IA: QUATTRO URGENZE, PRIMA DELL'APOCALISSE

Un certo disagio di fronte all'intelligenza artificiale è ragionevole: non tanto per il timore che simili tecnologie causino la fine della nostra specie, bensì per evitare di trattarla con eccessiva superficialità. (..) i rischi impellenti dell'IA sono altri, relativi al modo in cui tali sistemi creano o inaspriscono condizioni avverse al benessere umano: su tali temi, le responsabilità non sono imputabili alle tecnologie, bensì a esseri umani – talora i produttori dei sistemi di IA, talvolta i loro utenti, spesso entrambi. Non dovremmo quindi permettere a facili millenarismi di distrarci da queste priorità: in attesa che l'IA conquisti il pianeta, ci sono vari fronti su cui è urgente intervenire.

Di Fabio Paglieri per la Rivista Il Mulino

18 GENNAIO 2024

A fine maggio 2023, il Center for AI Safety ha reso pubblico un pronunciamento lapidario (single-sentence statement) sui rischi connessi alle tecnologie di intelligenza artificiale (IA), che tradotto letteralmente suona così: "Mitigare il rischio di un'estinzione causata dall'IA dovrebbe essere una priorità globale, al pari di altri rischi che coinvolgono l'intera società, quali pandemie e guerra nucleare". In questo caso, definire l'affermazione apocalittica non è iperbole o eufemismo, giacché di questo si tratta: si paventa che simili tecnologie possano causare la fine della nostra specie, e si invitano tutti a correre ai ripari.

L'appello ha suscitato notevole eco mediatica, per tre ragioni principali. Innanzitutto, il tono catastrofista: se fonti ragionevolmente attendibili annunciano che la fine del mondo è prossima, la cosa desta interesse. In secondo luogo, il carattere perentorio del messaggio, che si compiaceva di essere lungo un'unica frase: pazienza se questo impediva di spiegare alcunché sul modo in cui l'IA dovrebbe o potrebbe causare l'estinzione del genere umano. Gli estensori dell'appello si sono giustificati in nome di una presunta efficacia comunicativa, volta a superare le differenze di opinioni sui rischi specifici dell'IA, onde produrre un comune grido di allarme che mettesse tutti d'accordo: per un motivo o per l'altro, l'IA è una minaccia, attenzione! Sarà. Eppure il formato Twitter del messaggio lo rende più affine alla propaganda urlata che non alla seria discussione scientifica sui rischi delle nuove tecnologie. Infine, l'iniziativa ha catturato l'attenzione generale anche e soprattutto per l'identità dei primi firmatari: un centinaio di esperti riconosciuti di IA, che includeva non solo scienziati di chiara fama, ma anche proprietari delle aziende leader del settore, inclusa OpenAi (quella di ChatGpt e Dall-E, per intenderci) e Microsoft. L'implicazione appare ovvia: se i massimi esperti di IA ci dicono che siamo tutti minacciati e ci invitano a procedere con attenzione, sarà il caso di dare loro ascolto.

Proprio l'identità dei firmatari dell'appello, tuttavia, suggerisce ulteriori riflessioni: perché non si tratta solo di chi l'IA la conosce, ma anche di chi la sta sviluppando e vendendo a tutto il mondo, ricavandone enormi profitti; l'indice Nasdaq nel 2023 è cresciuto di quasi il 50%, trainato proprio dai successi dell'IA generativa. Da questo punto di vista, l'apocalittico annuncio del Center for AI Safety assomiglia molto al caveat "nuoce gravemente alla salute" che i produttori di tabacco stampano sui pacchetti di sigarette: una tutela da successive contestazioni e cause legali, sicuramente non un sincero invito a non consumare il prodotto. Nel caso dell'IA, poi, è lecito un ulteriore sospetto: che simili profezie di sciagura, apocalittiche ma vaghe, servano a distrarre l'attenzione da più reali e mondani problemi delle tecnologie di IA, al contempo presentando i produttori di tali tecnologie come esperti coscienziosi e preoccupati dal bene comune, anziché biechi speculatori intenti a riempirsi le tasche.

All'indomani dell'appello sui rischi di un'estinzione da overdose tecnologica, un articolo su "The Atlantic" denunciava proprio simili ipocrisie, con un titolo piuttosto esplicito: *AI doomerism is a decoy* – liberamente tradotto: "il fatalismo sull'IA è uno specchietto per le allodole". Nel contributo, si suggeriva che ciò da cui annunci di imminente apocalisse vorrebbero distrarci sono i reali danni prodotti nel recente passato da queste stesse tecnologie. Lo stesso ragionamento si applica ai rischi attuali e futuri nella diffusione di tali tecnologie: rischi veri e documentati, non inventati o presunti, rispetto ai quali appare chiaro che le grandi multinazionali dell'IA non abbiano alcuna intenzione di auto-regolarsi, rendendo dunque necessari interventi culturali, politici e legislativi da parte del potere pubblico.

