L’intelligenza artificiale può hackerare la democrazia (parola di Yuval Noah Harari)

di Luca Angelini (Il punto del Corriere della Sera di mercoledì 3 maggio 2023)

La cosa più sorprendente di ChatGpt-4, ha detto il francescano Paolo Benanti, pioniere dell’algoretica, nell’intervista a Paolo Ottolina di LogIn che abbiamo pubblicato anche nella Rassegna del weekend, è che «si dimostra drammaticamente simile all’uomo comune». Ed è una somiglianza potenzialmente foriera di grandi rischi. Perché, come scrive in un denso intervento sull’Economist lo storico, filosofo e saggista israeliano Yuval Noah Harari (autore di libri di successo planetario come Sapiens, Homo Deus e 21 lezioni per il XXI secolo) «se sto avendo una conversazione con qualcuno e non sono in grado di dire se sia un umano o un’intelligenza artificiale (AI), è la fine della democrazia». Insomma, il fatto che noi riusciamo a ingannare l’intelligenza artificiale è un problema (come hanno mostrato Milena Gabanelli e Simona Ravizza in un Dataroom), ma non quanto il fatto che l’intelligenza artificiale possa ingannare noi.

E il motivo è che «potremmo presto trovarci a condurre lunghe discussioni online sull’aborto, il cambiamento climatico o l’invasione russa dell’Ucraina con entità che pensiamo siano umani, ma in realtà sono AI. Il problema è che è assolutamente inutile passare il tempo a cercare di cambiare le opinioni dichiarate di un bot di intelligenza artificiale, mentre l’AI potrebbe affinare i suoi messaggi in modo così preciso da avere buone possibilità di influenzarci».

Per questo, a suo avviso, non dovremmo tanto preoccuparci dell’uso che di ChatGpt-4 e affini potrebbero fare per rimediare buoni voti a scuola, quanto di quello che potrebbe esserne fatto per rimediare buoni risultati alle elezioni.

La vera (e inquietante) novità portata da ChatGpt-4 e analoghi sistemi è la capacità di maneggiare il linguaggio in un modo che, appunto, sembra umano. E ciò, per Harari, è un rischio capitale per l’umanità. «Negli ultimi due anni sono emersi nuovi strumenti di intelligenza artificiale che minacciano la sopravvivenza della civiltà umana sotto un profilo inaspettato. L’AI ha acquisito alcune notevoli capacità di manipolare e generare il linguaggio, sia con parole, suoni o immagini. Ha quindi hackerato il sistema operativo della nostra civiltà».

Non sarà un po’ esagerato? Può darsi. Ma il filosofo israeliano invita a riflettere sull’importanza che il linguaggio ha per l’umanità, per quel che fa di noi umani quel che siamo. «Il linguaggio è la sostanza di cui è fatta quasi tutta la cultura umana. I diritti umani, ad esempio, non sono inscritti nel nostro Dna. Piuttosto, sono artefatti culturali che abbiamo creato raccontando storie e scrivendo leggi. Gli dei non sono realtà fisiche. Piuttosto, sono artefatti culturali che abbiamo creato inventando miti e scrivendo scritture (questo può suonare irritante per i credenti, ma anche padre Benanti dice: «L’uomo ha bisogno dell’Altro e se l’Altro non lo troviamo più nei Cieli, lo cerchiamo nel cloud? Non ho una risposta», ndr). Anche il denaro è un artefatto culturale. Le banconote sono solo pezzi di carta colorati, e attualmente oltre il 90% del denaro non è nemmeno banconote: sono solo informazioni digitali nei computer. Ciò che dà valore al denaro sono le storie che ci raccontano banchieri,

ministri delle finanze e guru delle criptovalute. Sam Bankman-Fried, Elizabeth Holmes e Bernie Madoff non erano particolarmente bravi a creare valore reale, ma erano tutti e tre racconta-storie estremamente capaci».

