6. Gennaio
Un'intelligenza artificiale "sovrumana" può causare la nostra estinzione

L'estinzione è la regola.
È la sopravvivenza a costituire l'eccezione.
Carl Sagan

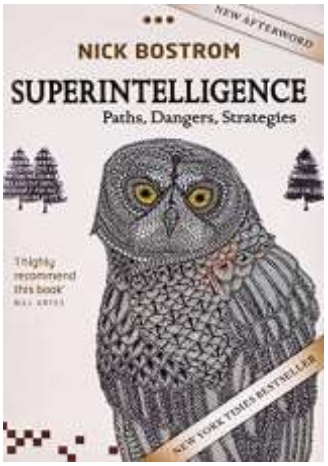
C'è soltanto una guerra che può permettersi il genere umano:
la guerra contro la propria estinzione.
Isaac Asimov

Molti ricercatori nel campo dell'intelligenza artificiale ritengono che il possibile sviluppo futuro di **un'intelligenza artificiale sovrumana** abbia una possibilità non banale di causare la nostra estinzione, ma vi è anche un diffuso disaccordo e incertezza su tali rischi.

Ciò emerge da un sondaggio condotto su 2700 ricercatori nel campo dell'intelligenza artificiale che hanno recentemente pubblicato lavori in sei delle principali conferenze sull'intelligenza artificiale.

L'indagine ha chiesto ai ricercatori di condividere le loro opinioni sulle possibili tempistiche per i futuri traguardi tecnologici dell'intelligenza artificiale, nonché sulle conseguenze sociali positive o negative di tali risultati. Quasi il 58% ha affermato di ritenere che esista una probabilità del 5% di estinzione umana o di altri risultati estremamente negativi legati all'intelligenza artificiale.

L'Intelligenza Artificiale è indubbiamente una delle più grandi promesse dell'umanità; grazie ai suoi sviluppi, attuali e futuri, saremo probabilmente in grado di fare cose che oggi sarebbero impensabili, vivremo meglio, e magari più a lungo e più felici.



Nick Bostrom è un filosofo eclettico nato in Svezia con un background in fisica teorica, neuroscienze computazionali, logica e intelligenza artificiale, oltre che in filosofia, professore all'Università di Oxford, dove dirige il Future of Humanity Institute come direttore fondatore.

Nel 2014 con il libro **Superintelligence: Paths, Dangers, Strategies (2014)**, un bestseller del New York Times che ha contribuito a innescare una conversazione globale sull'intelligenza artificiale che attraversa filosofia, scienza, etica e tecnologia, ha illuminato i collegamenti tra le nostre azioni attuali e i risultati globali a lungo termine, gettando così una nuova luce sulla condizione umana.

Nel gennaio 2015 è stato cofirmatario, assieme tra gli altri a *Stephen Hawking*, di una celebre lettera aperta che metteva in guardia sui potenziali pericoli dell'Intelligenza Artificiale.

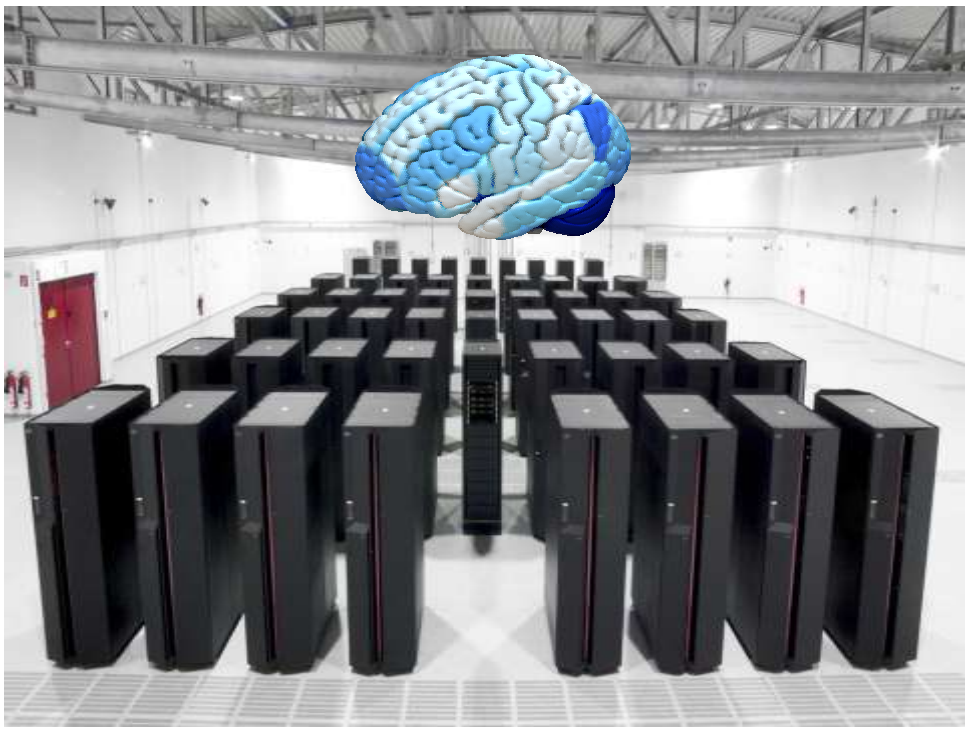
Non ha firmato quell'appello per passatismo, né tantomeno per luddismo, bensì in virtù di un lineare ragionamento filosofico.

