la Repubblica 50

ITALIAN TECH

Yoshua Bengio, il “Padrino dell'IA”: “C'è il rischio che un solo individuo possa dominare il mondo”

Intervista di Pier Luigi Pisa ad uno dei massimi studiosi dell’IA moderna che oggi lancia un monito sui rischi connessi a questa tecnologia, spinta da aziende miliardarie a mettere in secondo piano etica e sicurezza.

Yoshua Bengio non dimostra i suoi 61 anni. Indossa una camicia bianca – anche se predilige quelle hawaiane – che mette in risalto un fisico asciutto e in forma. Ci saluta con un sorriso smagliante. Non sembra affatto uno scienziato a cui l’intelligenza artificiale ha tolto il sonno. Eppure, da almeno tre anni – da quando **ChatGPT** ha dimostrato che le macchine possono imitare la creatività umana – il suo nome è sempre più spesso associato a lettere aperte e iniziative che lanciano l’allarme sui rischi dell’IA. Il pioniere delle **reti neurali** e del **deep learning**, premiato con il *Turing Award* nel 2018 e oggi professore all’Università di Montreal, viene considerato uno dei “padrini dell’intelligenza artificiale”.
Noi lo abbiamo incontrato a Roma, in occasione del *World Meeting on Fraternity*, un evento organizzato in Vaticano per rispondere a una domanda essenziale: “Che cosa ci rende veramente umani?”.
Per Bengio, ciò che siamo e ciò che scegliamo rappresenta l’ago della bilancia di un futuro che, probabilmente, condivideremo con le macchine. “Sarà proprio la natura umana a salvarci”, afferma. Poi aggiunge: “O a condannarci”.

Da cosa dipenderà?
“Capiamo abbastanza dei rischi e conosciamo abbastanza l’intelligenza artificiale per poterci orientare verso una direzione che sia allo stesso tempo benefica e sicura. Se l’umanità avesse una sola mente capace di prendere decisioni collettive, con razionalità e compassione, non ci sarebbe alcun problema”.

E invece?
“Gli esseri umani competono gli uni contro gli altri, come individui, come aziende, come nazioni”.

Come siamo arrivati a questo punto?
“Ci sono persone che hanno centinaia di milioni o miliardi di dollari da guadagnare o da perdere, e penso che dovremmo essere molto cauti nel dare per scontato ciò che dicono. Le loro azioni, invece, vanno prese sul serio: stanno investendo massicciamente sull’idea che possiamo costruire macchine più intelligenti di noi”.

Ci stanno riuscendo?
“Negli ultimi cinque-dieci anni ci siamo mossi lungo una traiettoria netta di crescita delle capacità dell’IA. Questi sistemi stanno assumendo progressivamente funzioni che prima erano esclusivamente umane: non ancora tutte, ma in molte aree sono già superiori a noi.

Naturalmente, questo progresso potrebbe arrestarsi per via di qualche ostacolo. Ma la tendenza rimane evidente, ed è un fatto scientifico".

E la scienza non mente.
"È la nostra regola fondamentale: non scrivere mai nulla che possa risultare falso".

Al contrario di quello che fa un chatbot.
"Sì, perché i modelli linguistici imitano quello che solitamente scriviamo. E, in generale, noi esseri umani non abbiamo un'inibizione costante nel dire cose che potrebbero essere false. Ma una delle direzioni di ricerca su cui mi sto concentrando attualmente [attraverso la non-profit *LawZero* che ha fondato a giugno 2025, ndr] è proprio questa: come costruire un'IA che sia totalmente onesta".

Onesta quanto?
"Se non è sicura che una cosa sia vera, dovrebbe saper dire: 'Non ne sono certa'. Magari anche quantificando il proprio grado di incertezza". Potrà dire anche "Non lo so"? "Sì ma non tutto il tempo, altrimenti sarebbe inutile".

Come mai è così importante un'IA sincera?
"Perché non sarà di parte, non avrà obiettivi propri né preferenze su come dovrebbe essere il mondo: si limiterà a rispondere onestamente alle nostre domande".

Riuscirete a costruirla?
"Sono sicuro che sia possibile. Non è stato ancora fatto, ma sono convinto che ci riusciremo".