Ricordando come teorie complottiste sgangherate come QAnon abbiano già influito sulla democrazia americana (vedi assalto del 6 gennaio 2021 al Campidoglio: qui un approfondimento di Sandro Modeo), Harari invita ad immaginare cosa potrebbe succedere con campagne di disinformazione molto più raffinate gestite dall'intelligenza artificiale generativa di ChatGpt, Bard e affini. Di qui il suo invito perentorio: «Lasciate perdere i compiti scolastici. Pensate alla prossima corsa presidenziale americana nel 2024 e provate a immaginare l'impatto degli strumenti di intelligenza artificiale realizzati per produrre in massa contenuti politici, notizie false e sacre scritture per nuovi culti».

Ad avviso di Harari, l'arma in più dei nuovi sistemi di intelligenza artificiale è la loro capacità di sedurci. Può sembrare un verbo fuori luogo. Ma c'è un episodio, che forse molti ricorderanno, al quale Harari guarda da un'ottica peculiare, che spiega bene cosa intenda. «Nel giugno 2022 Blake Lemoine, un ingegnere di Google, affermò pubblicamente che il chatbot AI Lamda, su cui stava lavorando, era diventato senziente. La controversa affermazione gli era costata il lavoro. La cosa più interessante di questo episodio non è stata l'affermazione del signor Lemoine, che probabilmente era falsa. Piuttosto, la sua disponibilità a rischiare il proprio lavoro redditizio per il bene del chatbot AI. Se l'AI può influenzare le persone a rischiare il posto di lavoro, cos'altro potrebbe indurle a fare?».

Per questo Harari sostiene che stiamo passando dai «mercanti di attenzione» (per usare il titolo di un libro di Tim Wu) ai «mercanti di intimità e confidenza» (intimacy). «In una battaglia politica per conquistare le menti e i cuori, l'intimità è l'arma più efficace e l'AI ha appena acquisito la capacità di produrre in serie relazioni intime con milioni di persone. Sappiamo tutti che negli ultimi dieci anni i social media sono diventati un campo di battaglia per controllare l'attenzione umana. Con la nuova generazione di AI, il fronte di battaglia si sta spostando dall'attenzione all'intimità. Cosa accadrà alla società umana e alla psicologia umana mentre l'AI combatte una battaglia per fingere relazioni intime con noi, che possono essere utilizzate per convincerci a votare per determinati politici o acquistare determinati prodotti? Anche senza creare "falsa intimità", i nuovi strumenti di intelligenza artificiale avrebbero un'enorme influenza sulle nostre opinioni e visioni del mondo».
E le possibili conseguenze sono di enorme portata. In primo luogo, per alcune aziende, settori e professioni. «Le persone potrebbero arrivare a utilizzare un singolo consigliere di intelligenza artificiale come oracolo unico e onnisciente. Non c'è da stupirsi che Google sia terrorizzata. Perché preoccuparsi di fare ricerche online, quando posso semplicemente chiedere all'oracolo? Anche l'industria dell'informazione e quella della pubblicità dovrebbero essere terrorizzate. Perché leggere un giornale quando posso semplicemente chiedere all'oracolo di dirmi le ultime notizie? E che scopo avrebbe la pubblicità, quando posso semplicemente chiedere all'oracolo di dirmi cosa comprare?».

Ma ciò è ancora nulla rispetto ai rischi che Harari vede per l'umanità come l'abbiamo finora conosciuta. «Ciò di cui stiamo parlando è, potenzialmente, la fine della storia umana. Non la fine della storia, solo la fine della sua parte dominata dall'uomo. La storia è l'interazione tra biologia e cultura; tra i nostri bisogni e desideri biologici per cose come il cibo e il sesso, e le nostre creazioni culturali come le religioni e le leggi. La storia è il processo attraverso il quale le leggi e le religioni danno forma al cibo e al sesso. Cosa accadrà al corso della storia quando l'AI prenderà il sopravvento sulla cultura e comincerà a produrre storie, melodie, leggi e religioni? Strumenti precedenti, come la stampa e la radio, hanno contribuito a diffondere le idee culturali degli esseri umani, ma non hanno mai creato nuove idee culturali proprie. L'AI è fondamentalmente diversa. Può creare idee completamente nuove, una cultura completamente nuova».
Un rischio sul quale ha insistito anche Geoffrey Hinton, padre delle «reti neurali», appena dimessosi da Google per poter denunciare i rischi dell'AI, è quello di non riuscire più a distinguere la realtà dalla finzione, tanto la seconda si farà raffinatamente simile alla prima.