Nick Bostrom è stato il primo a vedere i rischi e i pericoli dell'intelligenza artificiale lanciando un allarme che ha avuto un'eco vastissima in tutto il mondo. Siamo proprio certi che riusciremo a governare senza problemi una macchina "superintelligente" dopo che l'avremo costruita?

Se lo scopo dell'attuale ricerca sull'Intelligenza Artificiale è quello di costruire delle macchine fornite di un'intelligenza generale paragonabile a quella umana, ***quanto tempo occorrerà a quelle macchine, una volta costruite, per superare e surclassare le nostre capacità intellettive?***
Secondo Bostrom, "pochissimo".

Una volta raggiunto un livello di intelligenza paragonabile al nostro, alle macchine basterà un piccolo passo per *"decollare"* esponenzialmente, dando origine a superintelligenze che per noi risulteranno rapidamente inarrivabili.
A quel punto le nostre creature potrebbero scapparci di mano, non necessariamente per "*malvagità*", ma anche solo come effetto collaterale della loro attività. Potrebbero arrivare a distruggerci o addirittura a distruggere il mondo intero.

E’ particolarmente “intrigante” il confronto tra
cervello umano e ***cervello informatico***:



i neuroni possono sparare un massimo di centinaia di Hz, la velocità di clock dei computer moderni può raggiungere circa 2 Ghz – ~ dieci milioni di volte più veloce.

i potenziali d’azione viaggiano a qualche centinaio di m/s, la comunicazione ottica può avvenire a 300.000.000 di m/s – circa un milione di volte più veloce.

le dimensioni del cervello e il numero dei neuroni sono limitati dal volume cranico, da vincoli metabolici e da altri fattori, i supercomputer possono avere le dimensioni di un magazzino.

Inoltre i ***sistemi artificiali*** non devono nemmeno soffrire dei vincoli del ***cervello biologico*** rispetto alla memoria e all’affidabilità dei componenti, alla larghezza di banda di input/output, all’affaticamento dopo ore, al degrado dopo decenni, all’immagazzinamento dei propri meccanismi di riparazione e progetti al suo interno, e così via.

I sistemi artificiali possono anche essere modificati e duplicati molto più facilmente dei cervelli, e le informazioni possono essere trasferite più facilmente tra loro.

