MARIO
VERDICCHIO

LE AFFASCINANTI MA PERICOLOSE METAFORE DELL'INTELLIGENZA ARTIFICIALE

Ateneo, 27 marzo 2024

DEFINIRE L'INTELLIGENZA ARTIFICIALE

Il termine "Intelligenza Artificiale" ha sollevato diversi dibattiti da quando è stato utilizzato per la prima volta nel 1955 in una proposta per un workshop estivo al Dartmouth College (New Hampshire, USA) da coloro che sono considerati oggi i fondatori della disciplina. In quella proposta, è scritto che "lo studio deve procedere sulla base della congettura che ogni aspetto dell'apprendimento o qualsiasi altra caratteristica dell'intelligenza può in linea di principio[1] essere descritto in modo così preciso che una macchina può essere costruita per simularlo". Questa congettura sembra gettare le basi per una definizione di IA come disciplina dedicata a una descrizione precisa delle caratteristiche dell'intelligenza per farle simulare da una macchina. Esaminiamo il concetto di intelligenza utilizzato in questa proposta. C'è un presupposto secondo cui "l'apprendimento" è una caratteristica dell'intelligenza, il che non è controverso, poiché è accettato in molti altri campi oltre all'IA, come la pedagogia[2] o la psicologia[3]. Molto più controverso è l'altro presupposto sull'intelligenza, secondo cui essa è suscettibile di descrizioni precise e compatibili con le macchine. Il primo autore della proposta, il Prof. John McCarthy, ha sostenuto questo presupposto per tutta la vita, come dimostrato da un manifesto sotto forma di domande e risposte da lui pubblicato per la prima volta nel 2004 e rivisto l'ultima volta nel 2007, in cui la sua risposta alla domanda "che cos'è l'intelligenza?" è la seguente:

> L'intelligenza è la parte computazionale della capacità di raggiungere obiettivi nel mondo. Diversi tipi e gradi di intelligenza esistono nelle persone, in molti animali e in alcune macchine[4].

Queste considerazioni tracciano una linea molto netta attorno al concetto

1 John McCarthy, J., Marvin L. Minsky, Nathaniel Rochester, Claude E. Shannon, *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955*, in "AI Magazine", 27, 2006 (4), p. 12.
2 Joseph D. Novak, D. Bob Gowin, *Learning how to learn*, Cambridge 1984.
3 Jean Piaget, *The psychology of intelligence*, London 2003.
4 John McCarthy, *What is artificial intelligence*, 2007: http://www-formal.stanford.edu/jmc/whatisai.html (ultima consultazione: agosto 2024).

di "intelligenza", che è ritenuta intrinsecamente computazionale da McCarthy. In questa visione, esistono degli obiettivi nel mondo, e l'intelligenza è la parte computazionale (ovvero che consiste di operazioni su numeri) delle azioni che un agente (sia esso umano, animale, o artificiale) esegue per raggiungerli. Questa è un'affermazione così audace che il documento stesso inizia con una clausola di esclusione di responsabilità: "le opinioni espresse qui non sono tutte opinioni consensuali tra i ricercatori di IA". Le discussioni su che cosa sia l'intelligenza spaziano da resoconti aneddotici al buon senso a teorie scientifiche a pieno titolo. Una persona in grado di eseguire operazioni computazionali a una velocità notevole senza mai commettere errori sarebbe considerata intelligente se non fosse in grado di apprendere le regole più basilari per vivere in società? Tra le teorie più note sull'intelligenza che includono più del calcolo c'è la teoria delle intelligenze multiple dello psicologo Howard Gardner[5], che differenzia l'intelligenza umana in otto modalità specifiche: visuo-spaziale, linguistico-verbale, logico-matematica (presumibilmente quella su cui McCarthy ha eseguito le sue ricerche), corporeo-cinestetica, musicale, interpersonale, intrapersonale e naturalistica. Queste due visioni sull'intelligenza sembrano creare due scenari diversi, ognuno con una sua propria definizione di intelligenza e, di conseguenza, con un suo distinto ambito dell'IA come disciplina. Nel primo scenario, McCarthy ha torto o almeno ha una visione troppo ristretta dell'intelligenza, la quale, in questo caso, esiste anche in modalità che non sono computazionali. Piuttosto che Intelligenza Artificiale (IA), dovremmo chiamare la sua disciplina Intelligenza Artificiale Computazionale (IAC). Nel secondo scenario, McCarthy ha ragione e tutte le forme di intelligenza possono essere ricondotte a processi computazionali, e psicologi come Gardner propongono un quadro con forme multiple di intelligenza solo perché le loro fondamenta computazionali non sono ancora state scoperte. Chi ha ragione? Quale scenario è quello reale? Non esiste una risposta definitiva a queste domande, ma sicuramente molta energia è dedicata nel contesto della ricerca sull'intelligenza artificiale per perseguire entrambe le visioni. In particolare, molti ricercatori che credono nella definizione di intelligenza di McCarthy fanno parte di un sottocampo dell'intelligenza artificiale in cui l'obiettivo è costruire macchine che svolgano qualsiasi compito di cui un essere umano sia capace. Poiché la gamma di capacità intellettuali umane è così vasta e generale, questo sforzo di ricerca è chiamato Intelligenza Artificiale Generale (in inglese, *Artificial General Intelligence*, AGI)[6]. Resta da vedere se un giorno l'AGI avrà successo. Ciò che è interessante ora

