Elon Musk e altri 1.000 leader della Silicon Valley: «**Sospendere lo sviluppo dell'intelligenza artificiale**»

In una lettera aperta gli oltre 1.000 firmatari spiegano come «*negli ultimi mesi c'è stata una corsa fuori controllo dei laboratori per l'intelligenza artificiale a sviluppare potenti menti digitali che nessuno, neanche i creatori, possono capire, prevedere e controllare*».
Ora una pausa di almeno sei mesi - 29 marzo 2023

Elon Musk e oltre 1.000 fra ricercatori e manager **chiedono una "pausa" di sei mesi** nello sviluppo dei sistemi di intelligenza artificiale avanzati come ChatGPT - l'intelligenza artificiale generativa finanziata da Microsoft - per fermare quella che definiscono una «pericolosa» corsa agli armamenti.

In una lettera aperta pubblicata da **Future of Life Institute** e riportata dal **Financial Times** (qui sotto la traduzione in italiano) che rilancia gli interrogativi etici sull'intelligenza artificiale, gli oltre 1.000 firmatari spiegano come «negli ultimi mesi c'è stata una corsa fuori controllo dei laboratori per l'intelligenza artificiale a sviluppare e dispiegare potenti menti digitali che nessuno, neanche i creatori, possono capire, prevedere e controllare».

Fonte: Il Sole 24ore.
Articolo del 29 marzo 2023.
La traduzione in italiano della lettera aperta che rilancia gli interrogativi etici sull'intelligenza artificiale pubblicata da Future of Life Institute, con oltre 1.000 firmatari

Sospendere gli esperimenti avanzati di intelligenza artificiale: lettera aperta

I sistemi di IA dotati di un'intelligenza competitiva con quella umana possono comportare rischi profondi per la società e l'umanità, come dimostrato da ricerche approfondite [1] e riconosciuto dai migliori laboratori di IA [2].

Come affermato nei Principi di Asilomar per l'intelligenza artificale, ampiamente approvati, l'IA avanzata potrebbe *rappresentare un cambiamento profondo nella storia della vita sulla Terra e dovrebbe essere pianificata e gestita con cura e risorse adeguate*.

Sfortunatamente, questo livello di pianificazione e gestione non sta avvenendo, anche se negli ultimi mesi i laboratori di IA si sono impegnati in una corsa fuori controllo per sviluppare e impiegare menti digitali sempre più potenti che nessuno - nemmeno i loro creatori - è in grado di comprendere, prevedere o controllare in modo affidabile.

I sistemi di intelligenza artificiale contemporanei stanno diventando competitivi con gli esseri umani in compiti generali [3] e dobbiamo chiederci se sia il caso di lasciare che le macchine inondino i nostri canali di informazione: dobbiamo lasciare che le macchine inondino i nostri canali di informazione con propaganda e falsità? Dovremmo automatizzare tutti i lavori, compresi quelli più soddisfacenti? Dovremmo sviluppare menti non umane che alla fine potrebbero superarci di numero, essere più intelligenti e sostituirci? Dobbiamo rischiare di perdere il controllo della nostra civiltà? Queste decisioni non devono essere delegate a leader tecnologici non eletti.

I potenti sistemi di intelligenza artificiale dovrebbero essere sviluppati solo quando saremo sicuri che i loro effetti saranno positivi e i loro rischi gestibili. Questa fiducia deve essere ben giustificata e aumentare con l'entità degli effetti potenziali di un sistema. La recente dichiarazione di OpenAI sull'intelligenza artificiale generale afferma che "a 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 dei calcoli utilizzati per creare nuovi modelli". Siamo d'accordo. Quel punto è ora, lo abbiamo già raggiunto.

Pertanto, chiediamo a tutti i laboratori di IA di sospendere immediatamente per almeno 6 mesi l'addestramento di sistemi di IA più potenti del GPT-4. Questa pausa dovrebbe essere pubblica e verificabile. Questa pausa deve essere pubblica e verificabile e deve includere tutti gli attori chiave. Se tale pausa non potesse essere attuata rapidamente, i governi dovrebbero intervenire e istituire una moratoria.