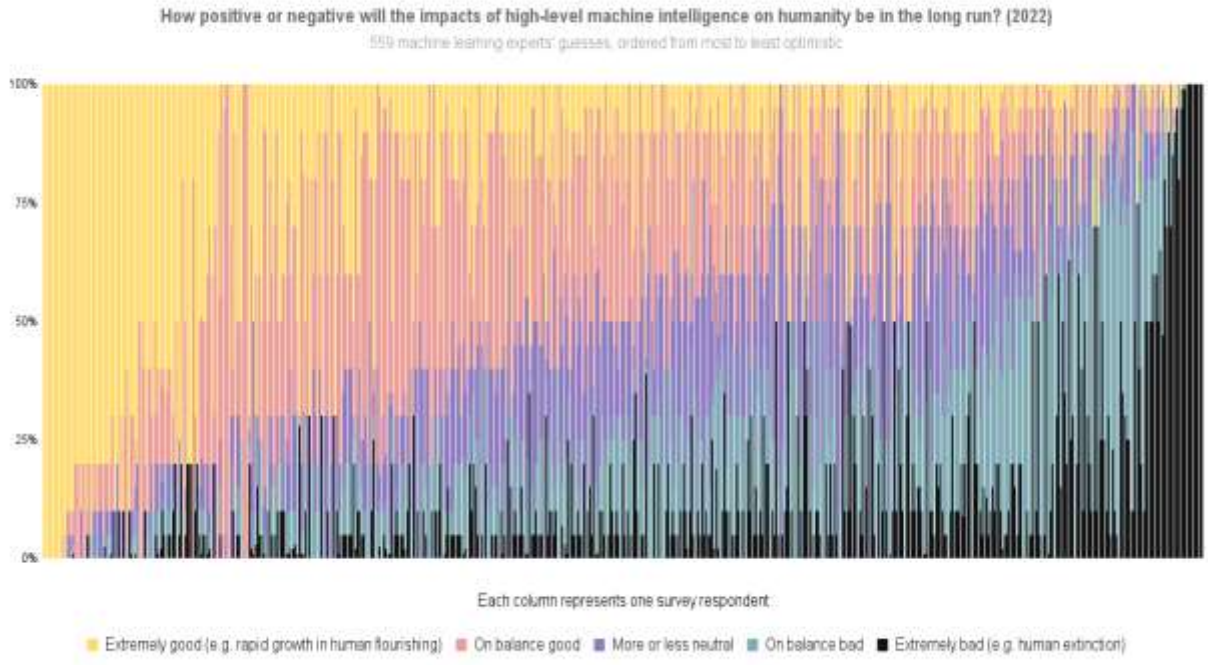
Per questo - sostiene Bostrom - dobbiamo preoccuparcene ora. Per non rinunciare ai benefici che l'Intelligenza Artificiale potrà apportare, è necessario che la ricerca tecnologica si ponga adesso le domande che questo libro pone con enorme chiarezza e chiaroveggenza.

Katja Grace del *Machine Intelligence Research Institute in California*
è convinta che la maggior parte dei ricercatori sull'intelligenza artificiale non trova del tutto implausibile che l'intelligenza artificiale avanzata distrugga l'umanità anzi pensa questa convinzione generale in un rischio non minuscolo sia molto più significativa dell'esatta percentuale di rischio."

Attraverso una serie di sondaggi ha testato l'opinione degli esperti nei confronti dell'intelligenza artificiali in una scala coloroimetrica riassuntiva

La domanda Quanto positivo o negativo ti aspetti che questo avrà un impatto complessivo sull'umanità, nel lungo periodo? Rispondi indicando la probabilità che ritieni che si verifichino i seguenti tipi di impatto, con una somma delle probabilità pari al 100%:

Estremamente buono (ad esempio, rapida crescita della prosperità umana)
Tutto sommato buono
Più o meno neutrale
Tutto sommato pessimo
Estremamente grave (ad esempio, estinzione umana)



Ma non c’è bisogno di farsi prendere dal panico per ora, afferma

Emile Torres della C*ase Western Reserve University in Ohio*. Molti esperti di intelligenza artificiale “non hanno una buona esperienza” nella previsione dei futuri sviluppi dell’intelligenza artificiale, dicono. Grace e i suoi colleghi hanno riconosciuto che i ricercatori nel campo dell’intelligenza artificiale non sono esperti nel prevedere la traiettoria futura dell’intelligenza artificiale, ma hanno dimostrato che una una versione del 2016 del loro sondaggio ha fatto un “lavoro abbastanza buono nel prevedere” i traguardi dell’intelligenza artificiale.

Rispetto alle risposte di una versione del 2022 dello stesso sondaggio, molti ricercatori sull’intelligenza artificiale hanno previsto che l’intelligenza artificiale raggiungerà determinati traguardi prima di quanto previsto in precedenza.

