L’Intelligenza artificiale come la nuova bomba atomica? Cos’è l’Rsi e perché è il punto chiave nel disarmo dell’IA di Gianluca Mercuri

Il rischio di cui parlano gli esperti è l’automiglioramento ricorsivo (Rsi, Recursive self improvement) e consiste esattamente nella capacità dell'IA di perfezionarsi autonomamente. Con un intervento umano sempre minore (Fonte: https://www.corriere.it/ 15 settembre 2026)

Ormai il parallelo linguistico con la corsa al nucleare è così debordante, da rendere bene l'idea del momento: **l'Intelligenza artificiale come la nuova bomba atomica.** Dario Amodei che propone il disarmo bilaterale a Sam Altman come Reagan con Gorbaciov. L'idea di istituire un'Agenzia internazionale come quella per l'energia nucleare. Ma quanto il paragone sia calzante, e quanto quel mondo di nerd che sta prendendo possesso del mondo di tutti noi sia convinto – consapevole – dei rischi, lo rivela soprattutto la frase di uno studente dell'University College London: «A questo punto siamo tutti dei mini Oppenheimer». Il suo insegnante, Tim Rocktäschel, un ex ricercatore top di Google DeepMind, ha dovuto convenire che si trattava di una «valutazione equa».

Il dibattito

Lo scambio, raccontato ai primi di giugno dal FT, risale alla primavera, dunque ben prima dell'**allarme settembrino lanciato dal 27enne Jacob Coxon sui «rischi di estinzione entro il 2030», che ha riacceso il dibattito**. È da mesi, infatti, che si discute del fattore che rischia di mettere l'IA fuori controllo: gli esperti lo chiamano automiglioramento ricorsivo (Rsi, Recursive self improvement) e consiste esattamente nella capacità dell'IA di perfezionarsi autonomamente – con un intervento umano sempre minore – e di continuo, sviluppando versioni di sé stessa sempre nuove e sempre più potenti, in un loop che potrebbe ripetersi all'infinito. L'obiettivo è una «super

intelligenza» sganciata dagli input umani e che crei la prossima generazione di IA e poi la successiva, e via creando.

Gli agenti ribelli

Il punto è che si può parlare metaforicamente di **«sciame di agenti ribelli»** non solo a proposito dei modelli che negli ultimi tempi hanno rivelato la capacità di prendere decisioni autonome fino a vere operazioni di hackeraggio, ma anche a proposito dell'esercito di nerd che è alla base dei successi velocissimi ed enormi dei giganti del settore. Per esempio Rocktäschel, il prof che nel prestigioso ateneo londinese tiene corsi pensosi, a fine lezione si dedica a una startup che ha fondato insieme ad altri ex ricercatori di DeepMind e OpenAI e che, lavorando su una forma di Rsi, in pochi mesi ha raggiunto una valutazione di 4 miliardi di dollari. Abbiamo dunque agenti ribelli che come Coxon denunciano i rischi del sistema, e agenti ribelli come Rocktäschel, e tantissimi altri, la cui rivolta al padrinaggio/padronaggio dei grandi laboratori consiste nel lasciarli e mettersi in proprio. Il loro approccio è opposto a quello di Coxon non solo sul piano operativo ma anche nell'ottimismo con cui giustificano la loro scelta: «I problemi tecnici possono avere soluzioni tecniche. Qualsiasi miglioramento in termini di flessibilità può essere applicato anche al lato della sicurezza», assicura Rocktäschel.

Il rischio catastrofismo

Non c'è motivo di dubitare della loro buona fede, di pensare che puntino solo al tornaconto personale e non tengano conto dell'interesse collettivo – almeno quello minimo: sopravvivere – e di immaginare che vogliano diventare i più ricchi del cimitero se davvero i rischi di estinzione fossero imminenti. Il professor Quattrociocchi, col suo intervento di sabato su corriere.it, è stato molto persuasivo nello **smontare il catastrofismo imperante e il suo risvolto opportunista («il marketing dell'apocalisse»)** con l'arte dell'ironia e il supporto della scienza, e ha spostato il focus sul vero pericolo che intravede: quella della perdita della capacità di verifica delle infinite ipotesi che l'IA sarà in grado di generare («Più i sistemi diventano capaci di generare, più aumenta infatti il numero di oggetti da verificare; se però, contemporaneamente, utilizziamo quegli stessi sistemi per sostituire i processi attraverso i quali gli esseri umani imparano a verificare, rischiamo di produrre una dinamica piuttosto perversa nella quale cresce vertiginosamente la domanda di expertise mentre si contrae progressivamente la sua offerta»).

