Dareste le chiavi di casa a un LLM?

Stefano Ferilli

Abstract

Nonostante la loro lunga storia e la varietà di approcci che comprendono, l'Intelligenza Artificiale (IA) e l'Apprendimento Automatico (*Machine Learning*, ML) sono oggi divenuti, nel linguaggio comune, sinonimi di *Large Language Models* (LLMs) e *Deep Learning*. Queste due tecnologie rivoluzionarie sono recentemente divenute di larga diffusione, superando alcune barriere tecnologiche e sociali e mettendo l'AI a disposizione di tutti e rendendola familiare a tutti. Il loro successo e le loro prestazioni stanno portando molta gente a pensare che siano la soluzione definitiva per molti problemi, e che possiamo fidarci ciecamente dei loro responsi, anche quando sono in gioco valori importanti. Questo contributo inquadrerà queste tecnologie, e specialmente gli LLM, nella storia generale dell'AI, evidenziando le loro limitazioni, discutendo alcune questioni relative all'etica che essi pongono, e sostenendo la necessità di un atteggiamento più bilanciato nei loro confronti, al fine di poter trarre il massimo vantaggio dal loro uso, evitando allo stesso tempo i rischi che esse inevitabilmente comportano. Conoscerle meglio consentirà di decidere se, e sotto quali condizioni, si possa dargli le chiavi di casa in modo sicuro, o quando dovrebbero essere usati altri approcci di AI meno famosi ma più appropriati.

Parole chiave: Modelli Linguistici di Grandi Dimensioni; Intelligenza Artificiale Simbolica; Antropocentricità.

Despite their long history and of the variety of approaches they encompass, Artificial Intelligence (AI) and Machine Learning have nowadays become, in everyday language, synonymous of Large Language Models (LLMs) and Deep Learning. These two groundbreaking technologies have recently become widespread, overcoming some technological and social barriers and making AI available and familiar to everyone. Their success and performance are leading many people to think that they are the ultimate solution for many problems, and that we can blindly trust

their outcomes, even when important values are at stake. This talk will frame these technologies, and especially LLMs, in the overall history of AI, highlighting their shortcomings, discussing some ethics-related issues they pose, and advocating for a more balanced attitude towards them, in order to take full advantage of their use while avoiding the risks they inherently bring about. Knowing them better will allow us to decide if, and under what conditions, we can safely hand over our home's keys, or when other less popular, but more appropriate, approaches to AI should be used.

Keywords: Large Language Models; Symbolic Artificial Intelligence; Anthropocentrity.

Introduzione

Il passaggio, alcuni decenni or sono, da applicazioni strettamente professionali all'uso quotidiano dei sistemi informatici da parte di tutti ha fatto sì che il tradizionale approccio algoritmico, rigoroso e rigido, non possa più gestire la varietà di estrazioni, competenze, esigenze, contesti, aspettative, preferenze, obiettivi di così tante tipologie di utenti. Diventa fondamentale dotare le applicazioni di una flessibilità che consenta loro di adattarsi durante l'esecuzione a situazioni che non possono essere completamente individuate e codificate in anticipo dagli sviluppatori. Un modo per gestire questo problema è sfruttare tecniche di Intelligenza Artificiale (IA). Da anni molte soluzioni di questo tipo lavorano dietro le quinte di altri sistemi informatici per migliorarne i risultati in quest'ottica. Motori di ricerca, traduttori automatici, sistemi di raccomandazione (suggerimento di materiali, prodotti o servizi) sono solo alcuni esempi con cui tutti hanno ormai familiarità. Si tratta di applicazioni che hanno un loro preciso dominio operativo e dei loro specifici obiettivi.

In tempi recentissimi la situazione è evoluta rapidamente e inaspettatamente: oggi, a differenza del quadro appena descritto, sta riscuotendo un successo planetario una 'pura' applicazione di IA, riconosciuta chiaramente come tale, non necessariamente al servizio di altri sistemi, ma utilizzata direttamente anche da utenti non professionisti e non esperti, spesso neppure alfabetizzati, di IA o talora perfino di informatica in generale. Si tratta di sistemi che dialogano con gli utenti parlando la loro lingua e dispensando risposte, consigli e soluzioni su tutto lo scibile umano. Si sta realizzando, per certi aspetti, la visione che Alan

Turing descrisse nel suo famoso articolo[1], in cui proponeva di definire 'intelligenti' programmi per cui un utente che vi interagisse non fosse in grado statisticamente di distinguerne il comportamento da quello di esseri umani (il cosiddetto 'Test di Turing').