5 Houard Gardner, *Frames of mind: The theory of multiple intelligences*, New York 1983.

6 Ben Goertzel, Cassio Pennachin (a cura di), *Artificial General Intelligence*, Berlin 2006.

nell'AGI, almeno nell'ambito di questa analisi, è la sua attenzione al confronto tra intelligenza umana e intelligenza artificiale.

CONFRONTARE L'INTELLIGENZA ARTIFICIALE

Da un punto di vista definitorio formale, un'analogia tra due entità A e B è una mappatura uno a uno tra oggetti, proprietà, relazioni e funzioni in A e quelli in B. Non tutto in A deve essere messo in corrispondenza con elementi corrispondenti in B: un'analogia è composta da corrispondenze solo tra un sottoinsieme di caratteristiche[7]. Chiaramente, l'analogia alla base dell'IA è tra un essere umano e una macchina computazionale. Tuttavia, quali aspetti debbano essere coinvolti nell'analogia è controverso e ciò porta a diverse varianti dell'IA: generale, ristretta, forte e debole. Le contrapposizioni generale/ristretta e forte/debole sono ortogonali ma non completamente indipendenti e la loro connessione ci riporta al riferimento di McCarthy alla simulazione dell'intelligenza umana con una macchina. In effetti, una simulazione imita un processo tramite un altro processo[8], ovvero una simulazione è intrinsecamente basata su analogie tra due processi, quello che viene simulato e quello che simula. Se l'intelligenza umana deve essere simulata tramite una macchina computazionale, quali aspetti devono essere riprodotti all'interno della macchina simulante? È qui che le contrapposizioni sopra menzionate mostrano la loro ortogonalità. AGI e intelligenza artificiale ristretta (in inglese Artificial Narrow Intelligence, ANI)[9], si basano sulla quantità, più precisamente, la quantità di compiti che una macchina deve essere in grado di svolgere. Nell'AGI, l'obiettivo è il più ambizioso: tutti i compiti che un essere umano può svolgere, nell'intera gamma teorizzata da Gardner, devono essere descritti in termini computazionali in modo che una macchina possa eseguirli. Nell'ANI, il contesto è invece più ristretto: una macchina è costruita per eseguire un compito specifico o un insieme molto piccolo di compiti. Se la realizzabilità dell'AGI è ancora dibattuta tra i ricercatori, ci sono stati diversi progetti ANI di grande successo in diversi campi, tra cui i giochi[10] e la medicina[11]. L'IA

7 Paul Bartha, *Analogy and Analogical Reasoning*, in Eduard N. Zalta (a cura di), *The Stanford Encyclopedia of Philosophy*, 2019, disponibile online: https://plato.stanford.edu/archives/sum2022/entries/reasoning-analogy/ (ultima consultazione: agosto 2024).
8 Stephan Hartmann, *The world as a process*, in Rainer Hegelsmann, Ulrich Mueller, Klaus G. Troitzsch (a cura di), *Modelling and simulation in the social sciences from the philosophy of science point of view*, Berlin 1996, pp. 77-100.
9 Ragnar Fjelland, *Why general artificial intelligence will not be realized*, in "Humanities and Social Sciences Communications", 7, 10, London 2020.
10 AA.VV., *Mastering atari, go, chess and shogi by planning with a learned model*, in "Nature", 588 (7839), London 2020, pp. 604-609.
11 AA.VV., *Machine learning applications in cancer prognosis and prediction*, in "Computational and structural biotechnology journal", 13, 2015, pp. 8-17.

