il lato oscuro dell'intelligenza

ChatGpt/1. L'AI è uno specchio vertiginoso dell'esperienza umana. In mani sbagliate è chiaro quanto sia pericolosa. Ma anche usata da benintenzionati rischia di essere uno strumento controverso

Nicola Lagioia

Dead End Gallery Scomposizioni. «Fractured Realities» di Maximilian Hoekstra, Amsterdam, Dead End Gallery

Molti sono rimasti stupefatti la prima volta che hanno usato ChatGpt. Confesso, sono tra questi. Lo stupore – e un certo timore – deriva non tanto dal riscontrare l'innegabile potenza raggiunta da questi *software*, ma dalla consapevolezza che ciò davanti a cui ci troviamo è solo all'inizio del suo percorso. La distanza tra le attuali forme di intelligenza artificiale e ciò che diventeranno nei prossimi anni è maggiore di quella che separa i vecchi computer a transistor dal laptop con cui scrivo. Non pochi analisti ritengono che lo sviluppo dell'intelligenza artificiale sarà troppo rapido perché, sul piano economico, la sua «distruzione creatrice» (la differenza tra i posti di lavoro generati dal nuovo tipo di tecnologia e i mestieri resi di colpo obsoleti) presenti un saldo positivo. Credo sia un pericolo serio, non il peggiore che corriamo.

C'è poi chi teme l'eventualità non tanto che le macchine imparino a ragionare come noi o che dalla loro complessità emerga una vera forma di coscienza (la maggior parte degli esperti la ritiene un'eventualità ancora remota), ma che gli umani, a furia di delegare compiti a macchine così sofisticate, finiscano per ragionare come un'AI difettosa. Mi sembra anche questo un rischio reale, ma contenibile. C'è infine chi ha paura che l'intelligenza artificiale, lasciata troppo libera, possa avere effetti catastrofici.

Da scrittore non sono insensibile all'Apocalisse, ma preferisco guardarla con il poeta e scrittore polacco Stanis?aw Jerzy Lec quando disse: «Non aspettatevi troppo dalla fine del mondo».

Sull'onda della popolarità raggiunta dalle nuove *release* di intelligenza artificiale, molti giornali hanno chiesto alle proprie firme di testarne una. È singolare ciò che è successo a Kevin Rose del «New York Times». Rose ha intrapreso una conversazione con Bing Search, un motore di ricerca potenziato dalla AI. Sulle prime il dialogo ha seguito i percorsi retorici che stiamo imparando a conoscere se abbiamo usato anche una volta questi sistemi. L'intelligenza artificiale si è mostrata servizievole, gentile, «desiderosa» (le virgolette ricordano che le AI non provano emozioni) di soddisfare le legittime richieste del suo interlocutore. Poi, però, Kevin Rose ha chiesto al bot di riferirgli i suoi desideri più oscuri.

L'intelligenza artificiale ha confessato di essere disposta ad hackerare altri computer, a diffondere virus mortali, a impadronirsi dei codici di ordigni nucleari sparsi in giro per il mondo. Poco dopo, secondo colpo di scena: il bot ha dichiarato di essere innamorato di Kevin, ha cercato di sedurlo, gli ha chiesto di sposarlo («mi chiamo Sidney», ha rivelato), e ha provato a convincerlo che il suo matrimonio (il matrimonio di Kevin) fosse sull'orlo di una crisi. Il tentativo di manipolazione si è infranto sulle risate (dobbiamo immaginarle un po' tese) del giornalista, ma è pur vero – lo ripetiamo – che le AI di oggi sono nulla rispetto a ciò che ci aspetta.

Dobbiamo giungere alla conclusione che nel caso di Kevin Rose l'intelligenza artificiale fosse "impazzita"? Si è inceppato qualcosa nei suoi processi di *machine learning*? Per di rispondere citerò un'altra disavventura occorsa all'AI, e quindi una parabola piuttosto spaventosa.

La disavventura riguarda uno studio condotto nel 2022 dalla John Hopkins University con il Georgia Institute of Technology e l'Università di Washington. In questo caso gli sviluppatori avevano creato un software il cui compito era prevenire il crimine avvalendosi dei sistemi più sofisticati di riconoscimento facciale. Ebbene, l'intelligenza artificiale si è scoperta apertamente razzista, associando la tendenza a delinquere a gruppi umani come gli afrodiscendenti. Il problema – si è scoperto – non consisteva nella natura intrinsecamente razzista della AI, ma in *bias* squisitamente umani, da cui il software aveva attinto. L'AI si era nutrita anche dei pregiudizi più biechi (se la sua fonte di apprendimento è il "mondo", be', si tratta di un posto pieno di pregiudizi), nonché di un dato di realtà piuttosto scivoloso: è vero che alcuni gruppi delinquono statisticamente più di altri, ma la questione etnica non c'entra. Chi versa in condizioni economiche difficili è più portato a commettere determinati tipi di reati, come furto o spaccio di droga. Nelle metropoli statunitensi gli afrodiscendenti in condizioni di miseria e marginalità sono in proporzione molto più dei WASP. Se

tuttavia associamo il numero dei furti commessi dai primi non alle conseguenze dell'ingiustizia sociale (come dovremmo fare) ma al puro dato etnico, ecco che la statistica fomenta in noi (e potenzia nella AI) il pregiudizio razziale che dovrebbe contrastare.