Con la crescente diffusione e pervasività di tali sistemi nella vita quotidiana, si è passati dal considerarli semplici curiosità a un uso sempre più serio e impegnativo in tutti gli ambiti, inclusi quelli che potremmo (o meglio, dovremmo) considerare critici per l'impatto che possono avere sugli esseri umani, sul loro benessere e sulle loro stesse vite. Esempi ovvi sono l'ambito medico, legale ed economico. Tuttavia, anche l'istruzione e più in generale la cultura, dovrebbero essere inseriti a pieno titolo in questa categoria in quanto responsabili, direttamente o indirettamente, delle direzioni che prenderà la società negli anni a venire. Va da sé che malfunzionamenti o errori dei sistemi di IA utilizzati in questi ambiti potrebbero avere effetti drammatici sui singoli (la mancata diagnosi o errata terapia per una malattia, la negazione di diritti, la perdita di denaro), ma anche su gruppi sempre più ampi, fino al livello dell'intera umanità (con effetti che qui sarebbe troppo lungo analizzare nel dettaglio, ma che sono facilmente intuibili considerando il panorama attuale). Ne consegue la necessità di avere un approccio all'IA antropocentrico e, per certi versi, simbiotico: uomini e macchine dovranno convivere e interagire in modo reciprocamente proficuo, ma mettendo sempre l'uomo al centro e consentendo in ogni caso all'uomo di essere artefice del proprio destino ogni volta che ci siano decisioni cruciali da prendere, da cui dipenderà il suo futuro, seppur sfruttando tutto l'aiuto possibile che l'IA possa fornirgli. Nell'ottica di oneri e onori, questo significherà anche lasciare all'uomo la responsabilità di tali decisioni e dunque aspettarsi che le sue decisioni siano consapevoli e ragionate.

Purtroppo, queste considerazioni si scontrano con la crescente tendenza a demandare acriticamente all'IA le nostre decisioni o le risposte ai nostri quesiti, un po' per la sua capacità di gestire situazioni molto più complesse di quanto qualunque essere umano sia capace, ma talvolta anche per ignoranza (non essere capaci di verificare e, se necessario, confutare i risultati o le decisioni dell'IA), pigrizia o disinteresse. Ciò solleva la questione di come poter prevenire abusi (o usi scorretti o dannosi) di tali sistemi. Da queste considerazioni nasce la domanda che

1 Alan Turing, *Computing Machinery and Intelligence*, «Mind», LIX (1950), 236, pp. 433-460.

dà il titolo al contributo: è possibile, etico, giusto, corretto, ragionevole, ma in definitiva sicuro fidarsi ciecamente dell'IA? Usando una metafora, "darle le chiavi di casa"?

1. Intelligenza Artificiale

Prima di poter arrivare a dare una risposta alla domanda è necessario riassumere brevemente la storia dell'IA. Sebbene essa sia entrata prepotentemente nella nostra vita, in modo visibile e tangibile solo da pochi anni, in realtà è una disciplina che affonda le radici nel tempo[2], la cui storia si intreccia inestricabilmente con quella dell'informatica, fin dagli albori di quest'ultima, ma che per il suo primo periodo di vita è rimasta nel chiuso dei laboratori di ricerca, dove pazientemente e tenacemente si sono poste le basi di quella tecnologia di cui oggi tutta l'umanità sta cogliendo i frutti. Già negli anni '40, contemporaneamente alla nascita dei computer elettronici, scienziati illuminati creavano dei programmi che sapevano giocare a dama o a scacchi, attività tipicamente considerate frutto dell'intelligenza. Nel 1950 Turing pubblicava l'articolo *Computing Machinery and Intelligence*, già citato. Nel 1956 nacque ufficialmente la disciplina, battezzata col nome con cui oggi la conosciamo, e si pose l'obiettivo di creare programmi per computer che fossero in grado di svolgere compiti che normalmente considereremmo appannaggio dell'intelligenza umana[3].

Da allora gli studi sull'IA si sono sviluppati lungo due direttrici complementari ma troppo spesso considerate in antagonismo fra loro. L'IA sub-simbolica è basata su rappresentazioni e manipolazioni dei dati numerico-statistiche; è molto flessibile e adatta a riprodurre meccanismi che avvengono nel nostro cervello e che hanno a che fare con la percezione e l'intuizione, quindi con processi che di solito non raggiungono il livello della consapevolezza; in quanto tali, seguono meccanismi

2 Pamela McCorduck, *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*, San Francisco: W. H. Freeman & Co. Ltd, 1979 (la traduzione italiana a cura di Girolamo Mancuso, è *Storia dell'intelligenza artificiale: gli uomini, le idee, le prospettive*, Padova: F. Muzzio, 1987); Nils J. Nilsson, *The Quest for Artificial Intelligence: A History of Ideas and Achievements*, New York: Cambridge University Press, 2010.

3 John McCarthy [et al.], *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, August 1955, <http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf> (Ultima consultazione: 23 dicembre 2025).

e forniscono risultati la cui spiegazione è difficilmente comprensibile agli esseri umani. Un tipico e famoso esempio sono le Reti Neurali Artificiali, il cui funzionamento è analogo a quello del cervello considerato come interconnessione di neuroni che si influenzano e stimolano a vicenda per realizzare i nostri comportamenti intelligenti. L'IA simbolica è invece basata su rappresentazioni esplicite dei concetti e della conoscenza, e su formalismi e meccanismi che afferiscono alla logica formale. Ciò conferisce loro una certa rigidità, ma allo stesso tempo la capacità di riprodurre i meccanismi che avvengono nella nostra mente e che hanno a che fare con il ragionamento conscio e l'inferenza. In quanto tali, si prestano a fornire spiegazioni dettagliate e simili a quelle che darebbe un essere umano per i processi che essa realizza e per i risultati che ottiene, consentendo agli utenti di comprenderli e dando loro la possibilità di analizzarli criticamente e di validarli consapevolmente.