forte e quella debole, d'altro canto, si basano sulla qualità, non in termini di esecuzioni perfette e mancanza di errori da parte della macchina, ma nel suo significato originale di "quale", di come una certa situazione viene percepita in maniera soggettiva dall'agente immerso in essa, in termini di caratteristiche qualitative. I sostenitori dell'IA forte credono che sia teoricamente possibile costruire una macchina che intrattenga esperienze coscienti come fanno gli umani, mentre i sostenitori dell'IA debole credono che ci sia una profonda differenza ontologica tra il cervello umano e le macchine informatiche, e solo i primi hanno le caratteristiche che li rendono in grado di percepire i "qualia" (plurale di quale). Gli scienziati concordano sul fatto che il sistema nervoso umano rende possibile la percezione dei qualia, ma il modo in cui quella sensazione soggettiva emerga dalla fisiologia umana rimane un mistero così profondo che è chiamato il "problema difficile" (in inglese, "hard problem") nel campo della filosofia della mente[12]. Nell'intelligenza artificiale debole l'analogia finisce qui: possiamo costruire macchine di calcolo sempre più sofisticate che svolgono sempre più compiti tradizionalmente svolti dagli esseri umani, ma la coscienza rimarrà per sempre una caratteristica sfuggente dell'esperienza umana che sfugge alla modellazione computazionale. Un'analogia è in effetti l'attacco più famoso contro l'idea di IA forte, fornito dal filosofo John Searle, che ha proposto l'esperimento mentale della "stanza cinese"[13], in cui si immaginava all'interno di una stanza, ad elaborare messaggi provenienti dall'esterno scritti in ideogrammi cinesi sulla sola base del loro aspetto grafico (dato che Searle non capisce il cinese), a formulare risposte seguendo le indicazioni scritte in un registro sotto forma di regole "se vedi XXX rispondi con YYY", e a inviare risposte che hanno senso in cinese, dando così l'impressione che la stanza capisca il cinese alle persone all'esterno. In questa analogia, Searle ha immaginato un'esperienza umana limitata, ovvero una in cui si elaborano solo i segni di cui è composto un messaggio ma non il suo significato, per darci un'idea di come funzionano le macchine informatiche: elaborano segni, ovvero numeri, in base ai loro valori e ad alcune regole, ma non hanno una mente in grado di associare idee e concetti a tali segni. Un altro filosofo, Hubert Dreyfus, usa argomenti simili per attaccare l'AGI: l'esperienza soggettiva che gli esseri umani hanno grazie alla loro coscienza non è solo necessaria per gli esseri umani per intrattenere significati, ma è anche un ingrediente fondamentale per formare ciò che è noto come buon senso. Sulla base delle loro esperienze passate, gli esseri umani sono in grado di tracciare analogie, affrontare con successo nuove

12 David Chalmers, T*he hard problem of consciousness*, in "The Blackwell companion to consciousness", Hoboken 2007, pp. 225-235.

13 John R. Searle, *Minds, brains, and programs*, in "Behavioral and brain sciences", 3, 1980, pp. 417-424.

situazioni e padroneggiare il complesso gioco della vita. Non tutti hanno successo allo stesso modo, ma tutti hanno il potenziale per apprendere qualsiasi tipo di compito compatibile con la natura umana. Questa intelligenza generale è possibile solo per gli esseri umani coscienti, mentre codificare tutte le possibili situazioni della vita reale in una macchina informatica è impossibile, semplicemente perché non c'è una memoria sufficientemente grande a contenerle tutte: gli esseri umani se la cavano col buon senso, costruendo analogie (tutt'altro che perfette, ma spesso funzionali) con loro esperienze passate[14]. Qui, qualità e quantità si incontrano: abbiamo bisogno dell'esperienza qualitativa della coscienza per sbloccare il potere di apprendere una quantità potenzialmente infinita di compiti. Anni dopo l'introduzione della stanza cinese, quando gli è stato chiesto se considerasse l'IA forte un'impossibilità logica, Searle ha lasciato la porta aperta, basandosi ancora una volta su un'analogia:

> (...) il cervello umano è una macchina, una macchina biologica, e produce coscienza tramite processi biologici. Non saremo in grado di farlo artificialmente finché non sapremo come lo fa il cervello e potremo quindi duplicare i poteri causali del cervello. (...) al momento non sappiamo abbastanza sul cervello per costruire un cervello artificiale[15].

