Mario Verdicchio

Che cosa genera davvero l'IA generativa?

Stiamo vivendo un momento di grande subbuglio attorno al concetto di intelligenza artificiale (IA), un clamore suscitato da alcuni notevoli successi nell'applicazione di questa tecnologia a imprese tradizionalmente considerate prettamente umane, poi cresciuto a dismisura sull'onda di un battage mediatico senza precedenti.

In questo turbinio, in cui si incontrano e scontrano discorsi tecnologici, psicologici, filosofici, culturali e anche politici, serve innanzitutto chiarezza. Serve un impianto concettuale ben delineato e condiviso tra coloro che partecipano attivamente a tali discorsi. Si tratta di una sfida non semplice, proprio a causa della vastità dei contesti coinvolti e dei loro rispettivi partecipanti: i pensieri, gli obiettivi e i modi di persone come, ad esempio, un adolescente che costruisce una parte significativa della sua vita sociale su TikTok da una parte, e un legislatore ultrasettantenne da decenni specializzato in questioni di proprietà intellettuale dall'altra, non potrebbero essere più diversi.

Il recente impatto delle tecnologie informatiche sulla società è stato talmente vasto e soprattutto trasversale che è stato imbastito un notevole numero di interrogazioni e audizioni allo scopo di creare a livello legislativo un sistema di protezione per limitare i potenzialmente catastrofici danni di una tecnologia fuori controllo. È proprio qui che il bisogno di un vocabolario comune si fa più sentire. Una commistione non ragionata di persone provenienti da parti molto diverse e distanti di questa galassia socio-tecnologica non può essere fruttuosa. Come avvenne nel 2018, quando Mark Zuckerberg,

l'amministratore delegato di Facebook (oggi nota come 'Meta'), fu il protagonista di un'interrogazione da parte del Senato statunitense che fu definita dai giornalisti più misericordiosi 'strana' e 'confusa' (Gutman-Wei, 2018). Le questioni sul tavolo erano numerose: la privacy, l'influenza sull'opinione pubblica, un potenziale monopolio... ma alla fine quello che è emerso maggiormente è stata la mancanza di una effettiva comunicazione tra le due parti. Questo buco nell'acqua fu allora attribuito all'eccessiva differenza di età, di circa 50 anni, tra l'interrogato e gli interroganti, i quali non avrebbero avuto una conoscenza sufficientemente approfondita dei meccanismi del social network per poter porre domande significative al suo amministratore.

Il problema, purtroppo, sembra essere destinato ad ampliarsi, perché con i nuovi sistemi IA che si stanno maggiormente diffondendo oggi, la mancanza di comprensione tra gli attori coinvolti non è più solo una questione di differenze di età e di abitudini, bensì di vera e propria conoscenza approfondita della tecnologia in uso. Per capire perché un certo artefatto IA abbia dato un determinato risultato, e quindi per poter valutare le cause e i meccanismi di tale funzionamento ed eventualmente attribuire responsabilità di malfunzionamenti o correggere e migliorare l'artefatto in modo che tali incidenti non si ripetano, occorre essere degli esperti non solo di IA, ma anche di molto altro. Non si tratta più semplicemente di seguire l'esempio dei più giovani e familiarizzarsi con l'interfaccia grafica e i meccanismi di interazione di una app, ma di avere una visione assieme vasta e profonda del funzionamento di un artefatto IA. Vasta come la rete che interconnette i diversi contesti che determinano tale funzionamento, e profonda come la gerarchia di livelli di astrazione attraverso cui esso si sviluppa.

In questo senso, si sono avuti negli ultimi anni almeno due casi a dir poco clamorosi, in cui non degli anziani senatori poco avvezzi agli smartphone, bensì gli stessi creatori di sistemi IA si sono ritrovati in grandissima difficoltà di fronte a risultati decisamente inaspettati forniti da prodotti che loro stessi avevano messo sul mercato. Il fatto che tali persone lavorassero per Google e Microsoft dovrebbe metterci tutti in guardia: nemmeno le

Nemmeno le più grandi aziende che lavorano con l'IA sembrano conoscerla

più grandi aziende che lavorano con l'IA sembrano conoscerla in maniera completa. Che cosa è successo?

