INTELLIGENZA ARTIFICIALE _Francesca Rossi_ Laterza (2024)[1]

CAPITOLO 1: Viviamo già nel mondo dell'IA

I social network sono interamente supportati da tecniche di IA, senza le quali noi utenti vedremmo con ugual frequenza i contenuti prodotti da chiunque faccia parte dei nostri contatti, nonché pubblicità generiche, come quelle che vediamo in tv, e anche molti più contenuti indesiderati.

Tecniche di IA simili, ma basate esclusivamente sulle nostre azioni e non sui contenuti che creiamo, vengono usate sulle piattaforme per acquisti (come Amazon), per streaming di brani musicali (come Spotify), di video (come YouTube), o di film e serie tv (come Netflix). Gli oggetti che compriamo, i film o video che guardiamo, le canzoni che ascoltiamo, insieme alle eventuali valutazioni che ne diamo, vengono usati dall'IA per categorizzarci e poi proporci nuovi prodotti e contenuti scelti da utenti con abitudini simili alle nostre (quante volte leggiamo "Potrebbe piacerti anche questo film…"?).

Per cominciare a diradare la confusione sull'IA, dobbiamo intanto renderci conto che questa tecnologia è presente nelle nostre vite da molto tempo. Secondo uno studio della società di analisi e consulenza Gartner, usando l'IA generativa (cioè i sistemi IA come ChatGPT e DALL-E) le aziende hanno già ottimizzato i loro costi (17%), supportato la crescita del loro revenue (26%) e migliorato il rapporto con i clienti (38%). Gartner stima che entro il 2025 il 30% delle aziende userà l'IA per lo sviluppo e il test di nuove soluzioni, entro il 2026 verrà automatizzato il 60% del progetto di nuovi siti web, ed entro il 2027 il 15% delle nuove applicazioni sarà creata dall'IA senza l'intervento di persone. Inoltre, lo studio stima che entro il 2025 oltre il 30% di nuove medicine e materiali sarà scoperto usando l'IA generativa, e il 30% dei messaggi pubblicitari di grandi aziende sarà generato dall'IA.

ChatGPT ha sorpreso molti e soprattutto ha influenzato molto la percezione che abbiamo dell'IA. Oltre a essere trasformativo nelle sue tecniche, ChatGPT ha battuto molti record anche nella velocità di diffusione del suo uso. Dopo la sua pubblicazione a novembre 2022, ha raggiunto un milione di utenti in cinque giorni, e cento milioni di utenti a gennaio 2023. Altre app usate a livello globale hanno impiegato un tempo molto più lungo a diffondersi: a Instagram sono serviti due mesi e mezzo per avere un milione di utenti, e tre anni e mezzo a Netflix.

Con l'avvento di ChatGPT, invece, per la prima volta nella storia dell'IA tutti pensano che sia improvvisamente arrivata, che abbia fatto il suo ingresso trionfale o catastrofico, a seconda dei punti di vista, nella nostra quotidianità. Secondo un report stilato nel 2023 dall'International Federation of Robotics (IFR) sono soprattutto i paesi asiatici a usare i robot (73% nel 2022), seguiti molto da lontano dall'Europa (15%) e dal continente americano (10%). Le culture asiatiche vedono i robot con maggiore simpatia rispetto ad altri paesi. Perciò installazioni di robot in tutti i settori, compresi quelli a diretto contatto con i consumatori, come alberghi, aeroporti, o case private, sono accolte con entusiasmo maggiore di quello che può essere riscontrato in altri paesi del mondo.

Oltre all'hardware, cioè i pezzi fisici di un robot, quello che è importante sono i programmi che dicono alle componenti meccaniche come comportarsi. Senza questi programmi, il robot non saprebbe cosa fare. È qui che interviene l'IA per fornire al robot la capacità di agire in modo utile. Tramite i suoi sensori, come telecamere o sensori audio, un robot acquisisce dati, che poi l'IA elabora per prendere decisioni che andranno ad attivare gli attuatori del robot, cioè le varie parti fisiche come le mani, le braccia, le gambe, le ruote o qualunque altro pezzo che può svolgere un'azione nel mondo reale. Naturalmente i programmi e i pezzi fisici del macchinario devono fare i conti con i loro limiti: anche il software più avanzato non può aiutare un robot a spostarsi, se a questo mancano sensori precisi che gli permettano di capire dove si trova e dove andare. Lo stesso discorso vale se il robot ha i migliori sensori disponibili, ma l'IA all'interno del robot non è in grado di interpretare bene i dati raccolti. Quindi, tra l'IA e le componenti meccaniche di un robot deve esistere una sorta di simbiosi, che serve a creare un sistema capace di agire in modo utile e intelligente.

[1] Riassunto per ECCOCI! da Gigi Bacchetta. Segnalazione errori: gigi.bacchetta@cgilpiemonte.it

A creare paure collettive sono le attuali tecniche più avanzate dell'IA, quelle basate sulla cosiddetta IA generativa, che oltre a interpretare contenuti ne sanno generare di nuovi, come fa appunto ChatGPT. Ma questo accade perché non c'è chiarezza su reali capacità, limiti e rischi della tecnologia in questione, e gli utenti finali, cioè tutti i cittadini, non vengono messi nelle condizioni di comprendere a fondo quali sono le opportunità, i vantaggi e le possibili implicazioni del suo utilizzo.

Le auto a guida autonoma

Di solito un'auto viene usata dal suo proprietario solo il 5% del tempo e per il restante 95% rimane parcheggiata. I guidatori umani vengono inoltre facilmente distratti per vari motivi, come la stanchezza o il fatto di svolgere altre attività mentre sono al volante, che causano circa il 90% degli incidenti mortali. Un'auto a guida autonoma è in pratica un robot che sa seguire autonomamente le regole della strada e portare i suoi passeggeri a destinazione.

Oltre alle difficolta tecniche, c'è anche una questione etica da considerare: se anche le auto a guida autonoma riuscissero a diventare più sicure di quelle a guida umana, e quindi se statisticamente potessero causare meno incidenti, siamo davvero pronti, come società, ad accettare l'eventualità che anche una sola persona muoia a causa delle decisioni di un robot anziché per un errore umano? Finché non avremo una risposta chiara a questa domanda, le auto a guida autonoma dovranno aspettare, o saranno rifiutate da una parte della società. Esistono cinque livelli di guida parzialmente autonoma.

Il livello 1 è chiamato guida assistita: il guidatore è una persona, la quale viene assistita dal sistema di cruise control che mantiene la velocità, da sistemi di frenata di emergenza o anche dal riconoscimento della segnaletica.
Il livello 2 è la guida semi-autonoma: oltre ai sistemi di supporto previsti nel livello 1, si aggiunge il parcheggio autonomo, l'assistenza nello sterzo e nella frenata, ma la responsabilità rimane sempre del guidatore umano.
Il livello 3 è detto di guida altamente automatizzata: qui il conducente umano può occuparsi di altro mentre l'auto è in grado di effettuare automaticamente le manovre di sorpasso e sterzata, di accelerare e di frenare. Però il sistema riconosce i propri limiti e, se necessario, il conducente deve essere in grado di riprendere il controllo della vettura.
Il livello 4 detto di guida completamente automatizzata, include i taxi a guida autonoma, che non hanno un conducente a bordo, e le auto autonome in aree delimitate e ben testate.
Le auto di livello 5, a cosiddetta guida autonoma, sono in grado di muoversi autonomamente in tutte le situazioni sulla strada, tramite sistemi di IA integrati in ogni auto e anche connessioni wi-fi tra le auto. Dalla partenza fino all'arrivo a destinazione, con auto di questo livello non è mai necessario l'intervento di un conducente.

L'impatto dell'IA sulla sanità

Come ha scritto un gruppo di ricercatori di Oxford in un articolo pubblicato sul blog del British Medical Journal: "Non abbiamo bisogno che l'IA superi il test di Turing perché sia utile nella sanità. L'equivalente del test di Turing nell'ingegneria aerospaziale sarebbe stata la creazione di un aeroplano che vola come un uccello, e chiaramente non è stato il caso. Dovremmo trattare la sanità come l'ingegneria e concentrarci sulla costruzione di sistemi di IA che aiutino gli stakeholder e i professionisti della sanità ad affrontare la complessità, per raggiungere l'obiettivo di diagnosi più accurate ed efficaci per i pazienti".

In radiologia, gli algoritmi di apprendimento profondo (deep learning) hanno già dimostrato di poter offrire prestazioni migliori di quelle dei medici in compiti come la classificazione e la rilevazione di noduli polmonari maligni nelle radiografie del torace. Inoltre, le applicazioni del deep learning nella valutazione della densità mammografica mostrano una precisione paragonabile a quella di mammografi esperti. È evidente che, nell'era della quarta rivoluzione industriale, le macchine eccellono in capacità di calcolo e risoluzione di problemi specifici, mentre gli esseri umani per plasmare il futuro possono far conto sull'intuito e sull'intelligenza emotiva.