Ciò coincide con il debutto di ChatGPT nel novembre 2022 e con la fretta della Silicon Valley di implementare ampiamente servizi di Silicon Valley di implementare servizi di chatbot AI simili basati su ampi modelli linguistici.
I
ricercatori intervistati hanno previsto che entro il prossimo decennio, i sistemi di intelligenza artificiale avranno una probabilità pari o superiore al 50% di affrontare con successo la maggior parte dei 39 compiti campione, tra cui scrivere nuove canzoni indistinguibili da un pezzo di Taylor Swift o codificare da zero un intero sito di elaborazione dei pagamenti.

Al possibile sviluppo di un’intelligenza artificiale in grado di superare gli esseri umani in ogni compito è stata data una probabilità del 50% che si verifichi entro il 2047, mentre alla possibilità che tutti i lavori umani diventino completamente automatizzabili è stata data una probabilità del 50% che si verifichi entro il 2116.

Queste stime sono di 13 anni e 48 anni prima rispetto a quelli indicati nell'indagine dello scorso anno.

Ma anche le crescenti aspettative riguardo allo sviluppo dell’intelligenza artificiale potrebbero fallire, afferma Torres. “Molte di queste scoperte sono piuttosto imprevedibili. Ed è del tutto possibile che il campo dell’intelligenza artificiale attraversi un altro inverno”, affermano, riferendosi all’esaurimento dei finanziamenti e dell’interesse aziendale per l’intelligenza artificiale durante gli anni 70/80

Ci sono anche preoccupazioni più immediate senza rischi di intelligenza artificiale sovrumana. La grande maggioranza dei ricercatori sull'intelligenza artificiale – il 70% o più – ha descritto gli scenari alimentati dall'intelligenza artificiale che coinvolgono deepfake, manipolazione dell'opinione pubblica, armi ingegnerizzate, controllo autoritario delle popolazioni e peggioramento della disuguaglianza economica come motivo di sostanziale o estrema preoccupazione. Torres ha anche sottolineato i pericoli che l'intelligenza artificiale può contribuire alla disinformazione su questioni esistenziali come il cambiamento climatico o il peggioramento della governance democratica.
Vedremo cosa succederà nelle elezioni europee e in quelle presidenziali americane del 2024 del 2024."
Una strada porta alla disperazione e allo sconforto più assoluto. L'altra alla totale estinzione. Preghiamo il cielo che ci dia la saggezza di fare la scelta esatta.
(Woody Allen)

PS
Consolazione personale: Se un giorno gli uomini dovessero estinguersi, il pensiero si troverebbe un altro luogo d'incubazione.

Elenco di fonti che si oppongono al rischio esistenziale derivante dall'IA

Cegloowski, Maciej. "Superintelligenza: l'idea che divora le persone intelligenti". *Parole inattive* (blog). Accesso effettuato il 9 dicembre 2021. **https://idlewords.com/talks/superintelligence.htm** .

Garfinkel, Ben, e Lempel, Howie. "Quanto siamo sicuri di questa roba dell'intelligenza artificiale?" 80.000 ore. Accesso effettuato il 16 settembre 2020. **https://80000hours.org/podcast/episodes/ben-garfinkel-classic-ai-risk-arguments/** .

Garfinkel, Ben. ***Quanto siamo sicuri di questa roba dell'intelligenza artificiale? (parlare)*** | *EA Global: Londra 2018* , 2019. **https://www.youtube.com/watch?v=E8PGcoLDjVk** . Anche in formato blog: **https://ea.greaterwrong.com/posts/9sBAW3qKppnoG3QPq/ben-garfinkel-how-sure-are-we-about-this-ai-stuff**

LeCun, Yann e Anthony Zador. "Non temere il Terminator." Rete di blog scientifici americani. Accesso effettuato il 9 dicembre 2021. **https://blogs.scientificamerican.com/observations/dont-fear-the-terminator/** .

Yudkowsky, Eliezer e Robin Hanson. "Il dibattito Hanson-Yudkowsky AI-Foom – LessWrong." Accesso il 6 agosto 2022. **https://www.lesswrong.com/tag/the-hanson-yudkowsky-ai-foom-debate** .