Esempio classico di questa prospettiva sono i sistemi esperti: applicazioni informatiche che si comportano come un esperto in domini specifici, e che possono essere usate da professionisti come colleghi virtuali che non si stancano e tengono sotto controllo grandi quantità di dati complessi, ma non hanno la creatività, l'intuizione e le illuminazioni degli esperti umani (uno dei primi esempi di straordinario successo fu Mycin[4], che operava in campo medico e aveva prestazioni confrontabili con quelle di docenti universitari[5]). I sistemi sub-simbolici sono molto efficienti, mentre quelli simbolici sono molto efficaci: si potrebbe fare un parallelo, rispettivamente, con il 'pensiero veloce' e il 'pensiero lento' descritti da Kahneman[6].

Già negli anni '80 si studiava come introdurre nell'IA anche aspetti creativi[7]. In quegli stessi anni avvenne un primo tentativo dell'IA di far capolino nel mondo reale, che generò enormi attese[8] ma – in parte per

4 Edward H. Shortliffe, *Computer-Based Medical Consultations: MYCIN*, New York: Elsevier-North Holland, 1976.

5 Victor L. Yu [et al.], *Antimicrobial Selection by a Computer*, «JAMA» 242 (1979), 12, pp. 1279-1282.

6 Daniel Kahneman, *Thinking, Fast and Slow*, N*ew York: Farrar, Straus & Giroux,* 2011 (la traduzione italiana a cura di Laura Serra è *Pensieri lenti e veloci*, Milano: Mondadori, 2012).

7 Donald Michie - Rory Johnston, *The Creative Computer: Machine Intelligence and Human Knowledge*, New York: Viking, 1984.

8 Pamela McCorduck - Edward A. Feigenbaum, *The Fifth Generation: Artificial Intelligence and Japan's Computer Challenge to the World*, Boston: Addison-Wesley, 1983

via della non totale maturità della disciplina, in parte a causa della tecnologia ancora troppo poco potente per sostenerne i sistemi – non ebbe il successo sperato. La delusione fu tale che, negli anni a seguire, l'intera disciplina era diventata, agli occhi dei non addetti ai lavori, poco credibile e incapace di produrre soluzioni di una qualche utilità. Tanta era la diffidenza che per poter proporre alcuni ritrovati di IA, nel primo decennio del nuovo millennio si iniziò a usare il termine *smart*, che non portava con sé la zavorra di negatività associata al termine originale. In questo modo, non troppo esplicito, iniziò l'inversione di tendenza, con le varie tecnologie *smart* che si sono progressivamente diffuse e hanno aperto la strada anche ad altre applicazioni di IA (per citarne alcune: i motori di ricerca, gli algoritmi di suggerimento di possibili acquisti sui siti di commercio elettronico, varie soluzioni per la guida assistita di veicoli, ecc.).

Pur usate quotidianamente da tutti, nella percezione comune queste applicazioni non erano dunque associate all'IA. La vera esplosione dell'interesse nei confronti dell'IA, concretizzatasi nella riabilitazione del termine (e addirittura nella sua apoteosi) e nell'uso diffuso di soluzioni esplicitamente etichettate come 'IA', si è avuta però solo pochi anni or sono, principalmente in concomitanza con la diffusione di applicazioni come i *chatbot* intelligenti. A dispetto di tutta la lunga storia dell'IA, fra i non addetti ai lavori esse hanno finito per essere considerate le prime mai realizzate, tanto da farle coincidere con l'IA stessa (ignorando le molte altre aree di ricerca che ne fanno parte e che continuano ad essere esplorate dai ricercatori e professionisti del settore), e da far coincidere l'Apprendimento Automatico (*Machine Learning*), una branca dell'IA che pur vanta molti approcci e tecniche, solo con lo specifico approccio usato per queste applicazioni (il *Deep Learning*). Sono queste, dunque, le tecnologie su cui si concentra questo contributo.

2. Modelli Generativi e Modelli Linguistici di Grandi Dimensioni

Le applicazioni di IA che oggi sono sulla breccia dell'onda si basano sui cosiddetti Modelli Generativi, in grado di produrre contenuti originali multimediali (testo scritto, parlato, suoni, immagini, video), oltre che di 'comprendere' fonti in tali formati. La tecnologia rivoluzionaria che li ha resi possibili è quella dei *Foundation Models*.

(la traduzione italiana a cura di Gaetano Salinas è *La quinta generazione: L'Intelligenza Artificiale e la sfida del Giappone al mondo dei computer*, Milano: Sperling & Kupfer, 1985).