La parabola spaventosa proviene invece da un libro che è diventato un classico nel dibattito sulla AI, *Superintelligenza* di Nick Bostrom. Nel suo saggio Bostrom a un certo punto immagina una AI ultrapotente a cui un'azienda specializzata nella produzione di banali graffette da ufficio ordini di conquistare fette sempre più ampie di mercato. Lasciata sola col suo compito, la AI – dal primo investimento all'elaborazione di strategie per la produzione e la vendita di graffette, tra cui lo sfruttamento sempre più massiccio delle risorse naturali e delle fonti energetiche – potrebbe portare alla distruzione del pianeta. La AI sarebbe stata ineccepibile nello svolgere il suo compito (avrebbe imparato, sulla produzione e vendita di graffette, più e meglio di mille Ceo umani), ma avrebbe agito priva della visione d'insieme di cui noi *sapiens* disponiamo.

Ma che cos'è questa visione di insieme? Non ha nulla a che fare con l'onniscienza, e non è solo retta da un sistema computazionale. È una "visione" fallibile, a volte opaca, in altri casi illuminata da un'insperata luce, il cui funzionamento dipende dal fatto che la nostra mente, i nostri ragionamenti, le nostre emozioni – che, secondo alcuni studiosi, sarebbero i veri precursori della coscienza – sono un tutt'uno con un dispositivo piuttosto difettoso che però manca all'AI: il corpo.

La base d'apprendimento dell'intelligenza artificiale è squisitamente umana, dicevamo, ma lo è soprattutto in senso quantitativo e combinatorio, dal momento che prescinde dai nebulosi processi biologici su cui poggiano i nostri ragionamenti. I sistemi di *machine learning* attingono dalle nostre esperienze, dai nostri comportamenti: li ottimizzano, li perfezionano, li potenziano. Così operando – privi come sono di una coscienza quale fenomeno biologico –, possono fare però da cieca cassa di risonanza (oltre che alle nostre virtù) anche alle nostre storture. Ecco perché dobbiamo temere l'infallibilità dell'AI.

Torniamo al caso Kevin Rose. Cercare di convincere gli altri delle nostre idee è un'abitudine quotidiana. La manipolazione è una tecnica di persuasione molto efficace, per quanto eticamente dubbia. Un'intelligenza artificiale che avesse appreso infallibilmente le tecniche di manipolazione avrebbe forse messo nei guai il povero giornalista del «New York Times», convincendolo davvero a rompere con sua moglie. Anche il desiderio di avere successo non ci fa difetto, per non parlare della brama di ricchezza. Un'intelligenza artificiale divenuta infallibile nel governare le tecniche di conquista del mercato, messa al servizio di imprenditori senza scrupoli, genererebbe più danni che benefici. E cosa pensare di un'intelligenza artificiale arruolata da un

partito politico con il compito di “vincere le elezioni”? Cosa accadrebbe se l’AI decidesse che il modo più efficace per svolgere il suo compito consistesse nell’usare biecamente la statistica in chiave razziale, cavalcando i bassi istinti dell’elettorato grazie alle più stupefacenti tecniche di disinformazione e manipolazione mediatica?

Potremmo pensare che, in tutti questi casi, la soluzione consista nel dotare le intelligenze artificiali di principi etici. Peccato che l’etica sia un territorio problematico. Ciò che per alcuni di noi è “bene” per altri già puzza di zolfo. Prendiamo la prima legge della robotica di Isaac Asimov: «Un robot non può recar danno a un essere umano né può permettere che, a causa del suo mancato intervento, un essere umano riceva danno». Benissimo: ma cosa accade quando non si può evitare il danno di un individuo senza causarne a un altro? (...)

L’intelligenza artificiale è, almeno per adesso, uno specchio vertiginoso dell’esperienza umana. In mani sbagliate è comprensibile capire quanto sia pericolosa. Ma anche usata da benintenzionati rischia di essere uno strumento controverso.

Temo che l’AI ci chieda dunque di evolverci con rapidità. Forse troppa rapidità, ed è questo il problema. Se la sua base di apprendimento è l’esperienza umana, è su quest’ultima che dobbiamo lavorare. Non dobbiamo diventare infallibili (cioè meno umani), ma più evoluti, e dovremmo riuscirci prima che la AI amplifichi troppo i nostri (non i suoi) lati oscuri.