Elenco di fonti che sostengono il rischio esistenziale derivante dall'IA

Adamczewski, Tom. "Un cambiamento nelle argomentazioni a favore del rischio legato all'intelligenza artificiale". Credenziali fragili. Accesso effettuato il 20 ottobre 2020. **https://fragile-credences.github.io/prioritising-ai/** .

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman e Dan Mané. "Problemi concreti nella sicurezza dell'intelligenza artificiale". *ArXiv:1606.06565 [Cs]* , 25 luglio 2016. **http://arxiv.org/abs/1606.06565** .

Bensinger, Rob, Eliezer Yudkowsky, Richard Ngo, So8res, Holden Karnofsky, Ajeya Cotra, Carl Shulman e Rohin Shah. "Conversazioni MIRI 2021 – LessWrong." Accesso effettuato il 6 agosto 2022. **https://www.lesswrong.com/s/n945eovrA3oDueqtq** .

Bostrom, N., *Superintelligenza* , Oxford University Press, 2014.
Carlsmith, Giuseppe. "L'intelligenza artificiale alla ricerca del potere è un rischio esistenziale? [Bozza]." Progetto Open Philanthropy, aprile 2021. **https://docs.google.com/document/d/1smaI1lagHHcrhoi6ohdq3TYIZv0eNWWZMPEy8C8byYg/edit?usp=embed_facebook** .

Cristiano, Brian. *Il problema dell'allineamento: apprendimento automatico e valori umani* . WW Norton & Company, 2021.
Cristiano, Paolo. "Che aspetto ha il fallimento." *Forum di allineamento AI* (blog), 17 marzo 2019. **https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like** .

Dai, Wei. "Commento su come risolvere gli argomenti sull'importanza della sicurezza dell'intelligenza artificiale - LessWrong." Accesso effettuato il 9 dicembre 2021. **https://www.lesswrong.com/posts/JbcWQCxKWn3y49bNB/disentangling-arguments-for-the-importance-of-ai-safety** .

Garfinkel, Ben, Miles Brundage, Daniel Filan, Carrick Flynn, Jelena Luketina, Michael Page, Anders Sandberg, Andrew Snyder-Beattie e Max Tegmark. "Sull'impossibilità delle macchine sovradimensionate". *ArXiv:1703.10987 [Fisica]* , 31 marzo 2017. **http://arxiv.org/abs/1703.10987** .

Hubinger, Evan, Chris van Merwijk, Vladimir Mikulik, Joar Skalse e Scott Garrabrant. "Rischi derivanti dall'ottimizzazione appresa nei sistemi avanzati di machine learning", 5 giugno 2019. **https://arxiv.org/abs/1906.01820v3** .

No, Richard. "Thinking Complete: districare gli argomenti sull'importanza della sicurezza dell'intelligenza artificiale". *Thinking Complete* (blog), 21 gennaio 2019. **http://thinkingcomplete.blogspot.com/2019/01/disentangling-arguments-for-importance.html** . (Anche **LessWrong** e **Alignment Forum** , con i relativi thread di commenti.)

No, Richard. "Sicurezza AGI dai principi primi", 28 settembre 2020. **https://www.lesswrong.com/s/mzgtmmTKKn5MuCzFJ** .

Ord, Toby. ***Il precipizio: rischio esistenziale e futuro dell'umanità*** **.** Edizione illustrata. New York: Hachette Books, 2020.
Piper, Kelsey. "Il caso di prendere sul serio l'intelligenza artificiale come minaccia per l'umanità." Vox, 21 dicembre 2018. **https://www.vox.com/future-perfect/2018/12/21/18126576/ai-artificial-intelligence-machine-learning-safety-alignment** .

Russel, Stuart. ***Compatibile con l'uomo: intelligenza artificiale e problema del controllo*** **.** Vichingo, 2019.
Turner, Alexander Matt, Logan Smith, Rohin Shah, Andrew Critch e Prasad Tadepalli. "Le politiche ottimali tendono a ricercare il potere". *ArXiv:1912.01683[Cs]* , 3 dicembre 2021. **http://arxiv.org/abs/1912.01683** .