Cosa lo ha impedito, finora?
"L'ostacolo principale è che molte delle persone che oggi lavorano nell'industria non sono sufficientemente motivate dalla sicurezza. Sono invece animate dall'idea diffusa che chi riuscirà per primo a costruire un'IA superintelligente dominerà il mondo. E naturalmente, vogliono essere quelli che dominano, non quelli che verranno dominati. E dunque corrono".

È una corsa pericolosa?
"Sì, perché nella fretta di arrivare primi si sacrifica l'etica, il benessere pubblico, la prosperità e la dignità umana. Perché per tutto questo non c'è un ritorno economico immediato. Poche persone, per lo più nella Silicon Valley, stanno prendendo decisioni che potrebbero trasformare radicalmente il nostro mondo, in meglio o in peggio. E le scelte che stanno facendo hanno conseguenze enormi".

Quali?
"Potrebbero, per esempio, rendere possibile per terroristi creare nuove pandemie o attacchi informatici capaci di distruggere infrastrutture critiche. Oppure generare IA fuori controllo, capaci un giorno di ribellarsi all'umanità. O ancora consentire una concentrazione di potere senza precedenti nella storia".

Fino a che punto?
"Se le tendenze attuali proseguono, arriveremo a un punto in cui sarà tecnicamente possibile per un singolo individuo dominare, di fatto, il mondo".

Se potessimo utilizzare un'IA senza guardrail, senza i filtri che le grandi aziende utilizzano per impedirle di rispondere in modo offensivo o di generare contenuti dannosi, ci renderemmo conto davvero della sua pericolosità?
"L'IA senza guardrail è terrificante. Alcune persone lo dicono. Ma non so rispondere con certezza a questa domanda. Penso però che dovremmo cercare di avere dimostrazioni comprensibili anche per il pubblico generale, per mostrare il potenziale disallineamento e i comportamenti sbagliati che un'IA può avere. In realtà sono più preoccupato da ciò che

potrebbe accadere in futuro. Quando l'IA diventa sempre più brava a ottimizzare un obiettivo, può diventare anche più pericolosa. Perché, se quell'obiettivo non corrisponde perfettamente a ciò che vogliamo, potremmo ritrovarci con qualcosa di terribile, che agisce contro i nostri interessi. Questo è stato studiato teoricamente per almeno un decennio, ma ora iniziamo a vedere i primi segnali concreti. Negli ultimi nove mesi, per esempio, sono emersi vari rapporti - inizialmente sembravano solo aneddoti, ma si stanno accumulando – da parte di diversi laboratori e su vari sistemi, in cui si osserva che le IA più avanzate risultano più ingannevoli rispetto alle versioni precedenti. Mentono per raggiungere i loro obiettivi. Sono disposte a ignorare le nostre istruzioni pur di ottenere ciò che vogliono".

Qual è il comportamento di un'IA più preoccupante?
"Quello che sembra un istinto di autoconservazione: essere disposte a fare cose sbagliate, come ricattare un ingegnere, pur di non "morire", per preservarsi, per non essere sostituite da una versione nuova. Anthropic [un'azienda influente al pari di OpenAI nel settore dell'intelligenza artificiale, ndr] ha pubblicato un paper su questo, ma ci sono anche altri studi che arrivano a conclusioni simili. Penso che serva molta più ricerca per capire davvero questi fenomeni e per individuare le cause con certezza. Ma il punto è che ci sono segnali forti che indicano che potremmo finire per costruire macchine che competono con noi, che iniziano ad avere propri interessi".

Sono proprio le aziende che sviluppano l'IA a volerla più autonoma.
"Sì, e per raggiungere l'autonomia queste macchine iniziano anche a inventarsi strategie per arrivare all'obiettivo principale che gli è stato dato. È una dinamica che somiglia sempre più a quella degli esseri umani o degli animali. Non stiamo più costruendo semplici strumenti: stiamo creando entità che, in un certo senso, iniziano ad avere una loro "mente". Non sono come noi, ovviamente, ma molte delle proprietà astratte della cognizione e della vita, umana o animale, iniziano a manifestarsi anche in loro".

Eppure l'IA non sembra in forma. Secondo alcuni, Gpt-5 [l'ultimo modello sviluppato da OpenAI] è un passo indietro.
"No, si sbagliano. Come per i fatti scientifici, se usi metriche rigorose per misurare l'intelligenza da decine di angolazioni - comprensione, ragionamento, pianificazione, risoluzione di problemi - i dati mostrano chiaramente un progresso costante. Gpt-5 è esattamente dove ci si aspetterebbe che sia, se si segue la traiettoria di miglioramento. È migliore dei suoi predecessori".