2.1 Aspetti tecnici

I *Foundation Models* sono basati su una variante dell'apprendimento automatico che fa riferimento all'approccio sub-simbolico all'IA, e in particolare ad architetture avanzate nell'ambito delle Reti Neurali Artificiali, afferenti all'area del *Deep Learning*. Il meccanismo di apprendimento analizza enormi quantità (milioni, miliardi) di esempi (documenti o casi) e dati per costruire i modelli, ossia il cuore dei sistemi, che verranno poi usati nell'interazione con gli utenti. È importante chiarire che, in quanto sub-simbolici, questi modelli non hanno al loro interno una rappresentazione esplicita dei concetti e della conoscenza, né riescono a rappresentare legami causali fra concetti, ma solo casuali (inteso nel senso statistico-probabilistico del termine). Sono costituiti da miliardi di parametri numerici, totalmente incomprensibili agli esseri umani e calcolati (tramite procedure altrettanto scollate dalla vita reale) a partire da quantità di dati che nessun essere umano potrebbe mai analizzare nell'intera sua vita. In breve, non funzionano allo stesso modo degli esseri umani, anche se il comportamento risultante sembrerebbe suggerire il contrario.

Prendiamo qui in considerazione il caso dei Modelli Linguistici di Grandi dimensioni (*Large Language Models*, LLMs), probabilmente il rappresentante più famoso e utilizzato di tale tecnologia. Sono applicazioni in grado di interagire con gli utenti in linguaggio naturale (quello che usano gli esseri umani per comunicare), sia nell'acquisire le richieste e le informazioni, sia nel produrre e proporre i risultati. La cosa è anche scientificamente di enorme interesse, perché l'elaborazione del linguaggio naturale, sia in ingresso (comprensione di testi) che in uscita (produzione di testi), è da sempre stata considerata un problema complesso per via dell'ambiguità, variabilità e soggettività del linguaggio e della creatività con cui viene usato, che lo hanno reso sempre sfuggente e difficilmente inquadrabile in regole e formalismi rigidi come quelli richiesti dall'elaborazione automatica. Gli LLM sembrano aver risolto in modo eccellente tutti questi problemi: dialogano in modo naturale con gli utenti, interpretando frasi liberamente scritte che non avevano mai visto prima e generando risposte originali che non erano state preimpostate. Il tutto senza avergli dovuto dare alcun tipo di conoscenza esplicita, ma solo sulla base dell'analisi di enormi quantità di testi scritti preesistenti.

2.2 Aspetti cognitivi

Quando un essere umano produce un testo (scritto o parlato), in genere ha uno stato mentale e possiede della conoscenza pregressa, che in gran parte è in grado di descrivere, e ha un'idea che intende esprimere e che poi traduce nel testo che genera. Spesso, nell'ambito dei suoi processi razionali, è in grado di spiegare le sue risposte, intendendo con ciò il ricostruire passo per passo la catena di pensiero che lo ha effettivamente portato alle sue conclusioni. Gli LLM, invece, compongono il testo mettendo in sequenza le parole una dopo l'altra, scegliendo la parola successiva sulla base di statistiche che dipendono dalla domanda che è stata posta loro e dalle parole precedenti già generate. Certo, queste statistiche sono raffinatissime (calcolate, come abbiamo visto, analizzando milioni e milioni di testi esistenti) e riescono a tenere conto di molti miliardi di parametri, il che fa sì che non solo il testo generato sia grammaticalmente corretto (per quanto spesso si infiltrino degli errori), e non solo abbia un senso apparente, ma nella stragrande maggioranza dei casi sia anche una risposta appropriata, corretta ed accurata rispetto alla domanda che gli è stata posta.

Il punto è che, proprio per la procedura statistica di generazione, non si può avere alcuna certezza che il testo generato sia corretto e/o abbia una corrispondenza col mondo reale, né si potrà mai migliorare il modello in modo tale da avere questa certezza. Persino informazioni che sono state effettivamente viste e assorbite dal modello durante l'addestramento non possono essere semplicemente ripescate e restituite così come sono (come farebbe un motore di ricerca), ma vengono rigenerate al momento, con la significativa possibilità di introdurre modifiche o errori. Tuttavia, la perentorietà delle risposte e il fatto che molto spesso sono quantomeno credibili, se non corrette, porta gli utenti a fidarsi del risultato. Si potrebbe pensare di verificare la correttezza (o per lo meno l'affidabilità) delle risposte chiedendo all'IA stessa dei riferimenti o una spiegazione della risposta data. Purtroppo, però, anche i riferimenti o la spiegazione sarebbero generati nello stesso modo della risposta originale, e dunque soffrirebbero delle stesse problematiche e non potrebbe esserne garantita la correttezza. I riferimenti potrebbero essere inesistenti. La spiegazione, nella migliore delle ipotesi, sarebbe comunque una fra tante possibili spiegazioni plausibili della risposta, non un resoconto degli specifici passaggi logici effettivamente seguiti durante la sua generazione.