L'IA ha inoltre il potenziale per migliorare significativamente il triage, ossia la valutazione del grado di priorità dei pazienti, affinché solo i casi gravi siano sottoposti ai medici. Attraverso algoritmi avanzati e modelli di

apprendimento automatico, i sistemi di IA possono analizzare rapidamente e accuratamente i dati dei pazienti, come sintomi, anamnesi e segni vitali.

Immaginiamo invece uno scenario migliore, in cui il settore pubblico adotti rapidamente l'IA all'interno dell'attuale quadro sanitario. In un futuro simile, contribuire con i propri dati sanitari a un sistema basato sulla condivisione dei dati sarebbe un gesto analogo alla donazione di organi. Scegliere di non farlo significherebbe, in altre parole, rinunciare all'opportunità di salvare altri pazienti e noi stessi.

CAPITOLO 2: Che cos'è l'intelligenza

L'avventura dell'intelligenza artificiale, in quanto campo scientifico e tecnologico, è iniziata negli anni Cinquanta con uno scopo ben preciso: quello di creare macchine intelligenti come le persone. Per capire se e quanto siamo effettivamente riusciti ad avvicinarci a questo obiettivo nel corso degli ultimi decenni, dobbiamo quindi iniziare chiedendoci che cosa si intende per intelligenza umana e quali sono gli aspetti che la definiscono.

Quando parliamo di intelligenza, in genere pensiamo alla capacità delle persone di acquisire, apprendere o applicare le proprie conoscenze. L'intelligenza umana ha, quindi, a che fare con le capacità intellettuali e cognitive, e richiede motivazione, creatività e coscienza di sé. Se è vero che le capacità razionali sono un elemento centrale, l'intelligenza umana è sempre supportata dalle nostre emozioni, dalle relazioni che intessiamo con gli altri e dalla nostra esperienza diretta del mondo.

Teoria del pensiero lento e veloce di Daniel Kahneman.
La teoria individua due modalità di attività cognitiva in una persona: il pensiero veloce, che attiviamo inconsciamente in molte delle azioni quotidiane, e il pensiero lento, che invece mettiamo in moto quando dobbiamo concentrarci per risolvere un problema difficile. Ad esempio, quando tra la folla riconosciamo il volto di una persona che conosciamo stiamo utilizzando il pensiero veloce, mentre quando dobbiamo risolvere un'operazione matematica difficile attiviamo il pensiero lento.
Il pensiero veloce è anche usato per percepire l'ambiente intorno a noi, per vedere, per sentire e in generale per acquisire informazioni tramite i nostri sensi e farci un'idea del contesto in cui ci troviamo. Coinvolge la sfera emotiva e percettiva, e ne abbiamo esperienza in tutte quelle attività che facciamo senza esserne consapevoli o che non sappiamo controllare.

Questa teoria offre anche una spiegazione del modo in cui le due dimensioni cognitive interagiscono tra loro e di come si passa dall'una all'altra. I due tipi di pensiero spesso si aiutano a vicenda. Il pensiero lento, che rappresenta la razionalità, a volte nell'esaminare tutti i dettagli di una questione da risolvere prende delle scorciatoie e si serve del pensiero veloce per individuarle. È ciò che fanno, ad esempio, i giocatori di scacchi: valutano attentamente la configurazione dei pezzi sulla scacchiera per scegliere la mossa successiva, ma si avvalgono anche di mosse predefinite, quelle che hanno studiato e provato così tante volte che le usano senza neanche rifletterci.

Le teorie cognitive, come quella del pensiero lento e veloce, cercano di spiegare come funziona la nostra mente, ma chiaramente l'intelligenza dipende anche dal nostro cervello, cioè da quell'intricata rete di neuroni che supporta la memorizzazione e l'evoluzione delle nostre conoscenze e azioni. Usando una terminologia da computer, il cervello è il nostro hardware, mentre la mente è il nostro software cognitivo.

Cosa vuol dire che una macchina è intelligente? All'inizio dell'avventura scientifica dell'IA, i ricercatori hanno avuto il problema di definire cos'è l'intelligenza di una macchina. Così, la prima domanda che si sono posti è se l'IA doveva possedere tutti gli aspetti dell'intelligenza umana. Ma riprodurre questa complessità sembrava porre difficoltà insuperabili per un computer. Perciò, soprattutto agli inizi degli studi sull'IA, si è scelto di puntare sull'intelligenza razionale, cioè su un comportamento analogo al nostro pensiero lento. L'importante era che, dato un obiettivo, un computer sapesse raggiungerlo nel modo migliore, secondo criteri di ottimalità prestabiliti, con gli strumenti a disposizione. Di fatto, per molti anni questo è stato il principio seguito per creare sistemi di IA. Nella pratica, questo voleva dire che i ricercatori creavano, con la loro intelligenza, dei metodi per risolvere problemi specifici, per poi codificarli in un linguaggio di programmazione in modo che un

computer potesse usarli. A far funzionare l'intelligenza artificiale era insomma l'intelligenza umana, quella usata nel lavoro di ricercatori e programmatori.

Riprodurre l'apparenza dell'intelligenza. Spesso questo concetto viene spiegato così: così come gli aerei possono volare, anche se non lo fanno con gli stessi meccanismi usati dagli uccelli (per esempio, non battono le ali), allo stesso modo l'importante è che un programma si comporti in modo intelligente, anche se non usa gli stessi meccanismi dell'intelligenza umana.

Questo approccio era alla base del ragionamento del matematico inglese Alan Turing, che già nel 1950 si chiedeva se le macchine potessero pensare e proponeva (indirettamente) un test per capire se una macchina fosse intelligente. Quello che poi è diventato noto come il test di Turing prevede che una persona, il valutatore, comunichi per mezzo di messaggi di testo con due entità: un computer e un'altra persona. Queste due entità sono in stanze separate e non visibili al valutatore. Se il valutatore, dopo aver comunicato con entrambi, non riesce a distinguere qual è l'umano e qual è il computer, vuol dire che il computer è (o almeno sembra) intelligente quanto una persona.

Se non possiamo dire in assoluto se un programma è intelligente, possiamo però misurarne l'efficienza nello svolgere un determinato compito. A questo servono i benchmark (letteralmente "banco di prova") creati per migliorare le tecniche di IA, per confrontarle tra di loro a partire da un termine di paragone comune e per capire quale tecnica è migliore. Un esempio è ImageNet, un database di circa 14 milioni di immagini, reso disponibile già dal 2009 e usato dai ricercatori di tutto il mondo per valutare le capacità dell'IA di riconoscere cosa è rappresentato in una immagine. Ciascuna immagine inclusa nel database è infatti stata manualmente annotata e catalogata sulla base di ciò che vi è rappresentato. Il database è liberamente accessibile, così che i ricercatori che lavorano allo sviluppo di nuovi software possano metterne alla prova le capacità di interpretare correttamente i contenuti di un'immagine.

Intelligenza umana e artificiale

L'era dell'intelligenza artificiale generale (AGI) avrà inizio quando il software raggiungerà un insieme così vasto di capacità. I progressi verso questo obiettivo stanno accelerando, ma c'è ancora molta strada da percorrere. Negli esseri umani, alcune operazioni mentali sono volontarie, consapevoli e richiedono uno sforzo (spesso vengono chiamate "sistema 2"), mentre altre sono automatiche ("sistema 1"). La nostra capacità di sforzo mentale è però limitata, e così le operazioni controllate vengono eseguite principalmente una alla volta, anziché simultaneamente. Le operazioni automatiche, cioè quelle che non richiedono la nostra attenzione, includono abilità percettive e motorie che condividiamo con altri animali, ma anche abilità apprese. Ad esempio, guidare un'auto è un'operazione che richiede sforzo per un principiante, ma in seguito diventa automatica. Allo stesso modo, la lunga esperienza accumulata consente ai campioni di scacchi o ai medici esperti di comprendere situazioni complesse con un solo sguardo.

I pensieri del sistema 2 sono prodotti da atti, spesso consapevoli, di osservazione intenzionale e ragionamento. Al contrario, la percezione del mondo esterno e molte delle nostre sensazioni e intuizioni sembrano cose che ci accadono, senza alcuna intenzione da parte nostra e senza la consapevolezza di come siano giunte alla nostra mente. La prima fase nello sviluppo dell'intelligenza artificiale si è concentrata sul ragionamento simbolico esplicito. Ma dai tempi del trionfo del machine learning nel 2012, l'intelligenza artificiale produce soluzioni attraverso le operazioni per nulla trasparenti di grandi reti neurali, che condividono molte caratteristiche con le intuizioni automatiche del sistema 1. Tuttavia, l'integrazione di intuizione e ragionamento nell'intelligenza artificiale non è stata ancora raggiunta durante questo primo decennio dell'era del machine learning.

CAPITOLO 3: Come siamo arrivati all'IA di oggi

Di solito si fa iniziare la storia dell'IA come disciplina scientifica nel 1956, quando un gruppo di matematici americani, tra cui John McCarty, un pioniere dell'IA e al tempo giovane professore di matematica, si riunisce per una estate presso il Dartmouth College, nel New Hampshire, e definisce un progetto di ricerca usando per la prima volta il termine "intelligenza artificiale". Il progetto era molto ambizioso e prevedeva la costruzione

di macchine intelligenti nell'arco di pochi anni, cosa che chiaramente non è avvenuta. Ma così è iniziata l'avventura dell'IA.