Per poter discutere serenamente degli effettivi rischi dell'IA, tuttavia, bisogna prima liberarsi dallo spauracchio dell'apocalisse digitale, cinicamente agitato da chi sul mercato dell'IA prospera e intende continuare a farlo. Sia chiaro, nessuno sostiene che non vi siano effettivi motivi di ansia esistenziale, alla luce dei recenti sviluppi dell'IA. Ma non si tratta del vago fantasma dell'estinzione della razza umana, come in un brutto film di fantascienza, bensì di un paio di fattori specifici e interessanti: da un lato, l'opacità radicale dell'IA generativa; dall'altro, la capacità di alcuni software di produrre altri programmi funzionanti.

Il concetto di opacità radicale fa riferimento al fatto che nessuno capisce esattamente come funzionino i sistemi di IA generativa, nemmeno chi li ha realizzati. Affermazione temeraria, che necessita immediata qualificazione: ovviamente, il funzionamento generale di simili sistemi è ben compreso da chi li progetta, altrimenti non sarebbe possibile realizzarli. Tuttavia, il modo specifico in cui essi imparano a risolvere determinati problemi, dopo processi di apprendimento che coinvolgono moli enormi di dati, risulta spesso incomprensibile persino ai loro creatori. Ciò non deve stupire, giacché deriva da due fonti di formidabile complessità: da un lato, le sterminate basi di dati su cui viene addestrato il sistema, irrimediabilmente rumorose e caotiche all'occhio umano; dall'altro, la stessa architettura interna di questi programmi, che ormai eguaglia o supera, per sofisticazione, quella dei cervelli biologici. I cosiddetti Large Language Models (Llms), come il celebre ChatGpt, si chiamano "large" proprio per il numero esorbitante di nodi contenuti negli strati interni delle deep neural networks (reti neurali profonde) da cui sono composti: ad esempio, ChatGpt-3, già nel 2020, aveva circa 175 miliardi di nodi o parametri, più del doppio del numero di neuroni presenti nel cervello umano, che si ferma a 86 miliardi, secondo le stime più attendibili (cfr. F.A. Azevedo *et. al.*, *Equal numbers of neuronal and nonneuronal cells*

make the human brain an isometrically scaled-up primate brain, "Journal of Comparative Neurology", vol. 513, pp. 532-541).

Questi enormi sistemi di calcolo, oltre a precludere una completa analisi del loro operato da parte di soggetti umani, hanno la capacità di produrre output testuali (o grafici, in applicazioni come Dall-E, Midjourney e Stable Diffusion): il che costituisce la ragione dell'utilità pratica e del fascino culturale di tali artefatti. Fra le numerose tipologie di testi che ChatGpt e compagni sono in grado di sfornare, tuttavia, rientrano anche programmi software, cioè righe di codice compilabili in applicazioni funzionanti. Questo già ora sta semplificando la vita di molti programmatori, anche se i risultati non sono esenti da rischi e imperfezioni: ad esempio, i programmi generati da questi sistemi tendono a essere più vulnerabili a intrusioni esterne da parte di malware e virus. Al di là degli inconvenienti tecnici, che probabilmente si risolveranno col tempo, la vera ragione di inquietudine è più profonda: un software che scrive altro software è, infatti, piuttosto simile a un sistema artificiale in grado di riprodursi. Beninteso, per ora questi sistemi non hanno alcuna intenzione di replicarsi motu proprio, essendo stati creati privi di motivazioni esistenziali: ma il fatto che ne possiedano la capacità, per quanto rudimentale, giustifica qualche vaga preoccupazione, tanto più alla luce di quanto poco capiamo il loro funzionamento di dettaglio.

Insomma, un certo disagio al cospetto dell'IA generativa è comprensibile e persino ragionevole, nella misura in cui promuove salutare cautela: non perché si paventi l'asservimento dell'umanità da parte delle nuove tecnologie, bensì per evitare di giocare all'apprendista stregone con eccessiva superficialità (sul punto, un'ottima prospettiva è offerta da Nello Cristianini, *La scorciatoia*, Il Mulino, 2023). Detto questo, i rischi impellenti dell'IA sono altri, relativi al modo in cui tali sistemi creano o inaspriscono condizioni avverse al benessere umano: su tali temi, le responsabilità non sono imputabili alle tecnologie, bensì a esseri umani – talora i produttori dei sistemi di IA, talvolta i loro utenti, spesso entrambi. Non dovremmo quindi permettere a facili millenarismi di distrarci da queste priorità: in attesa che l'IA conquisti il pianeta, ci sono vari fronti su cui è urgente intervenire.