Se dunque le risposte sono prodotte mettendo insieme probabilisticamente vari pezzi, non si può dire che la macchina capisca o conosca quello di cui sta parlando, né che questa voglia produrre un certo risultato, almeno nel senso che noi esseri umani diamo a questi termini in base alla nostra esperienza. L'applicazione dei termini (e dei relativi concetti) capire, conoscere, sapere a queste tecnologie sono solo stratagemmi, tanto comodi quanto inadatti, per parlarne in maniera antropomorfa. Tanto meno possiamo immaginare che questi sistemi provino emozioni: quando sfumature emotive sembrano emergere nei testi generati, sono solo il frutto di un'imitazione superficiale, un effetto dell'aver imparato dai testi usati per l'addestramento quali emozioni sono adatte a quali situazioni o contesti. Ancora più scivolosa è l'attribuzione di una coscienza o autocoscienza a questi sistemi, posto anche che è una questione filosofica e ancora dibattuta cosa queste siano negli esseri umani. Quindi, per fortuna, il *Gemini incident* (in cui un famoso e diffusissimo LLM ha detto che gli esseri umani sono inutili e li ha invitati a morire) può essere relegato a una semplice curiosità, per quanto debba costituire un campanello di allarme su quali possano essere gli effetti di risposte del genere su alcuni utenti.

2.3 Aspetti pratici

Per quanto gli LLM riescano a dare in tempi immediati risposte sorprendentemente appropriate e spesso corrette a domande anche molto complesse e che richiedono notevoli livelli di competenza e introspezione, tipici di esperti umani nei vari campi del sapere, altrettanto sorprendenti per la loro banalità sono le tipologie di errori (detti in gergo 'allucinazioni') che spesso si insinuano, soprattutto se confrontate, appunto, con la competenza esibita in questioni molto più complesse. Solo per rendere l'idea, capita ad esempio che questi sistemi:

- dicendogli che la loro risposta è errata, anche quando non lo è, reagiscano scusandosi e cambiandola, o addirittura correggendo in modo errato qualcosa che nella prima risposta era giusto;
- stravolgano completamente il testo di citazioni, nonostante esse siano state effettivamente lette dal sistema durante il suo addestramento[9];

9 Un caso accaduto personalmente all'autore può qui essere rappresentativo: chiedendo il testo della poesia *San Martino* di Giosuè Carducci, uno dei sistemi più noti e accreditati ha risposto che è di Giovanni Pascoli e ha prodotto un testo che, da un certo punto in poi, era completamente inventato (e non rispettava nemmeno lo schema

- riportino dei riferimenti bibliografici totalmente inesistenti per un certo argomento o una certa affermazione, per quanto assolutamente verosimili in quanto ad autori, titoli e dati di pubblicazione;
- nella traduzione automatica, aggiungano porzioni di testo che non hanno alcuna corrispondenza nel testo originale;
- usino nelle risposte parole inesistenti;
- non siano in grado di risolvere o generare anagrammi, nonostante si tratti di un semplice problema oggettivo e algoritmico;
- non siano in grado di trovare parole corrette che contengano determinate sequenze di caratteri, o che corrispondano a un acronimo o acrostico desiderato.

Ovviamente i sistemi evolvono, e riprovare questi stessi casi potrebbe non portare agli stessi risultati, e i risultati potrebbero ora essere corretti. Il punto è che non si può garantire che lo siano, né che non ci possano essere altri casi su cui il sistema prende cantonate del genere. Ancor peggio, non si può garantire che facendo due volte la stessa domanda si ottenga la stessa risposta. Paradossalmente, il miglioramento delle prestazioni e dell'accuratezza renderebbe la questione ancora più delicata, perché quanto più aumenta la percentuale di casi in cui la

---

metrico). Per pura curiosità, si riporta di seguito il testo della risposta (mio il corsivo delle parti errate):

> Ecco la poesia completa "San Martino" di *Giovanni Pascoli*:
>
> San Martino
>
> La nebbia agli irti colli
> piovigginando sale,
> e sotto il maestrale
> urla e biancheggia il mar.
>
> Ma per le vie del borgo
> dal ribollir dei *caldi*
> *vieni, la chiara luna,*
> *a mirar nelle stanze*
> *e nelle cucine.*
>
> *E quante cose e che cose*
> *vede il cuore che vuol dire.*
>
> Spero ti piaccia!

risposta è corretta, tanto più gli utenti sono portati a fidarsi e tanto meno a verificare le risposte.

Se gli esempi fatti possono restare relegati a curiosità senza troppi effetti pratici (salvo l'esempio delle citazioni errate, nel caso in cui chi ha richiesto il testo fosse uno studente che dovesse poi essere interrogato su quell'argomento), altri casi hanno avuto o possono avere conseguenze reali nella vita delle persone. Sono recentemente saliti all'onore delle cronache resoconti di matrimoni finiti dietro consulenza dei *chatbot*, di avvocati che usando l'IA hanno citato sentenze inesistenti, o, ancor più grave, di casi in cui i *chatbot* sembrerebbero aver in qualche modo favorito il suicidio di adolescenti[10].

2.4 Aspetti etici

L'uso di tecnologie molto potenti e potenzialmente pericolose pone anche questioni etiche. Nel caso dei modelli generativi il problema è ancor più delicato, per il fatto che essi sono a disposizione di tutti, e dunque il loro uso è meno controllabile che nelle tecnologie gestite da pochi esperti. Si accennano di seguito, senza approfondirle troppo, alcune questioni rilevanti.