Lo scopo in origine era quello di arrivare a una intelligenza artificiale generale (Artificial General Intelligence, AGI), cioè in grado di risolvere problemi in tutti i campi, come sanno fare le persone, ma presto i ricercatori si resero conto che dovevano mirare a un obiettivo più semplice e alla portata, cioè alla creazione di una intelligenza artificiale limitata (Narrow), ossia focalizzata sulla soluzione di problemi specifici. Lo scopo non era più quello di creare una macchina che sapesse fare tutto, ma quello di realizzare tanti programmi diversi, ognuno pensato per risolvere un tipo di problema specifico.

Il metodo, chiamato algoritmo, era spesso basato su regole del tipo "se <condizione>, allora <azione>". Quando la condizione si verifica, l'azione corrispondente viene eseguita. Il funzionamento di un algoritmo a regole è in realtà più semplice di quanto possa sembrare ed è qualcosa con cui tutti siamo familiari, se nella nostra vita abbiamo mai aperto un libro di cucina. Quando cuciniamo un dolce, infatti, stiamo eseguendo un algoritmo, cioè una sequenza di passi descritti da regole.

Negli algoritmi, sia le condizioni sia le azioni possono essere associate a una probabilità, espressa con un numero tra 0 e 1 che indica una soglia di attivazione che permette una maggiore flessibilità nel descrivere una situazione reale. Un ruolo centrale nello sviluppo dell'IA lo hanno svolto i giochi, spesso usati dai ricercatori per concentrare l'attenzione e gli sforzi su un problema semplice e ben definito, e spingere le abilità dell'IA il più avanti possibile giocando contro avversari umani. Un tipico esempio è la sfida a scacchi del programma Deep Blue di IBM, che nel 1997 vinse contro Garri Kasparov, all'epoca il miglior giocatore di scacchi al mondo. Deep Blue usava un sistema IA a regole, ma la carta vincente fu l'utilizzo di un computer velocissimo (per quell'epoca) che permetteva di analizzare molte mosse e contromosse dell'avversario prima di decidere la prossima mossa.

Il fatto che un programma potesse vincere contro il campione mondiale di scacchi suscitò al tempo molto scalpore, perché si riteneva, e si ritiene ancora oggi, che per giocare bene a scacchi una persona debba essere molto intelligente. La sorpresa della sconfitta di Kasparov sembrava mettere in dubbio che l'intelligenza umana fosse superiore a quella di un computer. In realtà, quello che Deep Blue ci ha insegnato è che, per giocare bene a scacchi, serve molta memoria e velocità, più che intelligenza. Dato che il cervello e la mente umana hanno memoria e velocità limitate, noi umani dobbiamo ricorrere ad altri mezzi per poter eccellere in questo gioco.

Una rete neurale consiste di tante unità di computazione molto semplici (i neuroni) collegate tra loro in modo che l'output di un neurone diventi l'input di altri neuroni. Ogni connessione parte da un neurone e arriva a un altro neurone, ed è associata a un numero (chiamato "peso") che determina l'informazione che viene data in input al neurone di arrivo. La rete neurale riceve in input un problema e deve produrre in output la soluzione del problema. I pesi sulle connessioni interne alla rete determinano l'output che la rete produce. Ad esempio, potremmo voler creare una rete neurale che interpreta correttamente i comandi vocali. In questo caso l'input sarebbe un comando vocale (come "Chiama Francesca") e l'output sarebbe la corretta azione corrispondente al comando (come il far partire una chiamata al contatto giusto tra quelli in rubrica, Francesca appunto).

Nella fase di training, alla rete vengono forniti numerosi esempi con le soluzioni associate. Se, come nell'esempio, l'output prodotto dalla rete è sbagliato (l'immagine di un cavallo viene identificata come scimmia) si attiva automaticamente l'algoritmo di backpropagation che modifica i "pesi" delle connessioni tra i neuroni per diminuire la probabilità di errore in futuro.

Caratteristica fondamentale delle reti neurali è che possono imparare da sole come risolvere un problema, senza richiedere che questo venga indicato tramite un algoritmo. Per questo l'approccio basato su reti neurali si chiama machine learning, cioè apprendimento automatico. Questa forma di apprendimento avviene tramite l'uso di tanti esempi formati da due componenti: una istanza del problema da risolvere e la soluzione corretta del problema.
Se questo output non è corretto, cioè non è uguale alla seconda componente dell'esempio (la soluzione giusta), la rete usa un algoritmo per aggiustare i pesi delle sue connessioni in modo da minimizzare l'errore

nei tentativi successivi. E così via per tutti gli esempi disponibili. Dopo la fase di apprendimento, detto apprendimento supervisionato per la presenza di soluzioni corrette generate da umani, la rete viene testata su altri esempi non inclusi nell'elenco iniziale, per controllare che la probabilità di errore sia piccola. A questo punto la rete, con i suoi pesi opportunamente modificati dalla fase di apprendimento, è pronta per essere usata.

Se, nella fase di test, la rete dimostra di avere un margine di errore sufficientemente piccolo, possiamo iniziare a usarla, in uno scenario di uso simile a quello descritto dagli esempi di training. Altrimenti dobbiamo continuare con il training, finché l'errore non sarà abbastanza piccolo.

Il machine learning include molte tecniche diverse, e la sua evoluzione ha portato a una variante molto utile, chiamata deep learning. Il machine learning tradizionale richiede un intervento umano significativo in fase di programmazione, per individuare le caratteristiche del problema da risolvere. Se si vuole creare una rete neurale per classificare immagini di cani e gatti, cioè per capire se in una immagine c'è un cane o un gatto, dobbiamo configurare manualmente la rete neurale per riconoscere caratteristiche come la forma degli occhi, della coda, delle orecchie e così via. Con il deep learning, invece, che usa reti neurali a vari livelli (da qui deep, cioè "profondo"), questo non è più necessario. Mentre il machine learning tradizionale funziona bene per analizzare dati strutturati, cioè con un formato standardizzato (ad esempio una tabella di dati sui clienti contenente colonne come nome, indirizzo e numero di telefono, il numero totale di clienti e la località con il numero massimo di clienti), il deep learning è una forma specifica di machine learning che riesce a gestire dati non strutturati, come le immagini o il linguaggio naturale. È con il deep learning che i sistemi come Alexa riescono a capire i nostri comandi vocali.

È importante notare la fondamentale differenza tra il machine learning e il metodo a regole. Con le reti neurali, l'IA inizialmente non sa come risolvere il problema di riconoscere la cifra nell'immagine, ma lo apprende tramite gli esempi che le forniamo. Ci sarà però sempre una piccola probabilità di errore. Se l'IA segue quelle istruzioni, avrà risolto il problema perfettamente e senza errori. Ma dobbiamo sempre darle lo stesso input (gli ingredienti) e sempre nelle quantità che abbiamo previsto, altrimenti semplicemente non saprà cosa fare. Invece, qualunque immagine diamo in pasto alla rete neurale, lei saprà sempre generare una risposta, anche se a volte non sarà corretta.

Un'altra differenza importante è la centralità dei dati nel machine learning. Senza dati, che fungono da esempi, le reti neurali non sanno come imparare a risolvere il problema in questione. Questi metodi sono stati quindi facilitati da Internet, che ha permesso la condivisione di grandissime quantità di dati (immagini, testi e altro), pubblicamente disponibili, che le aziende hanno potuto usare per mettere a punto le loro reti neurali. Per usare tutti questi dati, l'IA aveva anche bisogno di computer molto veloci, altrimenti ci avrebbe messo troppo tempo a imparare dagli esempi. A questo scopo, sono venute in aiuto piattaforme hardware sufficientemente veloci come le GPU (Graphics Processing Units), pensate originariamente per applicazioni grafiche come i giochi online ma poi usate con successo anche per l'IA in ragione della loro grande velocità.

L'apprendimento automatico permette alle macchine di imparare da sole a risolvere un problema (una volta che le abbiamo fornito gli esempi) e ciò rappresenta una tappa importante verso una qualche forma di effettiva intelligenza artificiale, perché la capacità di imparare dall'esperienza è un aspetto importante dell'intelligenza umana. Con le reti neurali descritte in precedenza, l'IA sa imparare dai dati come risolvere un problema e come interpretare correttamente dati di vario tipo, quali immagini, parlato e testi. Ma l'output è sempre pre-definito dalle persone. Ad esempio, se una rete neurale deve capire se un prodotto ha un difetto, essa prende in input un'immagine del prodotto e in output indica un semplice "sì" o "no".

L'evoluzione delle tecniche di apprendimento automatico ha recentemente permesso di costruire reti neurali che riescono anche a generare contenuti originali, cioè non pre-definiti da umani. Ad esempio, ChatGPT, il primo sistema usato su larga scala di IA di nuova generazione, prende in input il testo scritto da una persona e risponde con un altro testo generato dal sistema. Lo stesso avviene per sistemi di testo-immagini o testo-video, ad esempio DALL-E e Sora (dell'azienda OpenAI) o Midjourney (dell'azienda omonima), che prendono in input un testo che descrive una situazione e produce in output una immagine o un video (nel caso di Sora)

che rappresenta quella situazione. Questa nuova abilità accresce ed estende le applicazioni e l'utilità dell'IA praticamente in tutti i campi e settori produttivi.