Nel 2015, Google arricchì la sua app per la gestione di fotografie con un sistema di classificazione automatica che aveva lo scopo di etichettare le foto su uno smartphone con parole chiave basate sul contenuto delle foto stesse. Negli Stati Uniti una coppia di ragazzi di colore si è vista etichettata come 'gorilla', dando luogo al primo caso di disgustoso insulto razzista creato da un'IA (BBC, 2015). Nel 2016, meno di un anno dopo questo incidente, toccò a Microsoft, che lanciò sul social network Twitter (ora chiamato 'X') un chatbot di nome Tay, ossia un software specializzato nel dialogo con gli esseri umani, perché interagisse con gli utenti di Twitter tramite tweet e relative risposte, imparando da queste interazioni a scrivere testi sempre più sofisticati e rilevanti. Dopo meno di 24 ore, l'azienda è stata costretta a ritirare Tay da Twitter dopo che aveva già scritto decine di tweet in cui inneggiava contro il femminismo, approvava l'Olocausto e scriveva oscenità di vario genere (Hunt, 2016).

Questi incidenti sembrano puntare verso i peggiori scenari dei film di fantascienza su un futuro distopico in cui una tecnologia malvagia prende coscienza e sopraffà gli esseri umani. Questa estrapolazione è molto semplice da eseguire, almeno nelle nostre fantasie: se l'IA, già difficilmente comprensibile nelle sue forme più semplici ai potenti di turno, diventa imprevedibile anche agli occhi dei suoi stessi creatori nelle sue versioni più all'avanguardia, che cosa succederà quando, anziché classificare immagini o scrivere tweet, verrà usata per controllare centrali elettriche o, peggio ancora, sistemi di difesa e armamenti? Dal momento che non ci sono solo film su questo tema ma anche saggi di futurologi e filosofi della tecnologia talmente rinomati da poter essere consultati da governi riguardo a strategie e decisioni intorno all'IA (Bostrom 2018), il bisogno di chiarezza diventa un dovere morale: dobbiamo tutti comprendere meglio che cos'è davvero l'IA e che cosa intendiamo quando diciamo che essa è 'generativa' di risultati.

Questi incidenti sembrano puntare verso un futuro distopico

Per analizzare al meglio la tecnologia che ha messo in imbarazzo Google e Microsoft, e che un giorno potrebbe mettere a rischio l'intera umanità, occorre fare un salto indietro nel tempo, agli albori dell'IA, quando i computer e i programmi erano decisamente più primitivi

e semplici di quelli odierni, ma i semi concettuali che danno frutti ancora oggi furono piantati.

Il termine 'intelligenza artificiale' fu utilizzato per la prima volta nel 1955 in una proposta per un workshop estivo al Dartmouth College (nel New Hampshire, Stati Uniti) da parte di coloro che oggi sono considerati i fondatori della disciplina. In tale proposta espressero l'intenzione di lavorare «sulla base della congettura che ogni aspetto dell'apprendimento o qualsiasi altra caratteristica dell'intelligenza possa in linea di principio essere descritta in maniera talmente precisa che si possa costruire una macchina per simularla» (McCarthy et al., 2006, p.12, traduzione mia).

Cominciamo dal concetto di intelligenza utilizzato nella proposta. Che 'l'apprendimento' sia una caratteristica dell'intelligenza non sembra dar luogo a controversie, poiché tale presupposizione è presente in molti altri campi del sapere umano, come ad esempio la psicologia o la pedagogia. Molto più problematica, invece, è l'altra considerazione sull'intelligenza, secondo la quale l'intelligenza è suscettibile di descrizioni precise e compatibili con una macchina, almeno in termini di simulazione. Uno degli autori della proposta, John McCarthy, è rimasto fedele a questo assunto per tutta la sua vita, come dimostrato da un manifesto sotto forma di domande e risposte da lui pubblicato decenni dopo il suo lavoro pionieristico sull'IA, in cui la sua risposta alla domanda 'cos'è l'intelligenza?' è la seguente: «l'intelligenza è la parte computazionale della capacità di raggiungere obiettivi nel mondo. Vari tipi e gradi di intelligenza si riscontrano negli esseri umani, in molti animali e in alcune macchine» (McCarthy, 2007, p. 2, traduzione mia). Si tratta di un'affermazione molto forte, perché dà una connotazione estremamente precisa al concetto di intelligenza: essa è computazionale, ossia consiste nella descrizione dei problemi in termini numerici e nell'esecuzione di operazioni numeriche per raggiungere la loro soluzione.