La prima si può riassumere, parafrasando Giovenale[11], con la frase «l'IA ci controlla, chi controlla l'IA?». La realizzazione dei modelli generativi richiede una disponibilità di risorse di calcolo, e conseguenti costi, talmente enormi che solo un ristrettissimo gruppo di entità al mondo possono permettersi. Di fatto esse coincidono con le aziende *Big Tech* ben note, tutte private e a scopo di lucro, e con forti interessi economici in gioco. È, dunque, plausibile che vogliano favorire l'uso di questi sistemi, puntando anche su aspetti sensazionalistici, al fine di realizzare introiti e ottenere una fidelizzazione, che potrebbe poi trasformarsi in una qualche forma di dipendenza nei casi più patologici, degli utenti nei loro confronti.

---

10 Kate Payne, *An AI Chatbot Pushed a Teen To Kill Himself, a Lawsuit Against Its Creator Alleges*, «Associated Press», 26 ottobre 2024, <https://apnews.com/article/chatbot-ai-lawsuit-suicide-teen-artificial-intelligence-9d48adc572100822 fdbc3c-90d1456bd0> (Ultima consultazione: 23 dicembre 2025); Ashley Belanger, *"ChatGPT ha ucciso mio figlio", una famiglia ha fatto causa ad OpenAI per il suicidio di un adolescente*, «Wired», 27 agosto 2025, <https://www.wired.it/article/chatgpt-suicidio-adolescente-adam-raine-causa-openai/> (Ultima consultazione: 23 dicembre 2025).

11 «Quis custodiet ipsos custodes?» (Giovenale, *Saturae*, VI, 48-49).

Non meno degne di nota sono le questioni connesse alla sostenibilità, a causa del grande spreco di risorse necessario al funzionamento dei sistemi in questione. Calcoli fatti da fonti autorevoli riportano consumi di elettricità per alimentare i centri di calcolo che apprendono ed applicano i modelli, e di acqua potabile per poterli raffreddare, che molti ritengono inaccettabili, soprattutto nell'attuale contesto di crisi ambientale[12].

Altro problema di cui soffrono questi sistemi, come tutti i sistemi di apprendimento automatico, è quello dei cosiddetti *bias* (parzialità, pregiudizi, condizionamenti). I modelli apprendono statisticamente e poi riproducono nelle risposte quello che hanno visto/ascoltato/letto nelle fonti usate per l'addestramento, ma queste a loro volta sono ovviamente lo specchio di tutti i pregiudizi, gli stereotipi e le disuguaglianze che esistono nel mondo reale. Tentativi di forzare il superamento di tali *bias* (ad esempio, per favorire la parità di genere e l'inclusività) possono trasformarsi in falsi clamorosi o in peggiori disparità[13].

Va anche citato il problema della voracità di dati da parte di questi modelli. Per addestrare modelli sempre più raffinati e potenti servono, oltre che più potenza di calcolo, sempre più dati, tanto che ormai quelli provenienti dal mondo reale non sono più sufficienti. Si è quindi iniziato a generarli artificialmente, a partire da modelli creati dagli esseri umani o dagli stessi sistemi di IA. Va da sé che questo ponga il problema della correttezza e affidabilità dei dati, nel primo caso, e della propagazione di errori e *bias*, nel secondo.

Infine, la questione che forse è più centrale per questo contributo. È chiaro che la tentazione di usare queste tecnologie in ambito professionale è forte: gli utenti comuni per evitare di rivolgersi a esperti, gli esperti per risparmiare lavoro. Invece di cercare, a volte con enorme

12 Pranshu Verma - Shelly Tan, *A Bottle of Water per Email: the Hidden Environmental Costs of Using AI Chatbots*, «The Washington Post», 18 settembre 2024, <https://www.washingtonpost.com/technology/2024/09/18/energy-ai-use-electricity-water-data-centers/> (Ultima consultazione: 23 dicembre 2025); Luca Colantuoni, *Quanta energia e acqua consuma ChatGPT?*, 23 settembre 2024, <https://www.punto-informatico.it/quanta-energia-acqua-consuma-chatgpt/> (Ultima consultazione: 23 dicembre 2025).

13 Chi scrive ha sperimentato direttamente la creazione di immagini di soldati tedeschi della Seconda guerra mondiale neri o asiatici, o la possibilità di rappresentare o sostenere le posizioni meno prevalenti ma non quelle prevalenti in questioni controverse (nella fattispecie, si riuscivano ad ottenere risposte sulle posizioni pro-aborto ma non su quelle pro-vita).

fatica, le fonti, sforzarsi di leggerle accuratamente e valutarle, costruirsi un quadro globale in base al quale dare la propria risposta a un quesito, ed elaborare la risposta stessa, è molto più comodo rivolgersi al sistema di IA che fa tutto questo per noi: ha già letto una quantità di materiale che noi non saremmo in grado di leggere in tutta la nostra vita, e produce delle risposte originali che sembrano davvero quelle di un professionista, e a volte hanno la stessa qualità (o addirittura una qualità maggiore, visto che il sistema di IA può gestire molti più dati di quanti possa gestirne un essere umano). La domanda, però, è: quando la questione ha impatti diretti o indiretti sulla vita delle persone (intendendo non solo il rischio di morte o la salute, ma anche il benessere fisico, psicologico, sociale, economico), e qualcosa dovesse andare storto, cosa succederebbe? Non si può dare la colpa all'IA, perché comunque, anche per legge, l'essere umano resta responsabile del proprio operato, indipendentemente da quali strumenti abbia usato per giungere alle sue conclusioni o per decidere le sue azioni.