L'apprendimento automatico supervisionato ha bisogno di tanti esempi in cui una persona indica la risposta giusta. Qui invece si tratta di apprendimento non-supervisionato. Questo vuol dire che al sistema vengono mostrati dei testi, ma non c'è un umano che fornisca altre informazioni sulla soluzione giusta o sbagliata di un problema.

ChatGPT è un sistema di interazione con un cosiddetto large language model, cioè un modello del linguaggio umano. Nel caso specifico, con ChatGPT dialoghiamo con GPT-4 o un'altra sua versione (prima della 4 veniva usata la 3.5). GPT significa Generative Pre-trained Transformer: si tratta di IA generativa, cioè che sa generare contenuti (in questo caso testo), viene addestrata (trained) prima di essere usata, e usa una tecnica di apprendimento automatico basata sul concetto di transformer, ovvero una innovativa architettura basata sulle reti neurali. Questi sistemi vengono chiamati large, cioè grandi, perché vengono addestrati su enormi quantità di dati di testo. Ad esempio, GPT-4 è stato addestrato su testi, presi dal web, che sommati contengono circa 300 miliardi di parole.

Contrariamente a quanto si pensi, l'output del large language model è una parola, non una frase o un testo lungo. Quindi, il modello può essere usato per completare una frase dove manca una parola: l'input sarà la frase senza l'ultima parola, e l'IA genererà la parola mancante. Ad esempio, se scriviamo "Tanto va la gatta al lardo, che ci lascia lo", il programma risponderà con "zampino". Ma come si fa a guidare la generazione della parola in modo tale che, aggiunta alla frase usata come input, formi una frase sensata? Tramite i testi forniti durante la fase di addestramento. Le frasi contenute in questi testi, infatti, vengono usate dall'IA per associare una probabilità alla presenza di una parola dopo una certa sequenza di altre parole.

I large language model fanno esattamente questo: generano in output la parola che, secondo i calcoli statistici da loro compiuti nella fase di addestramento, è la più probabile continuazione del testo fornito in input. Più le probabilità sono calcolate bene e più la parola che viene generata sarà un completamento sensato della frase in input. Il passaggio da parola a testo è semplice, basta infatti usare il sistema ripetutamente. Se si vuole che l'IA risponda a una domanda, la prima volta verrà dato in input all'IA il testo della domanda, e l'IA risponderà con la prima parola della risposta. Poi verrà data in input la domanda più la prima parola generata, e l'IA risponderà con la seconda parola, e così via. La sequenza di parole generate formerà la risposta completa alla domanda iniziale.

Un meccanismo analogo viene usato dai sistemi text-to-image come DALL-E e Midjourney, che generano immagini digitali un pixel alla volta. Così come un testo è scomponibile in una sequenza di parole, una immagine digitale è invece una sequenza di pixel, cioè piccoli quadratini di immagine, ognuno con un suo colore.

Con i sistemi a regole l'IA risolve un problema definito da un umano, da cui riceve anche l'algoritmo per risolverlo; con l'apprendimento supervisionato, invece, l'IA riceve dall'umano il problema da risolvere e alcuni esempi di soluzioni di istanze del problema, e deve apprendere da questi esempi come risolvere future istanze dello stesso problema; con le tecniche generative, invece, l'IA riceve da un umano esempi di frasi (o di immagini) e apprende quali testi (o immagini) generare, una parola (o pixel) alla volta.

Mentre prima l'IA veniva addestrata o programmata per risolvere un problema specifico, e poi poteva essere usata solo per risolvere quel problema, l'IA generativa viene addestrata a risolvere un problema molto generale (quello di generare testi o immagini sensate) e poi può essere usata per molti problemi specifici relativi al linguaggio o alle immagini. Negli anni, l'intervento delle persone nella progettazione di questi sistemi di IA si è sempre più rivolto alle strutture di reti neurali, piuttosto che agli algoritmi che il sistema deve usare per risolvere i problemi, o agli esempi di soluzioni.

È però importante notare che i testi forniti ai sistemi di IA generativa per il suo addestramento sono comunque generati da umani, perché sono presi dal web e quindi, solitamente, sono stati scritti da persone. La quantità

di articoli scientifici sull'IA pubblicati ogni anno è costantemente in crescita e nel 2021 è arrivata a circa mezzo milione, una quantità più che doppia rispetto al 2010.

Per quanto riguarda i paesi asiatici: secondo l'AI Index 2023 Report oggi essi arrivano a contribuire per quasi il 40% delle pubblicazioni scientifiche nei convegni sull'IA, con 9 delle 10 istituzioni con più alto numero di pubblicazioni sul tema situate in Cina (e la decima negli Stati Uniti). Vi è stata una notevole evoluzione negli argomenti della ricerca in IA negli ultimi anni, con sempre più lavori su tecniche di IA per l'analisi e la generazione di testi con tecniche di machine learning, rispetto ad altre tecniche o altri domini applicativi.

L'evoluzione dell'IA è stata supportata da altre innovazioni tecnologiche, come Internet e le GPU. Tra le nuove frontiere di collaborazione tra l'IA e altre scienze e tecnologie, c'è anche la combinazione con le neurotecnologie, sempre più usate in campo medico e per il benessere fisico e psichico. Con le neurotecnologie è possibile leggere l'attività neurale, ad esempio per controllare dispositivi esterni, come le protesi per persone con arti mancanti, o anche scrivere, cioè modificare, l'attività neurale per riparare una parte del cervello colpito da disturbi neurologici.

La combinazione di IA e neurotecnologie rende i dispositivi più precisi ed efficaci. In Cina nel 2023 alcune scuole hanno iniziato a usare tecniche di IA attraverso fasce che, poste sulla testa degli studenti, acquisiscono informazioni dai loro segnali neurologici e cambiano di colore a seconda del livello di concentrazione stimato. Quando il colore è rosso, significa che lo studente è concentrato, mentre quando è blu significa che è distratto.

L'IA sarà mai in grado di vincere il Nobel?
Nel 1998 è stato creato il primo prototipo di AIBO (che sta per Artificial Intelligence RoBOt), un piccolo robot a forma di cane dal design molto accattivante che sa interagire con le persone e l'ambiente in modo simile a un cane, riconoscendo i suoni e i comandi di un umano, seguendolo e interpretando quello che vede intorno a lui con le sue telecamere. AIBO può anche essere programmato da chi lo acquista, in modo da fornirgli una personalità unica. Sony ha venduto AIBO con grande successo dal 1999 al 2006 e poi lo ha reintrodotto sul mercato nel 2017.

Un esempio eclatante dei passi avanti fatti nella ricerca scientifica grazie all'utilizzo dell'IA è AlphaFold, un sistema creato nel 2018 dall'azienda DeepMind che è in grado di fare predizioni sulla struttura delle proteine, tramite tecniche di deep learning. Le proteine sono fondamentali nella nostra vita, perché sono responsabili di tantissime funzioni dell'organismo, incluse le reazioni metaboliche e la replicazione del DNA. Chimicamente, una proteina è una catena di aminoacidi caratterizzata dal ripiegamento in una struttura tridimensionale che determina la sua funzione biologica. Conoscere il funzionamento delle proteine consentirebbe di aprire una nuova era nella ricerca e applicazione della biologia molecolare e accelererebbe la creazione di farmaci per curare molte malattie.

Il problema è che capire come gli aminoacidi determinino la struttura 3D di una proteina non è semplice, perché una proteina può essere composta da molti aminoacidi, che si possono piegare in tantissimi modi. AlphaFold ha praticamente risolto il protein folding problem, cioè questo problema di generare la struttura 3D di una proteina a partire dalle informazioni sulla sequenza di aminoacidi che la compongono. Testando AlphaFold su proteine di cui si conosce già la struttura 3D, nell'ambito del sistema di test chiamato CASP (Critical Assessment of Structure Prediction), il sistema ha mostrato una capacità predittiva del 90%, superiore a tutti gli altri metodi. Quello che fa AlphaFold veniva fatto anche prima, con tecniche di ricerca biologica più tradizionali, ma richiedeva molto più tempo e costi molto più alti, dovuti all'uso di procedure fisiche e apparecchi molto costosi. Ad inizio 2020, AlphaFold è stato anche in grado di prevedere la struttura di diverse proteine del SARS-CoV-2 ben prima che queste fossero definite con i metodi tradizionali, i quali hanno sostanzialmente confermato la qualità dei risultati ottenuti con l'IA.

