel vietare la ricerca. Invece, **"dobbiamo tornare indietro e pensare a una ricerca responsabile e incorporare quel tipo di pensiero molto presto",** afferma. Come parte di questo, afferma che le aziende dovrebbero invitare esperti indipendenti per hackerare e sottoporre a stress test i loro sistemi prima di implementarli.
Nota che le persone dietro la lettera sono fortemente immerse nel mondo della tecnologia, che secondo lei offre loro una prospettiva ristretta sui potenziali rischi. "*Hai davvero bisogno di parlare con avvocati, persone che si occupano di etica, persone che capiscono l'economia e la politica", dice. "La cosa più importante è che queste domande non vengano risolte solo tra i tecnici".*

A chi legge
Questo Report è costruito sui dati di Laurie Clarke giornalista free lance

Allegato 1

Metti in pausa gli esperimenti di intelligenza artificiale gigante: una lettera aperta

Chiediamo a tutti i laboratori di intelligenza artificiale di sospendere immediatamente per almeno 6 mesi l'addestramento di sistemi di intelligenza artificiale più potenti di GPT-4.

22 marzo 2023

I sistemi di intelligenza artificiale con intelligenza competitiva umana possono comportare gravi rischi per la società e l'umanità, come dimostrato da ricerche approfondite e riconosciuto dai migliori laboratori di intelligenza artificiale.Come affermato nei ***Principi Asilomar AI***
ampiamente approvati , *l'IA avanzata potrebbe rappresentare un profondo cambiamento nella storia della vita sulla Terra e dovrebbe essere pianificata e gestita con cure e risorse adeguate* .
Sfortunatamente, questo livello di pianificazione e gestione non sta accadendo, anche se negli ultimi mesi i laboratori di intelligenza artificiale sono stati bloccati in una corsa fuori controllo per sviluppare e implementare menti digitali sempre più potenti che nessuno, nemmeno i loro creatori, può capire, prevedere o controllare in modo affidabile.

I sistemi di intelligenza artificiale contemporanei stanno ora diventando competitivi per l'uomo in compiti generali, e dobbiamo chiederci: dovremmo *lasciare* che le macchine inondino i nostri canali di informazioni con propaganda e falsità? *Dovremmo* automatizzare tutti i lavori, compresi quelli soddisfacenti?
Dovremmo sviluppare menti non umane che alla fine potrebbero essere più numerose, superate in astuzia, obsolete e sostituirci?
Dovremmo rischiare di perdere il controllo della nostra civiltà? Tali decisioni non devono essere delegate a leader tecnologici non eletti.

Potenti sistemi di intelligenza artificiale dovrebbero essere sviluppati solo quando saremo certi che i loro effetti saranno positivi e i loro rischi saranno gestibili.

Questa fiducia deve essere ben giustificata e aumentare con l'entità degli effetti potenziali di un sistema. La recente dichiarazione di **Open AI sull'intelligenza artificiale generale** afferma che *"Ad un certo punto, potrebbe essere importante ottenere una revisione indipendente prima di iniziare ad addestrare i sistemi futuri e, per gli sforzi più avanzati, concordare di limitare il tasso di crescita del calcolo utilizzato per creare nuovi Modelli."*
Siamo d'accordo. Quel punto è adesso.

Pertanto, ***invitiamo tutti i laboratori di intelligenza artificiale a sospendere immediatamente per almeno 6 mesi l'addestramento di sistemi di intelligenza artificiale più potenti di GPT-4*** .

Questa pausa dovrebbe essere pubblica e verificabile e includere tutti gli attori chiave. Se una tale pausa non può essere attuata rapidamente, i governi dovrebbero intervenire e istituire una moratoria.
I laboratori di intelligenza artificiale e gli esperti indipendenti dovrebbero sfruttare questa pausa per sviluppare e implementare congiuntamente una serie di protocolli di sicurezza condivisi per la progettazione e lo sviluppo avanzati di intelligenza artificiale, rigorosamente verificati e supervisionati da esperti esterni indipendenti. Questi protocolli dovrebbero garantire che i sistemi che vi aderiscono siano sicuri oltre ogni ragionevole dubbioCiò *non* significa una pausa nello sviluppo dell'IA in generale, ma semplicemente un passo indietro dalla pericolosa corsa a modelli black-box sempre più grandi e imprevedibili con capacità emergenti.