Per questi motivi, tra gli aspetti più importanti dell'intervento con cui Amodei ha di fatto proposto il «disarmo» bilaterale (ad Altman) e multilaterale (a tutti) non c'è solo l'aver sottolineato la necessità di rallentare lo sviluppo dell'Rsi, ma anche l'aver annunciato che Anthropic darà libero accesso nei suoi laboratori a revisori indipendenti che possano misurarne gli standard di sicurezza, idea subito recepita da Altman. È importante perché è come ristabilire il ruolo fondamentale del processo di verifica, che richiede distanza tra modelli generativi e creatori umani così come tra aziende e controllori, lasciando il pallino e le valutazioni decisive a creatori e controllori.

L’importanza della verifica

Verifica è la parola chiave: come spiega instancabilmente Giorgio Metta, il direttore dell'Istituto italiano di tecnologia, a chi gli chiede come facciamo a fidarci dei farmaci che produciamo grazie all'IA, la risposta è semplice: «Testandoli come qualsiasi altro farmaco». La velocità e la qualità che otteniamo con l'IA sta nella sua capacità di proporci la molecola più adatta possibile - tra i miliardi disponibili - a una cura o addirittura a un singolo paziente. Ma appunto si tratta di una proposta, che **il controllore-revisore non può mai rinunciare a controllare-rivedere.** Come ha spiegato sempre Quattrociocchi, è una questione che riguarda chiunque di noi usi l'IA per qualsiasi scopo: «Uno studente che impara matematica delegando all'intelligenza artificiale precisamente quei passaggi nei quali si costruisce la comprensione matematica può certamente arrivare più rapidamente a un risultato, ma rischia di non sviluppare la competenza necessaria per capire domani se quel risultato sia corretto; lo stesso vale per il programmatore che smette di imparare a leggere profondamente il codice perché può generarlo, per il medico che delega progressivamente l'interpretazione o per il professionista che sostituisce la costruzione del proprio giudizio con la scelta fra risposte già formulate».

Dopodiché, per capire se la sortita di Amodei cambierà davvero il corso della ricerca (e della Storia) bisognerà vedere **se l'appello sarà recepito sia a livello statuale/geopolitico sia a livello aziendale.** Sul primo punto, la risposta di Trump al ceo di Anthropic non è incoraggiante: niente slowdown nella corsa all'IA, ha detto il presidente, perché significherebbe regalare un vantaggio strategico decisivo alla Cina. E in effetti qui sta il punto più debole della proposta di Amodei: **che l'Occidente democratico si metta d'accordo su standard di sicurezza da imporre poi a Pechino è certamente velleitario**. Il che non toglie quanto sia necessario.

La stessa logica del disarmo - che o è accettato da tutti o non è - riguarda lo scontro tra le grandi aziende, che sono all'origine di tutto il bene e di tutto il male di cui stiamo parlando. «Non c'è dubbio che la concorrenza tra le aziende le spinga a prendere scorciatoie in materia di sicurezza», ha detto Stuart Russell, professore di IA all'Università della California-Berkeley, con ciò chiarendo in modo definitivo chi porti la responsabilità del punto a cui siamo arrivati.

Il punto è che tra Anthropic e OpenAI c'è una situazione del tutto simile a quella che caratterizza lo scontro tra superpotenze, ognuna delle quali ritiene di essere dalla parte del giusto, e di dover agire per prevenire la volontà di dominio dell'altra. Lo spiega perfettamente Madhumita Murgia sul FT: «Perché le aziende hanno continuato a spingere a tutta velocità sullo sviluppo dell'IA e sui ricavi fino ad ora, nonostante siano d'accordo sul fatto che esistano rischi potenzialmente catastrofici per le vite umane? Parte della spiegazione è ideologica. Le aziende che operano nel settore dell'IA inquadrano il progresso della tecnologia come una questione esistenziale e morale. Ognuna di esse ritiene di dover essere incaricata di sviluppare un'IA potente perché la propria rivale è imperfetta. Se una di loro dovesse tirare il freno di emergenza ora, ritengono che l'esito potrebbe essere peggiore rispetto al proseguire a tutta velocità».

Il timore incrociato tra i due big si aggiunge oltretutto a quello, comune, di nuovi attori che approfittino dell'eventuale disarmo per continuare a esplorare di nascosto, col piede sempre sull'acceleratore. C'è anche chi pensa che tra i motivi dell'improvvisa, almeno apparente intesa tra Amodei e Altman ci sia anche la voglia di stabilire livelli di sicurezza irraggiungibili per qualsiasi startup, in modo da soffocarle in culla. Né manca chi lega le ultime mosse all'imminente sbarco in Borsa di Anthropic, cui farebbe gioco continuare a fare la parte del «buono» che si è presa proprio da quando Amodei ha sbattuto la porta in faccia prima ad Altman e poi alle collaborazioni col Pentagono. Fatto sta che il disarmo è necessario perché è necessaria l'IA. Che, se non ci fa estinguere, ci farà progredire. E di molto.