I risultati di AlphaFold furono commentati dicendo che "in mezz'ora di calcolo ha risolto quello che non ero riuscito a fare in dieci anni di continua sperimentazione". Sempre secondo Lupas, "questo cambierà sicuramente la medicina, la ricerca, la bioingegneria, cambierà davvero tutto". Uno studio della rivista Nature del settembre 2023 ha chiesto a 1.600 ricercatori (tra chi aveva pubblicato articoli scientifici nel 2022 o chi legge questa rivista, soprattutto in Asia, Europa e Nord America) quanto pensano che l'IA diventerà utile per i

loro campi di ricerca nei prossimi 10 anni: più della metà ritiene che sarà molto importante o addirittura essenziale. Secondo lo studio, 2 ricercatori su 3 hanno detto che l'IA permette loro di analizzare dati più velocemente e il 58% ha detto che velocizza anche calcoli che prima non erano possibili perché avrebbero impiegato troppo tempo. Lo studio di Nature mostra anche che l'IA permette di risparmiare tempo e denaro, automatizzare l'acquisizione di dati, scrivere programmi più velocemente, rispondere a domande che prima erano difficili da risolvere, fare nuove scoperte e generare nuove ipotesi di ricerca.

I 1.600 ricercatori intervistati hanno però indicato anche alcune preoccupazioni: il 69% ha detto che l'uso di deep learning può portare a usare tecniche di cui non si comprende il funzionamento; il 58% pensa che possa ereditare bias (cioè pregiudizi) dai dati di training; il 55% ritiene che faciliti casi di frode nella ricerca; e il 53% ha notato che può portare a processi di ricerca non riproducibili. I ricercatori hanno inoltre sottolineato che le più moderne tecniche di IA sono molto costose e quindi accessibili solo a pochi, il che incrementa la distanza tra le università o i centri di ricerca più finanziati e gli altri.

Dove sta andando l'intelligenza artificiale?
L'IA ha una lunga storia e si basa su teorie matematiche consolidate in diverse discipline, tra cui logica, teoria della probabilità, teoria del controllo, teoria dei giochi e teoria dell'utilità. Per la maggior parte della sua storia, i ricercatori in IA hanno sfruttato queste teorie per cercare di arrivare al loro obiettivo di sempre: creare sistemi dotati di una intelligenza generale. Tali sistemi trasformerebbero completamente la nostra civiltà.

Nell'ultimo decennio, però, la maggior parte di queste idee è stata abbandonata a favore principalmente di un approccio: il deep learning, ovvero l'apprendimento profondo, in cui circuiti con miliardi o trilioni di parametri vengono ottimizzati per imparare da enormi insiemi di dati. Grosso modo ora il paradigma dominante nello sviluppo dell'IA è questo: "se al primo tentativo non funziona, aumenta il numero di parametri e raccogli più dati".

Alcuni ricercatori ritengono che procedere seguendo questo approccio porterà inevitabilmente a sistemi di intelligenza artificiale generale. Altri sostengono il contrario perché ne sottolineano i ripetuti insuccessi, per esempio l'incapacità di imparare l'aritmetica di base da parte dei large language model (modelli linguistici grandi) nonostante l'input di milioni di esempi e di migliaia di spiegazioni e algoritmi. Questo genere di fallimenti suggerisce una debolezza intrinseca nella capacità di questo tipo di tecniche di IA di generalizzare accuratamente a partire da un numero ragionevole di esempi.

CAPITOLO 4: L'etica dell'intelligenza artificiale

L'equità è solo uno dei valori etici che ci aspettiamo vengano rispettati dalle persone quando prendono decisioni che possono avere un impatto sulla vita degli altri. L'etica è una branca della filosofia che si occupa di definire cosa è giusto o sbagliato fare in relazione a certi valori condivisi. Quindi l'etica, in particolare quella normativa, descrive come dovremmo comportarci per adeguarci a quei valori. Quando invece parliamo di etica in relazione all'IA, intendiamo lo studio e la definizione del modo in cui i sistemi basati su IA debbano comportarsi ed essere usati per essere compatibili con determinati valori sociali. In altre parole, le decisioni prese da una macchina devono essere tali che, se fossero state prese da una persona, il comportamento di quella persona verrebbe giudicato etico.

L'uso dell'IA: le decisioni prese da un algoritmo possono essere "perfette" da un punto di vista etico, ma l'IA può essere adoperata dalle persone in modo immorale. L'immissione di una tecnologia pervasiva come l'IA può avere conseguenze molto significative e socialmente inaccettabili, anche quando questa si comporta secondo norme etiche e viene adoperata eticamente. Non è l'IA a dover essere etica, anche perché non si può imputare a una tecnologia la responsabilità morale delle proprie azioni, piuttosto è l'intero ecosistema intorno all'IA che deve comportarsi in modo etico.

L'IA ha spesso bisogno di dati personali per poter fornire servizi personalizzati. Questi dati personali possono essere forniti esplicitamente da noi (come quando inseriamo i dati anagrafici in una richiesta di mutuo a una banca), oppure raccolti dall'IA stessa, che osserva le nostre azioni online (come i nostri like o follow, o i nostri post con testo e immagini, quando utilizziamo una piattaforma social). In questo secondo caso, spesso non ci

rendiamo neanche conto che stiamo dando informazioni che saranno usate dall'IA – ed è per questo che a volte ci stupiamo di trovare annunci pubblicitari molto allineati alle nostre preferenze. Questo è possibile proprio perché l'IA usata nelle piattaforme social riesce a ricavare informazioni utili dalle nostre azioni per identificare le nostre preferenze, spesso in maniera più efficace di quanto sapremmo fare noi.

Non c'è niente di illegale, naturalmente, perché l'accordo contrattuale con le app o le piattaforme social indica cosa viene fatto con i nostri dati. Ma chi si legge davvero tutte quelle paginate di testo che descrivono il trattamento dei dati, prima di cliccare sul pulsante in fondo e dare il proprio consenso? Con l'IA generativa, la privacy può essere violata anche in un altro modo. Oltre ad accumulare dati, l'IA generativa li può anche riprodurre, e questo fa sì che un sistema come ChatGPT possa generare un testo o porzioni di testo che dovrebbero invece rimanere riservati. Il sistema di IA può ad esempio ricreare un testo che era stato scritto da una persona in un precedente colloquio con lo stesso sistema, e può farlo mentre comunica con un'altra persona.

A maggio 2023 alcuni dipendenti della Samsung hanno usato ChatGPT per farsi aiutare a scrivere un programma informatico, inserendo nella casella di dialogo un pezzo di codice di un programma di proprietà dell'azienda e confidenziale. Questo ha creato una situazione di vulnerabilità, perché quel codice privato è stato condiviso con un'altra azienda (OpenAI, l'azienda che produce ChatGPT e che memorizza tutti i dialoghi che gli utenti hanno con il sistema). Se quel programma dovesse essere utilizzato per addestrare nuove versioni di ChatGPT, alcune sue parti potrebbero di nuovo essere generate in dialoghi futuri. Per questo, la Samsung ha vietato ai propri dipendenti di usare ChatGPT per scopi aziendali.

L'algoritmo COMPAS è stato uno dei primi sistemi di IA a mostrare un comportamento discriminatorio. COMPAS viene usato nel sistema della giustizia criminale negli Stati Uniti per fornire un supporto ai giudici quando devono decidere se tenere in prigione un criminale, decisione che va presa sulla base di una valutazione della sua possibilità di commettere nuovamente un crimine una volta rilasciato. Questo sistema prende come input le informazioni sulla persona che ha commesso un reato e genera un valore tra 0 e 1 che indica la probabilità prevista che questa persona possa ricommetterlo. Più è alto il valore, più vuol dire che il sistema crede che ci sia una alta probabilità di nuovi reati da parte di questa persona. COMPAS usa tecniche di machine learning ed è allenato su dati storici di decisioni prese dai giudici nel corso degli anni precedenti. Quando l'organizzazione di giornalismo investigativo ProPublica, nel 2016, ha analizzato questo algoritmo per vedere se discriminava tra persone bianche e di colore, per prima cosa ha calcolato la percentuale di errori compiuti dal sistema per le due categorie di persone, verificando che era praticamente la stessa (circa il 10%).

Le tecniche di machine learning si basano, infatti, sul calcolo probabilistico e, quindi, implicano sempre una certa percentuale, anche se molto piccola, di errore. A una prima analisi sembrava quindi che COMPAS funzionasse bene dal punto di vista dell'equità. Poi però sono stai analizzati nel dettaglio gli errori, discriminando tra falsi positivi (cioè errori in cui veniva generata una probabilità alta di recidiva, mentre in realtà doveva essere bassa) e falsi negativi (cioè errori in cui veniva generata una probabilità bassa di recidiva, mentre in realtà doveva essere alta). A questo punto si è visto che l'algoritmo generava soprattutto falsi positivi per persone di colore (cioè tendeva a giudicarle peggio di quanto meritassero), mentre forniva soprattutto falsi negativi per persone di pelle bianca (giudicandole meglio di quanto meritassero). Risulta quindi fondamentale che i dati di training con cui la macchina viene allenata siano analizzati attentamente, per individuare i bias e mitigarli.