Questa posizione appare coerente con il notevole sviluppo delle tecnologie informatiche nella seconda metà del XX secolo, il periodo che separa il primo manifesto di McCarthy dal secondo: le macchine calcolatrici elettroniche sono sempre più diffuse ed evolute, e forniscono una gamma sempre più vasta di servizi; ha senso, quindi, vedere questa tecnologia come sempre più 'intelligente'.

La problematicità di una tale prospettiva, però, emerge nel momento in cui facciamo paragoni con la concezione più tradizionale

di intelligenza, la quale, pur essendo stata sempre molto vaga, ha un suo riferimento ben concreto e preciso: l'essere umano. Noi siamo intelligenti solo perché sappiamo fare di conto? Decisamente no: pensiamo a una persona in grado di eseguire operazioni di calcolo a notevole velocità senza mai commettere errori ma incapace di apprendere le regole più elementari di convivenza con le altre persone in società. Verrebbe tale persona considerata 'intelligente'? Un 'genio matematico', probabilmente, ma non molto altro. Qui entra in gioco l'aggettivo 'artificiale': dobbiamo chiederci in che senso esso qualifica l'intelligenza nell'IA.

Noi siamo intelligenti solo perché sappiamo fare di conto? Decisamente no…

Possiamo immaginare almeno due interpretazioni dell'artificialità dell'intelligenza: parziale e universale. Tale distinzione emerge da un confronto della visione computazionale degli albori dell'IA con altre teorie sull'intelligenza, tra cui spicca quella delle intelligenze multiple dello psicologo Howard Gardner (2022), che analizza e distingue l'intelligenza umana in otto modalità specifiche: visivo-spaziale (la capacità di percepire accuratamente il mondo visivo e di eseguire trasformazioni e modifiche sulle proprie percezioni iniziali tramite immagini mentali), linguistico-verbale (la capacità di esprimersi in modo chiaro e pertinente con un vocabolario ampio e creativo per via orale o scritta), logico-matematica (la capacità di risolvere operazioni matematiche, individuare nessi logici, sperimentare idee e sviluppare argomentazioni logiche), corporeo-cinestetica (la capacità di controllare i propri movimenti corporali e di manipolare oggetti con abilità), musicale (la capacità di ascoltare e riprodurre con voce e strumenti suoni, note, ritmi e forme musicali), interpersonale (la capacità di comprendere gli altri e di interagire efficacemente con comunicazione verbale e non verbale), intrapersonale (la capacità di conoscersi e controllarsi, identificando i propri umori, sentimenti e pensieri) e, infine, naturalistica (la capacità di individuare ed entrare in contatto con la natura che circonda).

Anche senza esperimenti scientifici convincenti, possiamo convenire sul fatto che queste capacità caratterizzano in varia misura tutti gli esseri umani e, quando sono presenti in maniera spiccata in un individuo, suscitano ammirazione e sono alla base di numerose manifestazioni della cultura umana (concerti, eventi sportivi,

romanzi, e molto altro ancora). Come, dunque, si relaziona il concetto di 'artificiale' con queste forme di intelligenza? Nell'interpretazione parziale, i lavori che si fanno nell'IA si concentrano esclusivamente sull'intelligenza logico-matematica, ossia la capacità di eseguire operazioni matematiche: dal momento che le macchine dell'IA sono computer, ossia calcolatori, lavoriamo per affidare loro problemi di tipo logico-matematico sempre più complessi. Già da anni i computer superano gli esseri umani in termini di potenza di calcolo, con la capacità di eseguire miliardi di operazioni aritmetiche semplici al secondo. Ha senso, quindi, avvalersi di queste macchine per portare a termine nella maniera più efficiente possibile compiti espressi in termini numerici.