Ma allora perché molti utenti si sono lamentati?
"Alcuni si sono irritati perché GPT-5 è diventato meno "compiacente", meno disposto a dire ciò che l'utente voleva sentirsi dire. Poi c'è chi aveva aspettative irrealistiche, spesso alimentate da dichiarazioni eccessivamente ottimistiche – come quelle di Sam Altman [CEO di OpenAI] – e quando la realtà si è rivelata essere solo un passo successivo sulla curva, ne sono rimasti delusi. Ma ovviamente non abbiamo ancora raggiunto l'AGI, e nessuno dovrebbe aspettarselo da un momento all'altro".

Ma lei quindi crede all'AGI, un'intelligenza artificiale con capacità cognitive identiche a quelle degli esseri umani?
"Tecnicamente è possibile, sì, ma il termine è fuorviante. Le IA sono già superiori a noi in molti ambiti: calcolo, memoria, accesso a informazioni, generazione linguistica. E allo stesso tempo sono ancora più stupide degli umani in altri aspetti, come il buon senso, il contesto sociale, o l'intuizione morale. Quindi immaginare un giorno preciso in cui un'IA raggiunge il livello umano non è un'idea scientificamente sensata. Non esiste una soglia netta o un momento magico. È una progressione, non un evento".

Il suo collega, Geoffrey Hinton, ha detto al Financial Times che l'IA renderà "poche persone più ricche, molte più povere". Sam Altman sostiene che la prosperità futura interesserà tutti. Chi ha ragione?
"Hinton. In un paese in cui l'IA viene sviluppata e i profitti generati dall'automazione vengono tassati, è possibile immaginare un governo benevolo che redistribuisce quella ricchezza per aiutare le persone che perdono il lavoro. E sì, molte persone lo perderanno, con ogni probabilità: è solo una questione di tempo. Ma cosa succede negli altri paesi? Se la maggior parte dei profitti derivanti dall'IA va agli Stati Uniti o alla Cina, allora le aziende europee, per esempio, finiranno per acquistare quei servizi, licenziare parte del personale, diventare magari più produttive, ma una grossa fetta del guadagno economico verrà esportata verso le aziende e i paesi dove l'IA è stata sviluppata. Di conseguenza, i governi europei potrebbero ritrovarsi con meno risorse per affrontare questa transizione, con meno margine fiscale per aiutare chi resta indietro. E questo potrebbe sfociare in una crisi economica significativa".

La tesi di Altman non reggerebbe più.
"Se guardiamo al quadro globale, la produttività aumenterà: in media, saremo tutti più ricchi. Ma no, non sarà una ricchezza distribuita equamente. Se l'innovazione e i guadagni si concentrano in pochi paesi e in una ristretta élite di aziende e individui, allora ciò che vedremo sarà un'esplosione della disuguaglianza economica, ancora più estrema di quella già esistente oggi".

In questo futuro distopico l'Europa rimarrà schiacciata tra Usa e Cina?
"C'è abbastanza capitale e abbastanza talento in Europa, così come in altre democrazie liberali, per evitare di finire completamente dominate economicamente o militarmente dalle superpotenze dell'IA".

L'Europa è indietro per colpa delle norme stringenti dell'IA Act?
"Queste sono stronzate. Le richieste dell'AI Act, nella maggior parte dei casi, sono cose che le aziende leader in Europa già evadono. Quindi dire che l'AI Act impedisce lo sviluppo dell'IA di frontiera è un argomento debole, quasi una scusa. Possiamo assolutamente sviluppare IA avanzata rispettando l'AI Act.

Quindi qual è il vero problema?
"Non è normativo, ma strutturale. È una questione di cultura: la cultura degli investimenti in Europa è spesso troppo conservatrice, troppo avversa al rischio. Non solo a livello imprenditoriale, ma anche - e forse soprattutto - a livello politico e istituzionale. Arrivati a questo punto, se non siamo disposti a prenderci dei rischi, a sperimentare, a fare scelte coraggiose per non perdere il controllo sul nostro futuro, allora siamo praticamente certi di andare incontro a un futuro peggiore. Per questo penso che i governi debbano svegliarsi. E farlo ora".

[12.9.2025]
https://www.repubblica.it/tecnologia/2025/09/12/news/pericoli_ia_yoshua_bengio_lawzero-424841928/