Il suo attacco all'idea di computer coscienti può essere così inquadrato in un contesto tecnologico: nel modo in cui i computer sono costruiti oggi non è possibile l'emergere della coscienza; l'analogia tralascia alcune caratteristiche chiave. Naturalmente, questo non è un problema per coloro che perseguono gli obiettivi meno ambiziosi ma comunque potenzialmente molto impattanti di un'IA debole e ristretta. Tuttavia, concentrandosi sulle macchine elettroniche e digitali di elaborazione in uso oggigiorno si fa luce su un'altra minaccia per l'IA, forte, debole, generale o ristretta: i confini che la definiscono e distinguono come disciplina sembrano scomparire.

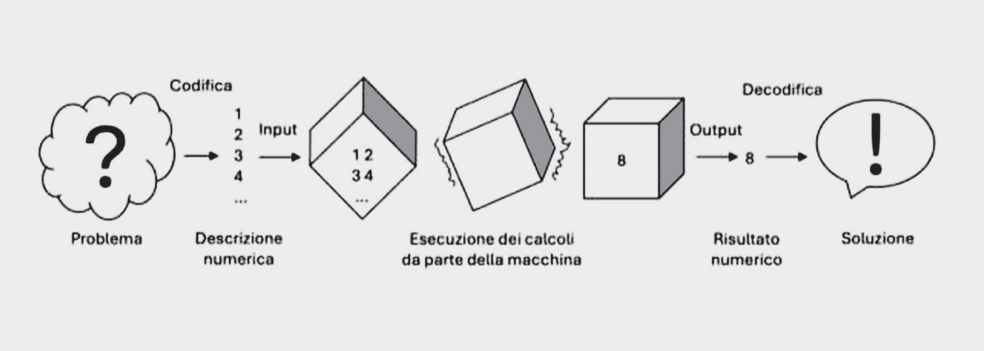
DISTINGUERE L'INTELLIGENZA ARTIFICIALE

I computer, le macchine computazionali più famose e diffuse al momento, sono un buon esempio di "realizzabilità multipla": per eseguire calcoli abbiamo diverse scelte su che tipo di dispositivi fisici costruire per rappresentare i numeri ed eseguire operazioni su di essi. Queste scelte si sono ampliate nel

14 Hubert L. Dreyfus, *What computers still can't do: A critique of artificial reason*, Cambridge 1992.
15 Dan Turello, *Brain, Mind, and Consciousness: A Conversation with Philosopher John Searle*, 2015, disponibile online: https://blogs.loc.gov/kluge/2015/03/conversation-with-john-searle/ (ultima consultazione: agosto 2024).