Qui troviamo il punto di incontro e scontro con l'altra interpretazione dell'IA, quella universale. La scelta tra incontro e scontro dipende dalla risposta che diamo alla seguente domanda: possono tutti i problemi essere espressi in termini numerici? Se la risposta è negativa, allora ci sono aspetti dell'intelligenza non passibili di una descrizione computazionale e, quindi, inaccessibili ai computer elettronici. In tal caso, c'è una chiara distinzione tra l'intelligenza umana e quella artificiale. Si intersecano in campo matematico, ma ci sono dimensioni del sapere che possono esistere solo nei cervelli umani. Quindi, quando parliamo di IA, la intendiamo secondo l'interpretazione parziale. Se, invece, la risposta è positiva, ossia tutti i problemi possono essere riportati all'ambito matematico, allora il ruolo dell'IA ha la potenzialità di espandersi a tutto lo scibile umano. Ciò che distingue l'intelligenza umana da quella artificiale è semplicemente una questione di realizzazione di modelli: i problemi già modellati in termini numerici possono essere affidati ai sistemi IA, quelli non ancora modellati in tal senso devono per il momento essere gestiti dagli esseri umani. L'IA, quindi, va interpretata in maniera universale: è la versione artificiale, ossia realizzata per mezzo di macchine, di tutta l'intelligenza umana.

Qual è la risposta corretta? Non si sa, e la questione è oggetto di accesi dibattiti in vari ambiti: filosofia, psicologia, neuroscienze e, naturalmente, anche informatica. In generale, chiamiamo 'computazionalismo' o 'teoria computazionale della mente' la posizione che propende per il sì, ipotizzando che si possa descrivere l'intelligenza umana in termini di operazioni di calcolo, potenzialmente realizzabili con un computer (Rescorla, 2020).

Qualunque sia la risposta, l'evoluzione tecnologica nell'ambito IA sembra puntare a espandere gli aspetti dell'intelligenza che possano essere trattati per mezzo di artefatti tecnologici computazionali. I sistemi IA attuali non sono più solo computer che effettuato operazioni matematiche, ossia sistemi intelligenti in senso esclusivamente logico-matematico. Il meccanismo alla base di questa espansione è l'accoppiata codifica/decodifica alla base di tutte le tecnologie informatiche: la prima per ottenere una descrizione numerica di un fenomeno fisico (come, ad esempio, la mappatura di uno spazio da esplorare su coordinate cartesiane), la seconda per trasformare i risultati dei calcoli di un computer in azioni concrete nell'ambiente (per esempio, controllare i pixel di un monitor con segnali numerici per far loro emettere una luce di un colore specifico). Con l'aggiunta di periferiche e accessori adatti per codificare la realtà fisica e decodificare i numeri del computer, sono stati costruiti robot in grado di orientarsi nello spazio e manipolare oggetti (intelligenze visivo-spaziale e corporeo-cinestetica) e, con un adeguato accesso ai miliardi di dati testuali presenti online, sono stati realizzati programmi, come l'oramai celeberrimo ChatGPT, che producono testi originali di ogni genere (intelligenza linguistico-verbale).

I sistemi IA attuali non sono più solo computer che effettuato operazioni matematiche, ossia sistemi intelligenti in senso esclusivamente logico-matematico

ChatGPT è il più recente sistema IA che aggiunge materiale di riflessione attorno alla domanda sulla natura dell'intelligenza, perché i suoi risultati sono davvero notevoli dal punto di vista della produzione testuale: saggi, poesie, soluzioni a domande d'esame, traduzioni, rapporti tecnici, creati automaticamente a partire da semplici indicazioni da parte degli utenti, che stanno mettendo in crisi il sistema scolastico (Yu, 2023) e certe parti del mondo del lavoro (Wach et al., 2023). La 'G' di GPT sta per 'generative' in inglese, 'generativo' in italiano: un ulteriore aggettivo il cui significato merita un'analisi approfondita per evitare di essere travolti dal turbinio mediatico intorno l'IA, anche perché, nella realtà della ricerca lontana dal marketing di certe aziende e dal sensazionalismo di certi media, ci si interroga su questo concetto da decenni, se non secoli.