Harari, ricorrendo alla storia della filosofia, parla di materializzazione del timore di vivere in un mondo illusorio, come nelle metafore del velo di Maya, della caverna di Platone o del genio maligno di Cartesio. Uno dei rimedi urgenti che il filosofo israeliano propone è una chiara e obbligatoria indicazione di tutti i testi prodotti dall'AI, anziché da umani. Anche padre Benanti, in un'intervista a Forbes Italia, ha spiegato: «Il diamante sintetico sarebbe indistinguibile da quello naturale se non fosse per due caratteristiche: non ha difetti al suo interno e per legge ha inciso al laser un numero di serie. Tuttavia vale quanto un diamante reale ed è in grado di ingannarci. La domanda è: abbiamo il diritto a essere avvisati che chi interagisce con noi è una macchina e non un essere umano?» (A noi viene da aggiungerne un'altra: chi e come potrebbe controllare la massa di testi che circolano online?).

Un'altra raccomandazione di Harari è l'urgente creazione («ne abbiamo bisogno da ieri, non da domani») di un equivalente della statunitense Food and drug administration (Fda), che passi al vaglio i nuovi prodotti basati sull'AI prima che possano essere resi di pubblico utilizzo, come la Fda oggi fa per i farmaci.
Harari non nega che l'AI possa avere anche una gran quantità di applicazioni positive, che possa aiutarci a sconfiggere il cancro, a frenare il cambiamento climatico e via elencando (un punto sul quale insiste, in un intervento sul Guardian, Ivana Bartoletti, esperta di sicurezza informatica e protezione dei dati, fondatrice di Women Leading in AI Network e più ottimista di Harari sulla capacità umana di «imbrigliare» l'AI con regole adeguate: «Scenari apocalittici di AI simili a quelli rappresentati nel film Terminator non dovrebbero renderci ciechi di fronte a una visione più realistica e pragmatica che veda il buono dell'AI e affronti i rischi reali. Regole del gioco sono necessarie e accordi globali sono fondamentali se vogliamo passare da uno sviluppo in qualche modo sconsiderato dell'AI all'adozione responsabile e democratizzata di questo nuovo potere»). Ma, aggiunge il saggista israeliano, «il compito degli storici e dei filosofi come me è indicare i pericoli».

Al riguardo, Harari non resiste a un paragone fra energia atomica e intelligenza artificiale: «Dal 1945 sappiamo che la tecnologia nucleare potrebbe generare energia a basso costo a beneficio degli esseri umani, ma potrebbe anche distruggere fisicamente la civiltà umana. Abbiamo quindi rimodellato l'intero ordine internazionale per proteggere l'umanità e per assicurarci che la tecnologia nucleare fosse utilizzata principalmente a fin di bene. Ora dobbiamo fare i conti con una nuova arma di distruzione di massa che può annientare il nostro mondo mentale e sociale».

Quanto al rischio che mettere limiti agli sviluppi dell'intelligenza artificiale nei Paesi democratici non faccia altro che dare un vantaggio competitivo ai Paesi autocratici che certi scrupoli non se li fanno, Harari spiega che le democrazie non hanno comunque alternative: «Implementazioni di AI non regolamentate creerebbero il caos sociale, che andrebbe a vantaggio degli autocrati e manderebbe in rovina le democrazie. La democrazia è una conversazione e le conversazioni si basano sul linguaggio. Quando l'AI hackera il linguaggio, potrebbe distruggere la nostra capacità di avere conversazioni significative, distruggendo così la democrazia».

Può essere che Harari ecceda in pessimismo. O che sottovaluti i rischi del rimanere indietro su quella che — come ha sottolineato anche Beppe Severgnini — è una rivoluzione, non una moda (rischi sui quali insiste spesso, sul Corriere, Massimo Sideri: ad esempio qui, qui e qui). Ma ci pare abbia, come minimo, dato più di uno stimolo alle meningi di chi voglia ancora tenere in esercizio la propria intelligenza. Senza l'«aiutino» di ChatGpt.