corso dei secoli. L'abaco, come appare oggi, fatto di legno e rinforzi metallici, è stato costruito nella Cina del XIII secolo. Per supportare le attività di suo padre come commercialista, Blaise Pascal inventò la prima calcolatrice meccanica con ingranaggi metallici rotanti, nota come Pascaline, nella Francia del XVII secolo. Charles Babbage modificò un telaio Jacquard e lo trasformò nel "Difference Engine", una macchina in grado di elevare i numeri alla seconda e terza potenza e di calcolare la soluzione di specifiche equazioni quadratiche nell'Inghilterra del XIX secolo. Una macchina di calcolo meccanica molto più sofisticata, con la rivoluzionaria caratteristica aggiuntiva della programmabilità, ovvero la capacità di archiviare non solo i dati ma anche le operazioni da eseguire su tali dati, con un conseguente notevole aumento dell'automazione, fu creata da Konrad Zuse all'inizio del XX secolo in Germania. Nello stesso periodo, con l'invenzione della valvola termoionica da parte di John Ambrose Fleming, divenne possibile controllare il flusso di elettricità attraverso componenti elettronici, e si aprì la strada ai primi computer elettronici, come quello di John Vincent Atanasoff, concepito nel 1937 e rilasciato 5 anni dopo all'Iowa State College, USA. La scoperta dei materiali semiconduttori, cioè materiali che abilitano o bloccano il flusso di elettricità a seconda della tensione con cui vengono stimolati, cambiò per sempre il mondo dell'informatica: i fisici americani John Bardeen, Walter Brattain e William Shockley inventarono il transistor nel 1947, consentendo così una miniaturizzazione senza precedenti degli interruttori che controllano il flusso di elettroni all'interno di una macchina da calcolo. La loro invenzione, che valse loro il premio Nobel per la fisica nel 1956, è il motivo per cui oggi possiamo portare in tasca computer molto potenti[16]. Nonostante l'enorme varietà tecnologica, in termini di progettazione, materiali, fenomeni fisici coinvolti, esiste un paradigma generale che guida la costruzione e l'uso di una macchina che comprende tutti i dispositivi sopra menzionati, inclusi abachi e i più recenti e veloci computer digitali. Ci sono ovviamente differenze radicali nelle prestazioni e nei livelli di automazione; tuttavia, il principio di funzionamento è lo stesso. Qualsiasi compito sia da eseguire, esso deve essere codificato, cioè descritto sotto forma di numeri; alcuni componenti della macchina sono utilizzati per rappresentare quei numeri; i numeri sono elaborati dalla macchina, cioè i suoi componenti che rappresentano i numeri sono modificati in base ad alcune operazioni sulla macchina (e dalla macchina stessa, se essa è programmabile); infine, quando le operazioni sono terminate, la macchina raggiunge uno stato finale e le quantità numeriche rappresentate in essa sono il risultato numerico che deve essere decodificato, cioè tradotto in un risultato pratico necessario per completare il compito [FIG. 1].

16 Martin Campbell-Kelly, William Aspray, Nathan Ensmenger, Jeffrey Yost, *Computer: A History of the Information Machine* (III ed.), London 2016.



[01]
RISOLUZIONE DI UN PROBLEMA PER MEZZO DI UNA MACCHINA COMPUTAZIONALE

Da questa prospettiva molto generale, usare un computer per ordinare un elenco di nomi o usarlo per simulare l'intelligenza umana non sembra fare una differenza significativa, poiché entrambe le attività si riducono allo stesso tipo di sequenza di operazioni di codifica-esecuzione-decodifica. Cosa, allora, distingue l'IA da altri rami dell'informatica, come l'ingegneria del software (in inglese Software Engineering, SE) o la teoria dei database (DB)? In effetti, può sembrare legittimo chiedersi se ha senso che l'informatica abbia dei sottocampi. La distinzione tradizionale in informatica tra hardware e software può essere un buon punto di partenza. L'hardware è fisico: è l'apparato materiale che costruiamo per eseguire calcoli. Il software è più astratto perché è la descrizione delle configurazioni da dare all'hardware per eseguire calcoli. Le caratteristiche dell'hardware determinano le sue possibili configurazioni e quindi limitano la portata del software. Ad esempio, non possiamo guardare un'immagine digitale su un abaco, perché all'abaco manca l'hardware per produrre i pixel luminosi e colorati che costituiscono un'immagine[17]. Tenendo a mente il breve e non esaustivo schizzo sulla storia delle macchine informatiche fornito in precedenza, apprezziamo almeno due direzioni lungo le quali l'hardware informatico può migliorare: può consentire nuovi tipi di operazioni (ad esempio l'elaborazione di immagini digitali) e può consentire un'esecuzione più rapida di tali operazioni. Poiché tutti gli informatici, siano essi ricercatori di IA, SE o DB, utilizzano lo stesso tipo di hardware, la distinzione tra i loro sottocampi, se esiste, deve derivare dal software, ovvero dalle operazioni che hanno scelto di far eseguire al loro hardware. Dobbiamo cercare criteri per classificare alcune operazioni come software di tipo IA per distinguerle dal software SE e dal

17 In generale, per le immagini digitali la codifica si basa su standard che creano una corrispondenza tra i numeri e i livelli di rosso, verde e blu di ogni pixel, e tra un sistema di coordinate numeriche e la posizione di ogni pixel all'interno dell'immagine.

software DB. Stiamo tornando alla questione di definire che cosa sia l'IA. Se guardiamo alle definizioni successive a quella di McCarthy, notiamo che l'attenzione verso ciò che fanno gli esseri umani è sempre presente. Secondo l'Enciclopedia Britannica:

> Il termine [intelligenza artificiale] è spesso applicato al progetto di sviluppo di sistemi dotati dei processi intellettuali caratteristici degli esseri umani[18].