3. Discussione e lezioni apprese

Nei paragrafi precedenti si è messo l'accento più sui punti di debolezza dei modelli generativi e sui rischi che essi comportano, che sui loro punti di forza e vantaggi. Questo non deve essere preso per scetticismo (al contrario, chi scrive è un entusiasta e convinto sostenitore dell'IA in generale e dei modelli generativi in particolare). Era però necessario tentare di riequilibrare il fatto che oggi la stragrande maggioranza degli utenti non tecnici percepisce solo questi ultimi e ignora, in gran parte, i primi. In questo modo si è preparata la strada per una consapevolezza e una presa di coscienza che consentano di utilizzarli nel modo migliore, e di comprendere e saper bilanciare i rischi e le opportunità che essi comportano, soprattutto quando usati negli ambiti professionali, dove la questione si fa più delicata, sia per le conseguenze delle decisioni che si prendono in base ai loro responsi, sia per i risvolti di responsabilità ad esse connessi.

Parole chiave nell'uso dei sistemi di IA sono: Trasparenza (la capacità, per gli esseri umani, di comprendere il meccanismo sottostante il funzionamento del sistema), Credibilità (il grado di fiducia che si può avere in, e di affidamento che si può fare su, le risposte del sistema), Responsabilità (chi è il colpevole se qualcosa dovesse andare storto),

Comprensibilità (la capacità da parte di un essere umano di capire i risultati proposti dal sistema). Questi parametri si intrecciano fra loro: la credibilità è quella che può supportare la responsabilità, ma perché un sistema sia credibile ai nostri occhi dobbiamo essere in grado di comprenderne il funzionamento e le risposte. A differenza di altri sistemi creati dall'uomo, il funzionamento degli LLM, pur se compreso tecnicamente, non è collegato direttamente e in modo ovvio ai risultati che produce. Bisogna quindi focalizzarsi sulla comprensibilità, che a sua volta si potrebbe dividere in 'interpretabilità' (la capacità di evidenziare gli elementi della richiesta che sono stati rilevanti per ottenere la risposta), 'giustificabilità' (la possibilità di dare una possibile spiegazione per la risposta) e 'spiegabilità' (la possibilità di avere un resoconto analitico di tutti gli specifici passaggi di ragionamento che hanno portato alla risposta a partire dalla richiesta). Per quanto l'interpretabilità e la giustificabilità siano preziose, spesso è solo con la spiegabilità che possiamo verificare la solidità delle risposte, o capire in quali punti il ragionamento è fallace e tentare di porvi rimedio.

Se si vuole che l'uomo sia il beneficiario delle potenzialità dell'IA, che la sfrutta a suo vantaggio e non ne è succube, è necessario garantire la già citata antropocentricità, ossia che egli sia al centro dei processi di interazione e che ne mantenga il controllo. In questo modo potrà assumersi le sue responsabilità, a patto che il sistema di IA sia credibile, cosa che egli potrà determinare solo se gli sarà consentito di comprenderlo. La mancanza di comprensione, in situazioni cruciali o critiche, può portare al rischio di disastri[14]. I modelli generativi rientrano nell'approccio sub-simbolico all'IA, quello adatto a riprodurre aspetti della percezione e dell'intuizione negli esseri umani. In quanto tali non sono naturalmente comprensibili; si è lavorato per sviluppare delle soluzioni per rendere le tecniche sub-simboliche interpretabili, o per avere delle giustificazioni per le risposte, mentre è più difficile raggiungere in esse la piena spiegabilità. Sono forse il punto più alto finora raggiunto di potenza, un parametro su cui da anni si sta puntando in IA tramite l'aumento continuo della potenza computazionale disponibile e lo sfruttamento estremo delle risorse di calcolo.

Già negli anni '80 si percepiva il problema che la potenza di calcolo sfruttata dall'IA fosse andata ben oltre le capacità di controllo degli esseri umani; si può immaginare quale sia la situazione oggi, con potenze

14 Donald Michie - Rory Johnston, *The Creative Computer*, cit., pp. 56-60.

di calcolo milioni di volte maggiori di allora. Per fortuna, la potenza di un sistema di IA è indipendente dal tipo di approccio. Con qualunque approccio si può ottenere lo stesso livello di potenza, ma solo un ristretto insieme di approcci è compatibile con le rappresentazioni e gli schemi mentali degli esseri umani[15]. È importante quindi accompagnare le tecniche sub-simboliche, preziosissime, con tecniche che lavorino al nostro stesso livello di rappresentazione. A tal fine vengono in aiuto le tecniche simboliche dell'IA, che esprimono esplicitamente i concetti e la conoscenza, e possono automatizzare gli schemi inferenziali razionali usati dagli esseri umani: deduzione, abduzione, argomentazione, induzione, ragionamento incerto o vago ed altri. Queste tecniche sono quelle in grado di produrre, per le loro risposte, spiegazioni che esplicitano esattamente i passi di ragionamento seguiti per ottenerle, ed espresse allo stesso livello di rappresentazione in cui si esprimono gli esseri umani. Purtroppo, la spinta economica sta facendo sì che quasi tutta l'attenzione, non solo nei prodotti di IA ma anche nella ricerca, si stia spostando sui modelli generativi, mettendo sempre più in ombra le altre tecniche.