La ricerca e lo sviluppo dell'IA dovrebbero essere riorientati per rendere i sistemi potenti e all'avanguardia di oggi più accurati, sicuri, interpretabili, trasparenti, robusti, allineati, affidabili e leali. Parallelamente, gli sviluppatori di intelligenza artificiale devono collaborare con i responsabili politici per accelerare drasticamente lo sviluppo di solidi sistemi di governance dell'IA. Questi dovrebbero come minimo includere: autorità di regolamentazione nuove e capaci dedicate all'IA; supervisione e tracciamento di sistemi di intelligenza artificiale altamente capaci e ampi pool di capacità computazionali; sistemi di provenienza e watermarking per aiutare a distinguere il reale dal sintetico e per tenere traccia delle fughe di modelli; un solido ecosistema di audit e certificazione; responsabilità per danni causati dall'IA; solidi finanziamenti pubblici per la ricerca tecnica sulla sicurezza dell'IA; e istituzioni dotate di risorse adeguate per far fronte alle drammatiche perturbazioni economiche e politiche (soprattutto per la democrazia) che l'IA causerà.

L'umanità può godere di un futuro fiorente con l'IA. Dopo essere riusciti a creare potenti sistemi di intelligenza artificiale, ora possiamo goderci un'*"estate di intelligenza artificiale*" in cui raccogliamo i frutti, progettiamo questi sistemi per il chiaro vantaggio di tutti e diamo alla società la possibilità di adattarsi. La società ha messo in pausa altre tecnologie con effetti potenzialmente catastrofici sulla società. Possiamo farlo qui. Godiamoci una lunga estate AI, non precipitiamoci impreparati in una caduta.

Un anno fa… Baedeker/Replay del 16 aprile 2022
Per far sopravvivere la ricerca in Ucraina

Mentre un esercito di "belve sanguinarie" della Federazione Russa descritti come "soldati" (la gloriosa Armata Rossa si vergognerebbe di questi miserabili assassini in divisa) sta invadendo e distruggendo i confini dell' Ucraina uccidendo donne e bambini, le istituzioni scientifiche di tutto il mondo stanno rispondendo a queste atrocità offrendo sostegno a centinaia di ricercatori ucraini che sono fuggiti dal loro paese dilaniato dalla guerra. Pochi (nessuno) aiuti sono purtroppo disponibili per quei ricercatori che, per