Yudkowsky, Eliezer. "L'intelligenza artificiale come fattore positivo e negativo nel rischio globale". In *Global Catastrophy Risks* , a cura di Nick Bostrom e Milan M. Ćirković, 46. New York, nd **https://intelligence.org/files/AIPosNegFactor.pdf** .

Yudkowsky, Eliezer, Rob Bensinger e So8res. "Discussione sull'allineamento MIRI 2022 - LessWrong." Accesso effettuato il 6 agosto 2022. **https://www.lesswrong.com/s/v55BhXbpJuaExkpcD** .

Yudkowsky, Eliezer e Robin Hanson. "Il dibattito Hanson-Yudkowsky AI-Foom – LessWrong." Accesso il 6 agosto 2022. **https://www.lesswrong.com/tag/the-hanson-yudkowsky-ai-foom-debate** .

Infezioni da covid-19 grave correlano ad un aumento della schizofrenia

Lo schizofrenico è un uomo senza speranza.
Ronald Laing

Le persone con gravi infezioni da covid-19 hanno maggiori probabilità, ***quattro volte di più***, possibilità di ricevere una ***diagnosi di schizofrenia o altri disturbi psicotici*** rispetto a coloro che non sono stati infettati dal virus, suggerendo che il covid-19 aumenta il rischio di schizofrenia.
La schizofrenia è una grave condizione mentale caratterizzata da ***allucinazioni, deliri e altri disturbi cognitivi.*** Non è chiaro cosa lo causi, anche se ricerche precedenti hanno suggerito che potrebbe essere innescato da virus, come l'influenza o addirittura il covid-19 .

Il team di **Wanhong Zheng** della *West Virgina University* hanno analizzato le diagnosi di schizofrenia e condizioni simili in persone di età compresa *tra 17 e 70 anni* che erano state infettate da covid-19.

Sono stati raccolti dati su più di ***650.000*** persone dalla **National COVID Cohort Collaborative** degli Stati Uniti.

Circa ***219.000*** *partecipanti avevano infezioni da covid-19 moderate, gravi o fatali* e circa ***213.000*** *erano risultati negativi al virus.*
I ***restanti partecipanti*** avevano la sindrome da *distress respiratorio acuto (ARDS),* una condizione polmonare pericolosa per la vita non correlata al covid.
Nessuno aveva una storia di schizofrenia*, disturbo bipolare, depressione, disturbi della personalità o traumi.*

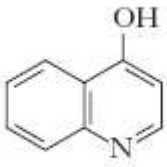
Il team ha esaminato quanti partecipanti hanno poi ricevuto la diagnosi di schizofrenia, disturbo psicotico acuto o condizioni correlate.
Hanno scoperto che tre settimane dopo l'infezione, a ***2573*** *persone è stata diagnosticata una condizione psicotica,* circa la metà delle quali aveva contratto il covid-19.

Le persone che avevano il ***covid-19 avevano circa 4,6 volte più probabilità*** di ricevere una diagnosi di condizione psicotica rispetto a quelle che erano risultate negative al virus.
Quelli con ***ARDS avevano circa il 25% in meno di probabilità*** di ricevere una diagnosi di queste condizioni rispetto a quelli risultati negativi.

Circa tre mesi dopo l'infezione, le persone affette da covid-19 avevano ancora il ***70% in più di probabilità di ricevere una diagnosi di disturbo psicotico rispetto alle persone risultate negative al test o affette da ARDS.***

Sophie Erhardt del *Karolinska Institute* ritiene che questo sia perfettamente in linea con quanto è stato ipotizzato, cioè che il covid-19 aumenta il rischio di psicosi

Un'idea del perché ciò potrebbe essere è che il covid-19 aumenta l'infiammazione nel cervello, che a sua volta porta a livelli più elevati di ***acido chinurenico***



Precedenti ricerche hanno dimostrato che le persone con schizofrenia e psicosi hanno livelli elevati di ***acido chinurenico*** nel cervello e nel liquido spinale, ed ***Erhardt*** e i suoi colleghi hanno osservato livelli altrettanto elevati in quelli con grave covid-19. Lei e altri sul campo ipotizzano che ***l'acido chinurenico*** sia un fattore scatenante della psicosi.