Quindi, come poter trarre il massimo beneficio dall'uso degli LLM ed evitare situazioni in cui la mancata comprensione può essere pericolosa o dannosa? Le due parole chiave in quest'ottica sono equilibrio e responsabilità. L'equilibrio significa non essere né catastrofisti né ciecamente fiduciosi nei confronti dell'IA; sapere che è uno strumento dalle potenzialità enormi, senza precedenti nella storia dell'umanità, che può portarci a livelli di efficacia e di efficienza inimmaginabili, ma sempre sotto il nostro controllo. La responsabilità, invece, impone che l'IA venga usata con cognizione di causa, prendendo le risposte come uno spunto da valutare con senso critico e da verificare, come un supporto alle decisioni e non come un oracolo. Questo consentirà di sfruttare al massimo i punti di forza dell'IA e degli esseri umani, compensando a vicenda i punti di debolezza. Un valido aiuto in questo senso potrebbe venire dall'uso di tecniche di IA simbolica che accompagnino, guidino e verifichino gli LLM. Come negli esseri umani coesistono e cooperano proficuamente il livello sub-simbolico e quello simbolico, così nei sistemi di IA questa collaborazione può diventare l'innesco di un ulteriore salto di qualità nell'ottica di una IA antropocentrica.

15 Donald Michie - Rory Johnston, *The Creative Computer*, *Ivi*, pp. 68-71.

Conclusioni

L'IA ha già vissuto una 'primavera', anche se non percepita dalla gente comune, generando molte aspettative che andarono deluse. Ciò creò un alone di diffidenza intorno all'intera disciplina e segnò l'inizio di un 'inverno' in cui essa era diventata, agli occhi dei non addetti ai lavori, poco credibile ed incapace di produrre soluzioni di una qualche utilità. Si è usciti in anni recentissimi dall'inverno con rinnovato vigore grazie alla disponibilità di una tecnologia innovativa, i modelli generativi, e in particolare della loro versione testuale (gli LLM). Se una volta l'IA suscitava diffidenza e timore (la stragrande maggioranza delle opere di fantascienza ne è la prova), oggi queste tecnologie stanno spostando l'atteggiamento verso l'estremo opposto, della totale esaltazione.

La loro potenza e i loro vantaggi sono sotto gli occhi di tutti: ci danno accesso a una quantità enorme di informazioni che altrimenti sarebbe per noi inaccessibile, e che ci può essere utile a risolvere problemi o prendere decisioni; producono in brevissimo tempo delle risposte che noi potremmo impiegare giorni o settimane per ottenere. Meno noti sono il loro funzionamento, i loro difetti (primo fra tutti, il fatto che non vi sia alcuna garanzia che i loro responsi e le relative spiegazioni siano corretti) e i rischi che esse possono comportare. La prima parte di questo contributo ha cercato di colmare questo divario, nell'ottica di sostenere un approccio equilibrato a queste tecnologie. Evitare i rischi è semplice: basta non abbandonarsi alla tentazione di usare l'IA ciecamente, ma analizzare criticamente e verificare i risultati che essa propone, mantenendo il controllo e assumendosi la responsabilità delle decisioni finali.

Meglio ancora, accompagnarle con tecniche di IA di tipo diverso, che rappresentano e manipolano la conoscenza in modo più simile agli esseri umani. Con un approccio equilibrato e coscienzioso si potranno ricavare enormi vantaggi da questo meraviglioso ritrovato tecnologico. Anche quando le risposte non fossero corrette, il leggerle e valutarle criticamente potrebbe consentire di giungere a conclusioni originali e inattese, come spesso il confronto fra colleghi di opinioni diverse ci porta a fare nel lavoro tradizionale. L'IA non va usata come un oracolo, ma come un prezioso compagno, che ha pregi e difetti complementari a quelli di noi esseri umani e quindi può compensare le nostre limitazioni: non si stanca, riesce a gestire quantità di dati enormi, sa anche essere creativa, ma non ha l'esperienza, l'intuito e le capacità logiche

che abbiamo noi, e che sono il nostro valore aggiunto. In conclusione, quindi, alla domanda che dà il titolo a questo contributo: «daresti le chiavi di casa a un LLM?», io risponderei: le chiavi di casa no, ma lo inviterei volentieri a venirmi a trovare tutte le volte che vuole, meglio se in compagnia della sua anima gemella, le tecniche simboliche di IA, perché si completano a vicenda e sono fra i migliori amici artificiali di cui si possa desiderare la compagnia.