La prima persona ad aver posto la questione in termini molto chiari,

ancora prima che i computer come li conosciamo oggi esistessero, è stata la matematica inglese Ada Lovelace, la quale, di fronte ai primi prototipi di telai modificati in macchine calcolatrici meccaniche, obiettava che tali artefatti sarebbero anche potuti diventare così potenti da eseguire qualunque cosa noi fossimo in grado di ordinare loro di eseguire, ma non avrebbero mai avuto la capacità di una persona che è in grado di trovare delle vie alternative alla soluzione quando si verifica un imprevisto (Lovelace, 1843). In altre parole, non possiamo aspettarci nulla di nuovo, di sorprendente, di originale da un dispositivo come il computer, il quale è programmato, ossia per la sua stessa natura (meccanica ai tempi di Lovelace, elettronica oggi) non può che eseguire le istruzioni che gli vengono date.

Un programma come ChatGPT, però, mostra un comportamento che sembra smentire tale presa di posizione: i testi che produce sono originali, almeno nel senso di essere una sequenza di parole mai prodotta prima, e spesso sorprendono gli utenti per la loro creatività ed efficacia. Questa è proprio la caratteristica di quei sistemi che vengono chiamati 'generativi'. Una delle definizioni più famose di questo concetto proviene dal mondo della Computer Art, o arte realizzata per mezzo di strumenti informatici, dove l'artista e studioso Philip Galanter scrisse che «l'arte generativa si riferisce a qualsiasi pratica artistica in cui l'artista utilizza un sistema, come un insieme di regole del linguaggio naturale, un programma per computer, una macchina o altro artefatto di tipo procedurale, che viene lanciato in una esecuzione caratterizzata da un certo grado di autonomia, che contribuisce alla creazione di in un'opera d'arte» (Galanter, 2003, p.4, traduzione mia).

È proprio il contesto della Computer Art a fornirci il primo esempio di sistema informatico generativo mai realizzato. Nel 1965, un decennio dopo che il termine "intelligenza artificiale" era stato coniato, per la prima volta nella storia, delle stampe eseguite per mezzo di calcolatori elettronici furono esposte in una galleria d'arte a Stoccarda, in Germania. Si trattava dei lavori di Frieder Nake e Georg Nees, oggi salutati come pionieri della Computer Art, ma allora considerati dal mondo dell'arte come meri matematici che hanno usato dei computer in maniera insolita. In particolare, un lavoro di Nake intitolato *Random Polygons* è molto utile per comprendere il concetto di arte generativa: si tratta di un disegno estremamente semplice, in cui una linea spezzata in più punti si snoda su un foglio

bianco. Dal punto di vista estetico, forse, questa opera non presentava nulla di particolarmente entusiasmante, ma il modo in cui era stata realizzata aveva caratteristiche rivoluzionarie. A parte la novità dell'esposizione in una galleria d'arte del risultato di un computer, quello che rendeva un lavoro come *Random Polygons* speciale era il modo in cui Nake aveva programmato la macchina perché desse questo risultato: la posizione dei punti in cui la linea si spezza nel disegno non erano stati scelti da Nake, bensì erano stati calcolati dal computer durante l'esecuzione del programma. È per questo che abbiamo l'aggettivo 'random' nel titolo dell'opera: la forma in essa rappresentata non è stata decisa dall'artista, bensì è stata creata dal caso, o meglio, da una procedura che risulta incomprensibile al programmatore della macchina. Il caso all'interno di un computer programmato non esiste, ma se gli diamo istruzioni che dipendono da parametri su cui non abbiamo né controllo né visibilità, allora i risultati di tali istruzioni potranno sorprenderci. Ad esempio, se vogliamo disegnare un segmento e vogliamo far decidere al computer la coordinata x del punto di inizio del segmento, possiamo istruire la macchina perché prenda il numero di millisecondi segnato dall'orologio al suo interno nel momento in cui il segmento sta per essere tracciato. A che coordinata x si troverà il punto di inizio di questa composizione? Il programmatore stesso deve attendere la stampa per saperlo.