scelta o necessità, sono rimasti in Ucraina. Molti hanno visto i loro stipendi ridotti o interrotti mentre i loro Istituti di appartenenza dirottavano le risorse per la ricerca alla difesa della nazione.
Non avevano nessuno a cui rivolgersi, fino a quando **il Wolfgang Pauli Institute (WPI),** un Centro austriaco di Eccellenza Internazionale in matematica e fisica ha iniziato a lanciare una modesta ma significativa iniziativa di assistenza. Il programma WPI è un programma che mette a disposizione piccoli aiuti economici per ricercatori bisognosi: 2000 euro e per gli studenti un dottorato di ricerca di 1500 euro senza vincoli , per sostenere le loro vite e il loro lavoro. Il budget complessivo iniziale del WPI è di soli 50.000 euro, ma sta crescendo di giorno in giorno aggregando e promuovendo iniziative di sostegno simili per provare a sostenere la scienza di base e applicata in Ucraina. Questa "piccola grande"iniziativa potrebbe aiutare a fermare la "fuga di cervelli" che molti temono lascerà l'Ucraina impoverita per molto tempo al termine di questa guerra Intanto si vanno organizzando iniziative di salvataggio più corpose per aiutare "quello che resterà" dell'Ucraina dopo la fine della guerra. Il governo degli Stati Uniti ha iniziato a definire gli ampi contorni di una sorta di Piano Marshall , che ricorda il massiccio sforzo degli Stati Uniti per ricostruire l'Europa dopo la seconda guerra mondiale. La scienza non sarà esclusa, afferma **Kenneth Myers**, presidente di CRDF Global, un'organizzazione no-profit statunitense che gestisce programmi di assistenza scientifica nelle nazioni dell'ex Unione Sovietica. L'idea non è solo di riportare la ricerca ai livelli precedenti l'invasione, ma di investire in aree in cui I ricercatori ucrainici possono eccellere e proporsi come leader . Per ora, l'obiettivo principale è la sopravvivenza fisica. Gli scienziati abbandonati nelle aree di conflitto sono quelli più a rischio, con le truppe russe che si stanno ammassando minacciose per un nuovo assalto all'Ucraina orientale.
L' Accademia francese delle scienze sta già organizzando una rete di supporto per i ricercatori nelle roccaforti della scienza orientale come Kharkiv e Sumy. Intato il progetto WPI sta "tenendo a galla" 17 scienziati in quelle città martoriate, permettendo così di raccontare le storie di alcuni ricercatori come quella di Ihor Shpetnyy, un fisico dello Stato di Sumy, rimasto indietro quando gli altri residenti sono stati evacuati che racconta: "Il mio compito era trovare e acquistare medicine per i miei genitori". Il trattamento di sua madre per l'ipertensione polmonare costa 500 euro al mese. Con il suo stipendio dimezzato a 220 euro al mese, " i soldi del WPI sono stati un vero toccasana." Una borsa di studio WPI è stata una boccata di ossigeno anche per Mykhaylo Mykhaylov. Alla fine dello scorso anno, questo ricercatore noto a livello internazionale per i suoi studi sullo stato solido della materia aveva appena finito di rinnovare il suo laboratorio al Verkin Institute for Low Temperature Physics and Engineering a Kharkiv. Appena 2 mesi dopo, con Kharkiv sotto assedio, Mykhaylov scortò la moglie e il figlio di 10 anni in un viaggio di 6 giorni a Uzhhorod, al confine con la Slovacchia, dove ha dovuto dire addio i suoi cari perché alla maggior parte degli uomini di età inferiore ai 60 anni è vietato lasciare l'Ucraina in attesa di essere chiamati a combattere; racconta che "sono andato da un notaio e ho fatto testamento. Poi ho dato a mia moglie tutti i nostri risparmi". Attualmente la famiglia di Mykhaylov ha raggiunto un rifugio nei Paesi Bassi e lui è tornato a casa in un appartamento distrutto dai bombardamenti con 300 euro in tasca. "Stavo pensando di abbandonare la ricerca", dice. Una borsa di studio WPI gli ha dato i mezzi per affittare un appartamento fuori Kharkiv e restare fedele alla scienza. I recenti bombardamenti hanno fatto esplodere le finestre del suo Istituto; oggi non sa ancora che fine ha fatto il suo laboratorio perché è troppo rischioso da visitare. Per ora, sta scrivendo un progetto di ricerca ed è desideroso di tornare al lavoro sperimentale, una volta ripristinata la normalità. Altre organizzazioni si stanno unendo ai soccorsi dei ricercatori. . Quest'anno la Fondazione per la scienza polacca finanzierà almeno sei collaborazioni di scienze sociali tra scienziati in Polonia e Ucraina; ogni gruppo di ricerca riceverà 58.000 euro in un anno per stipendi e spese di ricerca.

La Fondazione Krzysztof Skubiszewski, sempre in Polonia, prevede di erogare almeno 240.000 euro a studiosi in Ucraina e rifugiati in Polonia, la maggior parte dei quali andrà a quelli in Ucraina. Prima finirà la guerra, prima l'Ucraina potrà passare dalla sopravvivenza scientifica alla rinascita. "Il lavoro diplomatico sta già avvenendo", afferma Tom Callahan, vicepresidente per la strategia e l'innovazione di CRDF Global che ha in progetto di ripristinare una sorgente di neutroni presso l'Istituto di fisica e tecnologia di Kharkiv che è stata danneggiata in un attacco missilistico il mese scorso. **Bill Watterson** il vignettista satirico autore della striscia fumetti calvin &Hobbes diceva che Non facciamo abbastanza ricerca scientifica per trovare una cura per cretini, adesso anche per i criminali assassini come Vladimir Vladimirovič Putin.

A chi legge Le storie dei ricercatori ucrainici sono riprese dalle corrispondenze di Richard Stone redattore scientifico senior presso i Tangled Bank Studios dell'Howard Hughes Medical Institute a Chevy Chase (Maryland). da domani …:

GLOSSARIO PANDEMICO ESSENZIALE
(vai all'originale)

Un anno fa… Baedeker/Replay del 16 aprile 2021

Perché il vaccino J&J è in "pausa negli Stati Uniti, e cosa significa una "pausa"?