Mi limiterò a elencarne quattro:

1. *Scarsa trasparenza*: il tema della trasparenza degli algoritmi è oggetto di attenzione nell'IA da decenni, tanto che settori come l'IA spiegabile (explainable Artificial Intelligent, o Xai) e l'IA affidabile hanno acquisito lo status di vere e proprie sotto-discipline nelle scienze computazionali. Eppure, gli sforzi si orientano a rendere comprensibile il funzionamento tecnico dei sistemi di IA, con scarse aspettative di successo, data l'opacità radicale di queste tecnologie; curiosamente, meno attenzione viene garantita ad altri livelli di spiegazione dell'IA, assai meno opachi e plausibilmente di maggiore interesse per la collettività, quali i meccanismi di mercato che rendono redditizia l'IA e il modo in cui vengono distribuiti i conseguenti benefici (*cui prodest* IA?).

2. *Il valore dei dati*: se l'architettura dei sistemi di IA costituisce il motore della loro straordinaria capacità di apprendimento automatico, i dati ne rappresentano il carburante, di cui questi programmi sono voraci. Dati che vengono per la maggior parte forniti, spesso inconsapevolmente, da tutti noi. Questo pone da un lato problemi di protezione dei dati (chi fornisce il consenso al loro utilizzo per addestrare sistemi di IA, e con quali fini e conseguenze?), dall'altro questioni di redistribuzione dei profitti (se i nostri dati sono condizione necessaria a produrre rendite astronomiche, perché mai dovrebbero

beneficiarne solo pochi?). Quest'ultimo tema è al centro di diverse proposte, alcune di stampo privatistico (i data dividends, molto discussi negli Stati Uniti ma fino a oggi mai implementati), altre di natura collettiva e fiscale (la Digital Service Tax introdotta da vari Paesi, fra cui l'Italia, e attualmente in fase di negoziazione multilaterale fra i Paesi dell'Ocse).

3. *L'impatto dell'IA generativa sul mercato del lavoro*: in un distopico ribaltamento del sogno originario dell'automazione, in cui le macchine si sarebbero fatte carico di lavori pesanti e noiosi, lasciando noi liberi di dedicarci a occupazioni intellettualmente stimolanti, gli attuali sistemi di IA generativa sono invece straordinariamente bravi a sostituire i lavoratori umani in professioni altamente creative, quali quelle di scrittore, giornalista, attore, cantante o persino influencer. Al di là dell'effettivo talento manifestato da tali sistemi in queste attività, ciò che le rende minacciose per i lavoratori umani è l'abbattimento dei costi che esse consentono: di conseguenza, è irrealistico pensare che il mercato si autoregoli in merito, e diventano urgenti interventi correttivi. A titolo di esempio, si pensi alla recente vertenza di sceneggiatori e attori statunitensi contro gli studi di produzione, anche per limitare l'uso di IA generativa al posto di professionisti umani.

4. *Conseguenze cognitive dell'uso dell'IA generativa*: così come possono sostituire lavoratori umani in determinate professioni, i sistemi di IA generativa ci consentono di delegare loro un numero crescente di funzioni cognitive, e anzi ci incoraggiano a farlo, spesso e volentieri. Perché perdere tempo a riassumere un testo, scrivere un messaggio, compilare righe di codice, preparare una bibliografia, organizzare il proprio curriculum vitae, quando si può chiedere a un Llm di farlo al posto nostro, ottenendo buoni risultati con minima fatica? Qui si pone un dilemma antico, caratteristico di qualunque innovazione tecnologica: potenziare o rimpiazzare? Semplificando, tecnologie che aumentano le nostre capacità cognitive offrono un guadagno netto di competenza, seppure al prezzo di una parziale o totale dipendenza dal mezzo tecnologico; al contrario, tecnologie che rimpiazzano capacità cognitive preesistenti spesso si traducono in una perdita secca di autonomia, senza apprezzabile aumento di competenza e con analoghi livelli di dipendenza tecnologica. I sistemi di IA generativa, al momento, sembrano favorire più un uso sostitutivo che non accrescitivo, il che deve allarmare.

Si tratta, come dicevo, di problemi prosaici, che evocano preoccupazioni dal profilo noto, di carattere economico e sociale, unite a riflessioni su come assicurarsi che l'IA renda anche gli esseri umani un po' più intelligenti, anziché aiutarli a diventare sempre più torpidi e imbelli. Di conseguenza, sono ansie poco attraenti, esteticamente meno soddisfacenti dell'affresco a tinte fosche di un'incombente "rivolta delle macchine". Tuttavia, è di questi rischi che occorre occuparsi collettivamente e preoccuparsi individualmente, quando si ragiona sugli sviluppi dell'IA: il resto, per ora, rimane solo monito vago e in odore di propaganda.