Sicurezza, sorveglianza e uso militare

Esistono però applicazioni molto più problematiche. Per esempio, quelle legate alle armi autonome. In uno scenario di guerra, un drone armato può essere equipaggiato con una videocamera e un sistema di IA che sa riconoscere una persona specifica in una folla. Questo vuol dire che potrebbe essere usato per individuare un obiettivo umano e per azionare le sue armi contro quell'obiettivo in modo autonomo. Alla sorveglianza di massa e senza consenso, qui si aggiunge la questione ancora più seria di una entità artificiale che è messa nella condizione di decidere se uccidere o meno una persona.

L'esercito israeliano ha utilizzato sistemi di IA – noti con i nomi inglesi "The Gospel" e "Lavender" – per l'identificazione automatica dei target da bombardare. Ma non è stato il primo caso. Secondo un'analisi delle

Nazioni Unite, le armi autonome sono state usate in contesti di guerra per la prima volta nel 2020 in Libia quando, nell'ambito dell'operazione Peace Storm dispiegata nel corso della guerra civile, uno sciame di droni turchi ha identificato e attaccato forze legate al generale Haftar sotto il comando dell'allora primo ministro libico al-Sarraj.

CAPITOLO 5: Allucinazioni e deepfake

Le tecniche più avanzate di IA in circolazione permettono di generare contenuti come testi e immagini, riuscendo spesso a sostenere dialoghi e a rispondere a tono alle domande di una persona. A volte, però, generano delle cosiddette allucinazioni, cioè testi che contengono informazioni palesemente false, o immagini in apparenza realistiche ma che, ad esempio, contravvengono alle leggi fisiche, anche se nei dati di training ci sono solo informazioni corrette e immagini vere.

Possiamo per questo dire che l'intelligenza artificiale ci sta mentendo? In realtà, non è corretto parlare di bugie. Farlo implicherebbe che l'IA sappia quali sono le informazioni vere e che possa scegliere consapevolmente di darci una versione falsa. Le cose non stanno così: anche quando genera informazioni false, l'IA non sa cosa sta facendo, né può essere mossa da intenzioni malevole.

ChatGPT. Non ha una esperienza del mondo, ma solo informazioni che vengono dai testi di esempio usati nella fase di training. A partire da questi, assegna quindi una probabilità a ogni parola e decide quale parola scrivere dopo il testo già scritto dall'utente, selezionando quella con la probabilità più alta.

Il termine deepfake mette insieme deep, dalla tecnica di IA del deep learning, e fake, cioè "falso", per indicare la generazione tramite IA di immagini, video o audio intenzionalmente falsi, che però possono risultare estremamente convincenti e realistici. Tra alcuni degli esempi famosi che potete trovare in rete ci sono immagini di papa Francesco che indossa un piumino bianco di una nota casa di moda, o di Donald Trump catturato dalle forze dell'ordine.

La produzione e distribuzione di materiale falsificato non è un fenomeno nuovo. Ciò che c'è di nuovo è il fatto che ora quasi tutti possono facilmente generare contenuti falsi e disseminarli su scala globale in maniera estremamente rapida tramite le piattaforme di social media. Inoltre, le più recenti tecniche di IA permettono di farlo con livelli di realismo sorprendenti: i falsi sono così indistinguibili dalla realtà al punto di rendere molto difficile riconoscere se un certo materiale è autentico oppure no. Se con le allucinazioni dell'IA le informazioni false vengono generate in modo involontario, cioè senza che ci sia qualcuno che voglia intenzionalmente divulgare il falso, le deepfake sono invece sempre intenzionali: c'è un utente umano che usa l'IA proprio allo scopo di creare materiale indistinguibile dal vero ma falso, a danno di altri. Se a questo si combina la possibilità di usare l'IA per profilare le persone, cioè per raccogliere informazioni sulle loro preferenze e abitudini, c'è allora di che preoccuparsi.

CAPITOLO 6: L'IA al lavoro e tra i banchi di scuola

L'IA ci ruberà il lavoro?

Dal punto di vista del lavoro, l'IA è forse la tecnologia più pervasiva che si sia mai vista nella storia dell'umanità, perché ha applicazioni utili in qualunque settore e modifica praticamente ogni mansione. È chiaro quindi che le trasformazioni a cui dà, e darà, luogo sono enormi. A ciò si aggiunge il fatto che tanto l'evoluzione quanto l'adozione dell'IA sono incredibilmente veloci: nella maggior parte dei casi non ha bisogno di artefatti fisici per realizzarsi, come invece avveniva per i grandi macchinari della prima rivoluzione industriale. Ha bisogno solo di programmi software che vengono eseguiti su un computer e che possono essere replicati istantaneamente. È proprio questa rapidità a generare grandi preoccupazioni: il pericolo all'orizzonte è che la società nel suo complesso non abbia il tempo per capire come gestire la trasformazione e sia destinata a subirla. Le tempistiche per adeguare le strutture sociali ai nuovi scenari sono molto più lente delle trasformazioni fattuali. Uno studio di McKinsey stima che entro il 2030 il 21,5% delle ore lavorative negli Stati Uniti potrebbe essere automatizzata, senza l'IA generativa. Con l'IA generativa, questa percentuale potrebbe salire al 29,5%. Ma allora, se non siamo più utili né nei lavori manuali né in quelli cognitivi, cosa ci rimane da fare? In realtà, la situazione non è così drammatica come viene descritta, anche se è vero che l'IA, perfino nelle sue versioni più

primitive, può avere conseguenze su molte mansioni, a partire da quelle più ripetitive o in cui il contributo umano non richiede particolare creatività.

Niente di più sbagliato che guardare al problema come a una competizione uomo-macchina, quando invece è possibile e auspicabile una collaborazione proficua. L'IA può infatti aiutare le persone, qualunque lavoro facciano, a farlo meglio e più velocemente, lasciando svolgere all'IA gli aspetti più semplici o ripetitivi, e usandola per affrontare al meglio quelli più complessi. ChatGPT potrebbe generare testi che contengono intere parti tratte dai libri inclusi nel gruppo di dati usato per allenare il sistema, nel caso in cui ritenesse che quei brani costituiscano il seguito più probabile dopo le parole generate in precedenza. Se questi dati sono coperti da diritto d'autore, l'IA sta generando contenuti che non dovrebbe poter utilizzare, almeno non senza citare la fonte e riconoscerne la paternità, secondo le leggi che governano il diritto d'autore.

La prima causa legale importante a questo proposito è stata intrapresa dal New York Times che, nel dicembre 2023, ha accusato OpenAI e Microsoft di usare milioni di suoi articoli per allenare ChatGPT e Copilot. Il motivo dell'accusa è chiaro: questi sistemi hanno generato risposte in cui venivano riprodotti pezzi, riassunti o anche interi articoli che provengono dal New York Times, senza averne il permesso e senza pagare per questo uso di materiale di proprietà del giornale. La causa legale chiede un rimborso per miliardi di danni e chiede anche che le aziende in questione rimuovano il materiale di proprietà del New York Times dai dati di training.

Il problema è che sia il percorso formativo sia i metodi di valutazione sono stati definiti quando non esistevano ancora strumenti di IA disponibili che gli studenti potessero usare per svolgere i compiti loro assegnati. È vero che già da tempo potevano usare materiali disponibili sul web, ma se la ricerca online di per sé non richiedeva l'impiego delle capacità critiche, queste entravano comunque in gioco in un secondo momento, quando cioè gli studenti dovevano operare una scelta delle informazioni trovate e utilizzarle per argomentare la propria tesi. Adesso invece è possibile usare sistemi come ChatGPT per generare interi temi su qualsiasi argomento e, ad oggi, non esistono metodi affidabili e accurati per riconoscere se un testo argomentativo è stato scritto da una persona o dall'IA.

Secondo uno studio compiuto a gennaio 2023, cioè solo due mesi dopo che ChatGPT era diventato disponibile, negli Stati Uniti il 30% degli studenti lo aveva usato per fare i loro compiti. E ad agosto 2023, quando lo studio è stato aggiornato, il 10% degli studenti lo aveva usato per scrivere i temi di ammissione all'università. È ovvio che uno studente che genera un tema in questo modo sta aggirando tutto il processo formativo e di crescita descritto prima. Questo potrebbe fruttare anche un bel voto allo studente, ma viene a crearsi una dissociazione tra la valutazione e la sua effettiva crescita e formazione.

Cosa fare allora? La proibizione non è mai stata una soluzione efficace e non è nemmeno chiaro come si potrà effettivamente limitare o disciplinare l'uso di questi strumenti. Si può pensare a un uso dell'IA che sia di supporto alla formazione degli studenti? Idealmente, gli studenti potrebbero usarla non per scrivere al posto loro, ma per scrivere temi migliori, per capire meglio i vari aspetti di una storia e per arrivare a produrre un elaborato finale molto più originale e creativo. Si tratta insomma, come nel campo del lavoro, di indirizzarci verso forme di collaborazione con l'IA, invece che verso una sostituzione integrale delle attività creative e intellettuali svolte dagli umani.