Galanter parlò di 'autonomia' del sistema perché si possa parlare di 'generatività': l'essere umano non è in grado di prevedere il risultato delle operazioni della macchina; quindi, è come se la macchina avesse agito in maniera autonoma. Si tratta, però, di un'illusione: il computer ha comunque e come sempre seguito le istruzioni con cui è stato programmato; non c'è autonomia da nessuna parte. Semmai, c'è ignoranza da parte dell'essere umano, come sottolineato da una definizione contemporanea a quella di Galanter ma alternativa, elaborata nell'ambito di un lavoro sulla creatività nell'IA da parte del filosofo Selmer Bringsjord e colleghi: «una relazione epistemica restrittiva tra un agente artificiale A, il suo output O, e l'architetto umano H di A – una relazione che,

Si tratta, però, di un'illusione: il computer ha comunque e come sempre seguito le istruzioni con cui è stato programmato; non c'è autonomia da nessuna parte

grosso modo, si ottiene quando H non è in grado di spiegare come A abbia prodotto O» (Bringsjord et al., 2003).

Nella sua geniale semplicità, il lavoro pionieristico di Nake in ambito artistico ci aiuta a capire l'IA generativa di oggi. La definizione di creatività artificiale va accompagnata da ulteriori considerazioni. È ragionevole immaginare che quando Nake vide la stampa di *Random Polygons* non trasecolò: l'operazione prevista dal calcolo gli era nota, anzi, lui stesso aveva deciso di programmare il computer perché disegnasse una linea spezzata, ma affidò ai meccanismi interni della macchina la determinazione dei dettagli di tale linea.

Quello che non poteva prevedere è, come già spiegato, la posizione esatta dei punti in cui la linea si sarebbe spezzata, perché si trattava di un parametro a lui sconosciuto. Riprendendo la definizione di Bringsjord, è esagerato affermare che il programmatore umano non è in grado di spiegare come la macchina programmata abbia prodotto il risultato. Sarebbe più preciso dire che, quando facciamo dipendere alcune operazioni che diamo in esecuzione alla macchina da parametri che vengono stabiliti in maniera fuori dal nostro controllo, allora non siamo in grado di prevedere con la massima precisione il risultato della computazione.

Oggi l'IA generativa ci stupisce molto più di quanto *Random Polygons* avesse stupito Nake perché i parametri fuori dal nostro controllo sono cresciuti a dismisura, non solo in quantità ma anche in qualità. Tale crescita è legata a un altro fondamentale strumento tecnologico che ha preso avvio proprio nei primi anni dell'IA e della Computer Art e ha finito per supportarle entrambe in maniera determinante: la rete di telecomunicazione e scambio dati di Internet. L'apprendimento a cui McCarthy fa riferimento nella sua prima definizione di IA è il meccanismo alla base dei sistemi generativi che oggi creano testi e immagini. Tale 'apprendimento' consiste nell'analisi statistica automatica di quantità significative di dati all'interno di un preciso contesto determinato dai programmatori della macchina analizzatrice, alla ricerca di correlazioni tra dati con lo scopo di scoprire schemi ricorrenti che possano assurgere a modello matematico del fenomeno studiato in tale contesto. I meccanismi di base sono quelli fondamentali dell'informatica: codifica per la descrizione numerica di un fenomeno fisico, operazioni matematiche sui numeri così ottenuti, decodifica per riportare i risultati di tali operazioni nel mondo reale. La differenza introdotta dalla connettività

di Internet consiste nella tipologia di operazioni che le macchine sono in grado di eseguire: non più operazioni puntuali per risolvere un problema locale, ma analisi di dati a livello globale per costruire modelli numerici di tutto ciò che attraversa la rete. Gli esperti di IA hanno quindi potuto costruire, ad esempio, modelli computazionali di visi umani (Booth et al., 2016), di regole dei linguaggi naturali (Jurafsky e Martin, 2023), e di stili di pittura (Lecoutre et al., 2017). È grazie a questi modelli, costruiti sulla base di miliardi di dati online, che i computer di oggi possono essere programmati per creare visi di persone mai esistite, testi coerenti ma senza autore, disegni nello stile di artisti morti da tempo, e molto altro ancora.