I laboratori di IA e gli esperti indipendenti dovrebbero utilizzare questa pausa per sviluppare e implementare congiuntamente una serie di protocolli di sicurezza condivisi per la progettazione e lo sviluppo di IA avanzate, rigorosamente controllati e supervisionati da esperti esterni indipendenti.

Questi protocolli dovrebbero garantire che i sistemi che vi aderiscono siano sicuri al di là di ogni ragionevole dubbio.[4]

Ciò non significa una pausa nello sviluppo dell'IA in generale, ma solo un passo indietro rispetto alla pericolosa corsa verso modelli black-box sempre più grandi e imprevedibili con capacità emergenti. La ricerca e lo sviluppo dell'IA dovrebbero concentrarsi sul rendere i potenti sistemi all'avanguardia di oggi più accurati, sicuri, interpretabili, trasparenti, robusti, allineati, affidabili e leali.

Parallelamente, gli sviluppatori di IA devono lavorare con i 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; sorveglianza e monitoraggio di sistemi di IA altamente capaci e di grandi bacini di capacità computazionale; sistemi di provenienza e watermarking per aiutare a distinguere i modelli reali da quelli sintetici e per tracciare le fughe di notizie; un robusto ecosistema di auditing e certificazione; responsabilità per i danni causati dall'IA; solidi finanziamenti pubblici per la ricerca tecnica sulla sicurezza dell'IA; istituzioni ben finanziate per affrontare i drammatici sconvolgimenti economici e politici (soprattutto per la democrazia) che l'IA causerà.

L'umanità può godere di un futuro fiorente con l'IA. Essendo riusciti a creare potenti sistemi di IA, possiamo ora goderci una "estate dell'IA" in cui raccogliere i frutti, progettare questi sistemi per il chiaro beneficio di tutti e dare alla società la possibilità di adattarsi. La società ha messo in pausa altre tecnologie con effetti potenzialmente catastrofici per la società [5] e possiamo farlo anche in questo caso. Godiamoci una lunga estate dell'IA, non precipitiamoci impreparati in un autunno.
Traduzione realizzata con l'ausilio di DeepL

Note e referenze
[1] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610-623). Bostrom, N. (2016). Superintelligence. Oxford University Press.Bucknall, B. S., & Dori-Hacohen, S. (2022, July). Current and near-term AI as a potential existential risk factor. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (pp. 119-129).Carlsmith, J. (2022). Is Power-Seeking AI an Existential Risk?. arXiv preprint arXiv:2206.13353.Christian, B. (2020). The Alignment Problem: Machine Learning and human values. Norton & Company.Cohen, M. et al. (2022). Advanced Artificial Agents Intervene in the Provision of Reward. AI Magazine, 43(3) (pp. 282-293).Eloundou, T., et al. (2023). GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models.Hendrycks, D., & Mazeika, M. (2022). X-risk Analysis for AI Research. arXiv preprint arXiv:2206.05862.Ngo, R. (2022). The alignment problem from a deep learning perspective. arXiv preprint arXiv:2209.00626.Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. Viking.Tegmark, M. (2017). Life 3.0: Being Human in the Age of Artificial Intelligence. Knopf.Weidinger, L. et al (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.

[2] Ordonez, V. et al. (2023, March 16). OpenAI CEO Sam Altman says AI will reshape society, acknowledges risks: 'A little bit scared of this'. ABC News.Perrigo, B. (2023, January 12). DeepMind CEO Demis Hassabis Urges Caution on AI. Time.

[3] Bubeck, S. et al. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712.OpenAI (2023). GPT-4 Technical Report. arXiv:2303.08774.

[4] Ample legal precedent exists – for example, the widely adopted OECD AI Principles require that AI systems "function appropriately and do not pose unreasonable safety risk".

[5] Examples include human cloning, human germline modification, gain-of-function research, and eugenics.