Dobbiamo aiutare le generazioni future a capire le opportunità e i rischi di questa tecnologia, e i valori umani che vanno protetti e supportati nell'usarla; sono poche le università che offrono corsi di etica dell'IA, o che parlano dei rischi dell'IA e di come affrontarli nel modo migliore. Insomma, nel sistema formativo l'IA è presente, ma è ancora vista solo come una tecnologia, in un ambito monodisciplinare, oppure come uno strumento utile per lo sviluppo delle altre discipline. L'etica e la gestione dei rischi, quando presenti nei corsi di IA, sono di solito una specie di appendice alla fine del corso, un argomento curioso da trattare se rimane del tempo (ma spesso non ne rimane).
Fornire agli studenti la conoscenza necessaria per saper creare IA responsabilmente e gestire i suoi rischi non basta. Come è importante che i team di programmatori di IA includano membri diversi tra loro per genere, esperienze e conoscenze, è altrettanto importante che sia così anche per gli studenti. Senza una diversità nelle discussioni socio-tecnologiche, si tende a cadere in pregiudizi inconsci che non aiutano a capire i veri rischi e come affrontarli.

Negli Stati Uniti, anche se gli studenti di pelle bianca sono ancora la maggioranza negli studi informatici, altri gruppi etnici stanno gradualmente aumentando nella popolazione studentesca: secondo l'AI Index 2024 Report, nel 2011 i bianchi erano il 71,9%, mentre sono scesi al 44,58% nel 2022. Queste sono buone notizie. Meno buone sono le notizie sul sesso degli studenti: nel 2022, il 77,78% degli studenti che avevano conseguito un dottorato in IA era di sesso maschile, con una popolazione femminile che sta crescendo ma rimane una piccola minoranza nell'educazione universitaria in IA. Simili percentuali le troviamo anche tra i docenti universitari, che sono uomini per il 75,6%.

CAPITOLO 7: I pericoli di una super-intelligenza

L'IA che viene usata oggi è molto diversa da quella che i media dipingono come una tecnologia senza limiti e completamente autonoma. Questa è quella che, nel gergo dei ricercatori di IA, viene chiamata intelligenza artificiale generale (AGI, che sta per Artificial General Intelligence), ossia una tecnologia capace di risolvere non solo alcuni problemi specifici per cui è programmata, ma tutti i problemi che incontra. Quindi, l'AGI è intesa come una forma di IA che sa fare esattamente tutto quello che sanno fare le persone, e di più. Questo era in effetti l'obiettivo che i primi ricercatori in IA si erano posti negli anni Cinquanta, prima di focalizzarsi su un approccio più limitato.

Un concetto legato all'AGI è quello di super-intelligenza. Le domande da porsi a questo punto sono soprattutto tre. La prima: saremo mai in grado di generare tecnologia AGI o super-intelligente, e quando? La seconda: se la risposta alla prima è affermativa o se, comunque, avremo una IA più potente di quella di oggi, quali sono i veri rischi per l'umanità? La terza: c'è modo di prevenire e mitigare questi rischi?

Una delle paure principali nei confronti dei sistemi di IA molto potenti è relativa al loro abuso, cioè alla possibilità che vengano utilizzati perseguendo intenzionalmente il male. I sistemi come ChatGPT sono addestrati su grandissime quantità di dati presi dal web e sanno rispondere a domande molto complesse su qualsiasi argomento. È quindi possibile che questi sistemi sappiano accedere a informazioni già disponibili sul web meglio di quanto potremmo fare noi, e forniscano istruzioni per rispondere a richieste i cui fini non siano sempre benevoli. Possiamo chiedere all'IA istruzioni per cose "innocenti", come la ricetta di un dolce o l'itinerario da seguire per raggiungere un certo posto, ma anche per cose più pericolose, come la messa a punto di un attacco cibernetico o la creazione di un'arma biologica.

Cosa potrebbe succedere a una popolazione che si affida sempre più alla tecnologia per svolgere qualsiasi attività, semplicemente perché questo rende la vita più semplice. A mano a mano che la tecnologia diventa più efficiente, possiamo essere proprio noi a scegliere di passare le redini all'IA, perché ci rendiamo conto che sa prendere decisioni migliori delle nostre. Questa tendenza è già in atto, a partire dalle calcolatrici che fanno le operazioni aritmetiche per noi, per finire con la ricerca di informazioni su web, l'uso del navigatore satellitare per trovare la strada giusta o l'analisi di grandi quantità di dati storici per le analisi di mercato. In queste attività, come in molte altre, l'IA è più brava ed efficiente di noi; quindi è naturale usarla per poterci dedicare ad altre attività più "umane".

Una terza tipologia di paure che coinvolgono l'IA riguarda l'ipotesi che questi sistemi possano sfuggire al nostro controllo. Finora, se non ci piace quello che scrive ChatGPT o le azioni di altri software di IA, possiamo semplicemente spegnerli. Ma che succede se in futuro l'IA diventa in grado di replicare sé stessa in tante copie, o di scrivere nuovi software e di eseguirli, o ancora di automigliorarsi senza l'intervento umano? Spegnere una copia non servirà, perché ce ne saranno altre; se l'IA dovesse diventare più capace di noi, e non avesse i nostri stessi limiti morali ed etici, come potremo mantenere il controllo sulla nostra vita e sul pianeta? Nel 2023 sono state pubblicate due lettere, con molti firmatari, che mettono in guardia dai rischi dell'AGI. La prima è stata pubblicata dal Future of Life Institute nel marzo 2023 e chiedeva a tutti i laboratori di IA di interrompere per almeno 6 mesi l'addestramento di sistemi di IA più potenti di GPT-4, citando tra i motivi di tale richiesta i rischi legati alla disseminazione di disinformazione, alla sostituzione delle persone in tutti i lavori e allo sviluppo di "menti artificiali" più intelligenti di noi.

Personalmente non ho firmato questa lettera, per vari motivi. Primo, il documento si focalizza appunto su sistemi di IA più potenti degli attuali, come se non esistessero rischi importanti, da affrontare con metodi efficaci, già nei sistemi IA di oggi: rischi legati all'equità, alla privacy, alla trasparenza e all'impatto sul lavoro e sull'ambiente. Inoltre, la proposta di una pausa nell'attività di training dell'IA è secondo me sia fuorviante sia dannosa. Fuorviante perché non è l'addestramento a essere la sorgente di rischi, ma l'uso che si fa dell'IA. Dannosa perché è solo incentivando, e non mettendo in pausa, le attività di ricerca e sviluppo, tra cui l'addestramento dell'IA, che possono essere individuate e messe a punto nuove tecniche che permettono di mitigarne i rischi. Infine, la pausa proposta è praticamente impossibile da implementare o da controllare, perché avrebbe senso solo se attuata da tutti i soggetti interessati e non solo da alcuni.

Quale è stato il risultato di questa lettera? La pausa non c'è stata, ma certamente la lettera e la risonanza mediatica che ha ricevuto hanno avuto l'effetto positivo di sensibilizzare l'opinione pubblica e anche i governi sui rischi di forme più sviluppate di IA e sulla necessità di definire e attuare misure per affrontare questi rischi. L'IA non è etica o non etica, ma è l'intero ecosistema intorno all'IA, e quindi l'insieme dei vari soggetti interessati, che deve comportarsi in modo etico. L'IA attuale non rappresenta un rischio esistenziale, non è come la bomba atomica, non è un asteroide, non è una persona, e non siamo vicini all'AGI o alla super-intelligenza.

L'IA e l'ambiente

Le tecniche di machine learning, soprattutto quelle alla base dell'IA generativa, richiedono grandi quantità di dati di training per funzionare bene. Durante la fase di training, l'IA deve considerare un dato alla volta per calibrare i "pesi" della rete neurale e quindi migliorare il proprio comportamento. Per effettuare questa fase in tempi ragionevoli, sono quindi necessari computer molto veloci. I dati sappiamo dove prenderli: vengono da Internet, dove li carichiamo noi sotto forma di post, like, immagini, video ecc. I computer sono diventati sempre più veloci grazie anche a speciali strutture hardware, le GPU (Graphics Processing Units). Ma nonostante l'uso di un hardware molto veloce, il training di un large language model può durare molti mesi.

Secondo alcuni ricercatori di OpenAI (l'azienda che ha prodotto ChatGPT), dal 2012 la quantità di potenza computazionale usata per la ricerca in IA è raddoppiata ogni 3,4 mesi. Quando OpenAI ha generato GPT-3, sono state generate circa 500 tonnellate di CO2. La CO2 è il principale responsabile, tra i gas serra, del cambiamento climatico, in quanto intrappola calore nell'atmosfera. Si stima che entro il 2040 le emissioni dell'industria tecnologica raggiungeranno il 14% del totale globale. Il training di GPT-3 ha anche consumato 700.000 litri di acqua: acqua che evapora a causa del riscaldamento dei computer, e che quindi non può essere riusata.

Nonostante ciò, non ci sono solo aspetti negativi. L'IA può anche essere un utile alleato per affrontare il cambiamento climatico. Anche le Nazioni Unite considerano l'IA uno strumento che può aiutare a capire meglio l'impatto ambientale di varie attività antropiche e i loro effetti sul cambiamento climatico. L'IA può ad esempio analizzare grandi quantità di dati, come le immagini satellitari che i ricercatori usano per monitorare i cambiamenti del clima. Aiutati dall'IA, gli scienziati possono creare modelli più precisi per descrivere l'evoluzione del clima, identificare gli andamenti nel tempo e fare predizioni, in modo da definire strategie di mitigazione più efficaci. E così può anche essere usata per capire come sprecare meno risorse naturali, come prevenire e risolvere gli incendi, identificare materiali riciclabili e ottimizzare le reti di produzione dell'energia.