Con la complessità di questi modelli, però, aumenta a dismisura anche l'ignoranza di coloro che li usano: quale essere umano è in grado di comprendere, gestire e correggere rappresentazioni numeriche con milioni di parametri? Anche gli ingegneri più esperti, a questo punto, possono solo accontentarsi di controllare la qualità dei risultati delle macchine da loro programmate e, ed è qui che si nasconde l'aspetto critico di questa impresa, modellate dai dati provenienti dalla rete. Torniamo, quindi, ai famigerati incidenti di Google e Microsoft, accomunati dall'avere parametrizzazioni senza un adeguato controllo da parte dei programmatori sui dati in arrivo, anche se tale inadeguatezza si è manifestata in due modi diametralmente diversi. Nel caso di Google, si è scoperto che i programmatori hanno creato modelli matematici dei visi umani solo basandosi sul personale presente nella divisione dell'azienda dedita allo sviluppo del sistema IA per l'etichettatura delle immagini. Si trattava di persone esclusivamente di etnia caucasica, il che comportò la creazione di un modello capace di riconoscere persone bianche, ma senza parametri adeguati alle altre. I risultati imbarazzanti della app erano, quindi, dovuti a modelli costruiti male a causa di troppo pochi dati. Viceversa, il chatbot di Microsoft ha ricevuto numerosi dati con cui allenare il proprio modello computazionale linguistico per la formulazione di tweet. Il problema è stato che molti utenti di Twitter hanno voluto tirare un brutto scherzo all'azienda e hanno fornito dati tendenziosi, caratterizzati da fortissime connotazioni razziste, sessiste e antisemitiche, e tali caratteristiche sono state apprese dall'IA del chatbot, che le ha riprese nei suoi tweet, creando scandalo. Questa volta, il modello non ha sofferto per la scarsa quantità dei dati che lo hanno allenato, ma per la loro scarsa qualità.

Un'altra particolarità che i due casi hanno in comune è strettamente legata a questo approccio statistico dell'IA generativa: poiché i creatori del modello computazionale non sono in grado di interpretare il significato dei singoli parametri numerici che lo caratterizzano, non hanno nemmeno modo di capire dove si annidano i problemi all'interno del sistema. Tra le decine di migliaia di parametri della funzione di riconoscimento dei visi della app di Google, dove si trovano quelli che hanno determinato l'associazione razzista? All'interno del modello del linguaggio del chatbot di Microsoft, dove sono i componenti che hanno portato alla creazioni di tweet sessisti? Il processo di codifica, che trasforma le parole e le regole grammaticali rispettivamente in numeri e funzioni numeriche, offre il vantaggio di rendere queste operazioni compatibili con il funzionamento di un computer ma, naturalmente, rende il sistema sempre più incomprensibile agli occhi dei programmatori umani, soprattutto se non stanno lavorando con programmi semplici come quelli degli anni '60 che hanno creato arte generativa, ma con software estremamente complesso che interagisce con milioni di utenti ed elabora miliardi di dati. Questo è il motivo per cui sistemi IA di questo genere, quando danno risultati inadeguati, non possono essere corretti o ritoccati: il modello va ricostruito da zero con nuovi dati.

L'IA generativa, come dicono le definizioni di generatività e di creatività in ambito informatico, sembra essere davvero in grado di riservare delle sorprese. Nel recente passato risultati inspiegabili dagli stessi creatori dei sistemi ci hanno messo in guardia contro i potenziali pericoli di questo tipo di tecnologia. Le lezione è stata, almeno in parte, imparata: se mettiamo alla prova i sistemi di IA generativa più famosi, come il già menzionato ChatGPT per la creazione di testi o DALL-E per la creazione di immagini, con richieste ai limiti della decenza, i sistemi le respingono e non procedono con la computazione. Lontano dagli esperimenti generativi fatti con i nostri laptop, però, sistemi sempre più 'autonomi' (nell'ambiguo senso usato da Galanter) sono sviluppati per eseguire operazioni in circostanze troppo rischiose per gli esseri umani, sistemi a cui vengono collegati non monitor per la visualizzazione di testi e immagini, bensì droni per lo sgancio di bombe, oppure sistemi di difesa contraerea. Il delicato equilibrio tra controllo programmato e generatività creativa diventa ancora più critico in questi contesti,

dove il rischio non è costituito da contenuti offensivi, ma da veri e propri attacchi alla vita delle persone.