CAPITOLO 8: Come si controlla l'IA

Nel 2020 un progetto dell'Università di Harvard ha cercato di raccogliere tutti i principi che erano stati pubblicati dal 2016 da vari organismi, incluse aziende, governi e altre istituzioni, contandone più di 100 su temi come la privacy, la sicurezza, la trasparenza, l'equità, il controllo umano della tecnologia e la promozione dei valori umani. Nel 2020 il Berkman Klein Center dell'Università di Harvard ha pubblicato uno studio intitolato Principled Artificial Intelligence in cui sono stati analizzati 36 documenti che definiscono principi per l'uso e lo sviluppo etico e socialmente utile dell'IA. Si tratta di documenti pubblicati tra il 2016 e il 2019 che provengono da soggetti molto diversi tra loro: governi (verde), società civile (giallo), agenzie intergovernative (arancione), aziende private (fucsia) e iniziative multistakeholder (azzurro). Dalla comparazione dei testi sono

emersi nove temi chiave ricorrenti, tra cui particolare rilevanza sembra avere la tutela internazionale dei diritti umani.

I principi sono importantissimi. Danno le linee guida, servono come ispirazione e riassumono in modo conciso gli aspetti più importanti da considerare. Ma quella che è la loro caratteristica più importante, cioè la capacità di esprimere posizioni generali e largamente condivisibili, rappresenta anche un loro limite. Per questo non vanno presi come un punto di arrivo, ma piuttosto come un punto di partenza per definire azioni concrete specifiche, che vari soggetti possano eseguire. Ad esempio, se un principio condiviso è che l'IA non deve creare discriminazioni, cosa significa questo per i programmatori?

Uno studio IBM del 2022 ha chiesto a più di 1.000 aziende in 22 paesi nel mondo cosa stessero facendo per mitigare i rischi dell'IA. Ne è emerso che più del 50% delle aziende aveva sottoscritto principi per l'etica dell'IA, ma meno del 25% aveva messo in piedi azioni concrete al riguardo. Quindi c'è ancora una distanza significativa tra le intenzioni e le azioni; quello che serve è un approccio che coinvolga l'intera azienda e non solo la parte che crea tecnologia.

Se tutti i membri di un team che programma un sistema di IA sono maschi, potrebbero avere difficoltà a riconoscere possibili sorgenti di discriminazione nei confronti delle donne (e questo vale per tutte le differenze, naturalmente). Avere un team in cui sono presenti membri "diversi" tra loro aiuta a riconoscere i propri pregiudizi e quelli degli altri membri e a evitare che in modo inconscio questi vengano inseriti nella progettazione e la programmazione di un sistema di IA. In fondo, l'IA discrimina soprattutto perché noi umani siamo i primi a creare discriminazioni, a volte anche senza rendercene conto. Quindi è fondamentale agire sul capitale umano oltre che sulla tecnologia stessa.

Quello che serve è una governance centralizzata che sia visibile e ben collegata con tutte le parti dell'azienda. Il difetto di un team è che non riesce di solito ad avere un impatto significativo su tutta l'azienda per mancanza di collegamenti diretti e di potere decisionale. Le altre divisioni finiranno per essere meno sensibili a questi temi e a adottare meno soluzioni e controlli sull'IA. Ciò che può assicurare un impatto più significativo è invece un meccanismo centrale con potere decisionale sulle azioni di tutte le divisioni. Quando si verificano trasformazioni significative della società, è essenziale che tutti i soggetti interessati lavorino insieme: aziende, governi, istituzioni accademiche e anche organismi della società civile. È insomma necessario adottare quello che si chiama un approccio multistakeholder.

Cosa stanno facendo i governi sull'IA? A questo punto, praticamente tutti i governi del mondo hanno pubblicato una strategia nazionale per l'IA e stanno pensando a leggi mirate al controllo dell'uso della tecnologia e al contempo al supporto dell'innovazione tecnologica. Oltre a nuove leggi sull'IA, i governi mettono in campo anche altri meccanismi, quali investimenti in ricerca e sviluppo in IA, facilitazioni per accesso e condivisione di dati, creazione di infrastrutture per l'IA, ambienti per i test di nuovi sistemi di IA, fondi per le start-up di IA e accesso a piattaforme collaborative per imprese e università.

L'Unione Europea è particolarmente attiva nel definire nuove leggi pensate per mitigare i rischi dell'IA. Per proteggere la privacy dei dati, la legge più rilevante è la GDPR (General Data Protection Regulation), entrata in vigore nel 2016 e incentrata sui diritti di chi fornisce i propri dati. Anche se con un po' di confusione e alcune controversie, la GDPR è molto visibile e influente sia in Europa che in altre zone del mondo, anche grazie alle multe previste per chi non la rispetta, che possono arrivare fino al 4% del fatturato mondiale di un'azienda.

Partendo dall'analisi della Carta dei diritti fondamentali dell'UE e dei principi per l'IA, il documento del gruppo di esperti ha individuato sette requisiti per l'IA affidabile, che includono il controllo umano, la sicurezza dei sistemi di IA, la privacy, la trasparenza e la possibilità di spiegare le decisioni prese dall'IA, la non-discriminazione e il supporto al benessere della società e dell'ambiente. L'AI Act, la cui prima versione è stata pubblicata ad aprile 2021, prevede quattro livelli di rischio per le applicazioni di IA. Il livello più alto, chiamato del rischio inaccettabile, include applicazioni che non saranno permesse in Europa. Il secondo livello, detto ad alto rischio, include scenari applicativi come ad esempio l'uso di IA per la sicurezza di infrastrutture critiche, o per la gestione del personale, o per attività legate alla formazione, per i quali è necessario usare l'IA affidabile.

Per le applicazioni a rischio minore o nullo, la legge impone delle richieste di trasparenza, come il fatto che un chatbot deve dichiarare di essere tale nelle interazioni con utenti umani. Nel suo insieme, l'AI Act intende regolamentare di più dove c'è più rischio, e associare il rischio agli usi dell'IA e non alla tecnologia stessa. Gli Stati Uniti di solito tendono a non legiferare sulle nuove tecnologie, per non intralciare l'innovazione tecnologica e il suo uso nella società. Ma con l'IA le cose stanno andando diversamente. Vari Stati e anche città americane hanno fatto da apripista, pubblicando leggi di controllo su alcuni usi specifici dell'IA.

Un esempio è la legge della città di New York, entrata in vigore nella primavera 2023, che richiede che le aziende che vogliono usare l'IA per decisioni sulle assunzioni o promozioni debbano effettuare un audit, cioè una verifica esterna dei comportamenti di questi sistemi, con particolare riferimento alla possibile presenza di discriminazioni. Un altro esempio è la legge California Privacy Rights Act, entrata in vigore nello Stato della California a gennaio 2023, che vuole offrire ai consumatori un maggiore controllo sulle proprie informazioni personali raccolte dalle aziende, mutuando molti elementi dal GDPR europeo.

EPILOGO

Il punto non è contenere l'IA, come sostiene Mustafa Suleyman, o addirittura di fermarla, come suggeriscono altri, ma piuttosto creare un ecosistema intorno a questa tecnologia che assicuri che il suo comportamento e il suo uso siano allineati ai valori che riteniamo importanti da proteggere. Se non ci sarà un allineamento, questi valori potrebbero essere danneggiati e la società potrebbe andare verso uno dei tanti futuri distopici di cui spesso leggiamo sui media, dove l'umanità ha perso il controllo sul suo futuro. Si tratta, insomma, di creare un ecosistema che garantisca il rispetto dei valori umani imprescindibili.

Ma quali valori? Chi li decide? E chi decide se l'IA è allineata a questi valori? Nel 2015, tutti gli Stati membri dell'ONU hanno adottato l'Agenda 2030 per lo sviluppo sostenibile, che include 17 obiettivi. Presi insieme, essi definiscono un futuro a cui mirare, possibilmente entro il 2030. Coprono aspetti come la fine della povertà, l'accesso all'educazione e alla salute per tutti, la riduzione delle disuguaglianze, la crescita economica, la soluzione dei problemi climatici e la pace. Sono valori che possiamo definire universali e come tali condivisibili, e ritengo che sia un auspicio di tutti arrivare a un futuro così fatto.

Negli anni, l'IA è stata usata per favorire questi obbiettivi. Ad esempio, per accelerare la scoperta di nuove medicine, per rendere il sistema educativo più efficiente e inclusivo, per ottimizzare il consumo di energia e per accelerare le azioni contro il cambiamento climatico. Non andiamo veloci come dovremmo, ma si stanno facendo passi in quella direzione. La tecnologia è al servizio del progresso umano e non viceversa. Quando usiamo l'IA dobbiamo avere sempre presente che il fine ultimo non è la tecnologia stessa, ma l'umanità e la sua crescita.