L'uso delle armi è deciso, almeno nei paesi democratici, dai politici che governano tali paesi. La loro evidente ignoranza su come funziona l'IA generativa, unita ai rischi di un'IA generativa non adeguatamente costruita, sembra essere un fortissimo monito contro l'automazione informatica delle armi. Già in passato esperti di IA si sono riuniti per firmare una lettera contro le applicazioni belliche di questa disciplina (Future of Life Institute, 2016), ma forse più che ammonire è efficace spiegare. In questo senso, razzismo e sessismo generati in ambito digitale siano i benvenuti, se possono aiutarci a capire meglio in che direzione ci stiamo muovendo.

Bibliografia

J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway, *A 3D morphable model learnt from 10,000 faces*, in «Proceedings of the IEEE conference on Computer Vision and Pattern Recognition», 2016, pp.5543-5552.

N. Bostrom, *Superintelligenza. Tendenze, pericoli, strategie,* Bollati Boringhieri, Torino, 2018.

S. Bringsjord, P. Bello, and D. Ferrucci, *Creativity, the Turing test, and the (better) Lovelace test*, in J. H. Moor (ed.) *The Turing test: The elusive standard of artificial intelligence*, Springer, Dordrecht (WN), 2003, pp.215-239.

P. Galanter, *What is Generative Art? Complexity Theory as a Context for Art Theory*, in «Proceedings of Generative Art 2003», 2003. http://www.philipgalanter.com/downloads/ga2003_paper.pdf

H.E Gardner, *Formae mentis. Saggio sulla pluralità dell'intelligenza*, Feltrinelli, Milano, 2022.

D. Jurafsky, J.H Martin, *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, *3rd edition,* Web Standford Edu Online, 2023.

https://web.stanford.edu/~jurafsky/slp3/ed3book_jan72023.pdf

A. Lecoutre, B. Negrevergne, F. Yger, *Recognizing art style automatically in painting with deep learning*, in «Asian Conference on Machine Learning, PMLR», 2017, pp.327-342.

A. Lovelace, *Translator's notes to L.F. Menabrea's memoir*, *Scientific Memoirs Selected from the Transactions of Foreign Academies of Science and Learned Societies*, 1843, vol. 3, pp.691-731.

J. McCarthy, M.L. Minsky, N. Rochester, C.E Shannon, *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955*, in «AI Magazine», *27*(4), 12, 2006.
https://doi.org/10.1609/aimag.v27i4.1904

M. Rescorla, *The Computational Theory of Mind", The Stanford Encyclopedia of Philosophy* (Fall 2020 Edition), 2020.
https://plato.stanford.edu/archives/fall2020/entries/computational-mind/

K. Wach, C.D Duong, J. Ejdys, R. Kazlauskaitė, P. Korzynski, G. Mazurek, J. Paliszkiewicz, E. Ziemba, *The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT*, in «Entrepreneurial Business and Economics Review», 11(2), 2023, pp.7-24.

H. Yu, *Reflection on whether Chat GPT should be banned by academia from the perspective of education and teaching*, in «Frontiers in Psychology», 14, 2023.
https://doi.org/10.3389/fpsyg.2023.1181712

Sitografia

BBC. 2015. "Google apologises for Photos app's racist blunder", *BBC News Tech*.
https://www.bbc.com/news/technology-33347866
Future of Life Institute. 2016. "Autonomous

Weapons Open Letter: AI & Robotics Researchers". https://futureoflife.org/open-letter/open-letter-autonomous-weapons-ai-robotics/

Gutman-Wei, R. 2018. "The 13 Strangest Moments from the Zuckerberg Hearing", *The Atlantic*. https://www.theatlantic.com/technology/archive/2018/04/the-strangest-moments-from-the-zuckerberg-testimony/557672/

Hunt, E. 2016. "Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter", *The Guardian*. https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter

McCarthy, J. 2007. "What is Artificial Intelligence". http://www-formal.stanford.edu/jmc/whatisai.html