Il professore di informatica Wolfgang Ertel è piuttosto critico nei confronti di questo tipo di definizione, perché non riesce a distinguere l'IA dal resto dell'informatica: dopotutto, ricordare grandi quantità di testo e calcolare numeri sono processi intellettuali eseguiti dagli esseri umani, e quindi secondo questa definizione ogni computer sarebbe un sistema di IA[19]. Ertel considera la seguente definizione di Elaine Rich di gran lunga superiore:

> "L'intelligenza artificiale è lo studio di come far fare ai computer cose in cui, al momento, le persone sono più brave"[20].

Mi unisco a Ertel nell'elogiare la definizione di Rich perché introduce molte dimensioni nel discorso sull'IA con poche parole. Innanzitutto, si riferisce a un confronto o meglio a una competizione tra umani e macchine che è stata introdotta per la prima volta da Alan Turing (un pioniere *ante litteram* dell'IA) nel suo "gioco dell'imitazione", un esperimento mentale in cui un chatbot fa credere a una persona di parlare con un essere umano, per definire un criterio per riconoscere una macchina come intelligente[21]. In secondo luogo, ma non meno importante, questa definizione inquadra l'IA come un'attività di inseguimento di un bersaglio mobile, dove il movimento non è determinato solo dallo sviluppo tecnologico dell'hardware per il calcolo, ma anche dal cambiamento concettuale che investe ciò che è considerato un'attività intellettuale intrinsecamente umana. Tale cambiamento è significativamente influenzato dagli stessi sviluppi tecnologici dell'IA. Da questo punto di vista, memorizzare testi e fare calcoli erano attività intrinsecamente umane in un'epoca in cui i computer non esistevano o erano macchine estremamente rudimentali e lente, ma non più. Ora che, in queste attività, i computer superano gli umani di miliardi di volte in

18 Brian Copeland, *Artificial Intelligence*, in "Encyclopedia Britannica", London 2022: https://www.britannica.com/technology/artificial-intelligence (ultima consultazoine: agosto 2024).
19 Wolfgang Ertel, *Introduction to Artificial Intelligence*, New York City 2017.
20 Elaine Rich, *Artificial Intelligence*, New York 1983.
21 Alan M. Turing, *Computing machinery and intelligence* in "Mind", 59, 1950, pp. 433-460.

termini di efficienza e correttezza, qualcos'altro è considerato intrinsecamente umano e l'attenzione dell'IA si è spostata di conseguenza. Quindi, possiamo distinguere l'IA dagli altri rami dell'informatica per la sua natura dinamica: sempre in prima linea nella modellazione computazionale delle attività intellettuale umane, l'IA affronta problemi ancora irrisolti, solo per risolverli e trasformare le relative soluzioni in software di informatica ordinaria e andare avanti. Cercare di capire verso dove si sta muovendo l'IA ci porta all'ultima parte di questa analisi: quella rivolta verso il futuro.

ESTRAPOLARE L'INTELLIGENZA ARTIFICIALE

Cercare di prevedere il futuro dell'IA è parte integrante dell'IA stessa: gli sforzi di ricerca sono intrinsecamente orientati al futuro sotto il segno di una modellazione computazionale sempre più completa dell'intelligenza umana. Dopotutto, l'AGI e l'IA forte sono sottocampi dedicati a macchine intelligenti che non esistono o, almeno, non esistono ancora. Il rischio qui è quello di scrivere fantascienza piuttosto che prevedere gli sviluppi futuri della ricerca sull'IA. In effetti, una quantità significativa di storie di fantascienza coinvolge entità IA che sono diventate senzienti e aiutano gli umani oppure si ribellano a loro. Ovviamente, queste storie non forniscono mai una spiegazione scientifica su come queste macchine computazionali abbiano raggiunto la caratteristica umana della piena coscienza, ma è interessante notare che gli scrittori di fantascienza hanno immaginato scoperte rivoluzionarie sia basate su software sia su hardware: nella serie TV "Humans", ad esempio, gli umanoidi diventano completamente coscienti grazie a un codice speciale che viene caricato su Internet[22] mentre nella saga di "Terminator" i robot fanno quel salto grazie a un microchip particolare[23]. Sfortunatamente, esiste una quantità non trascurabile di ricerche futuristiche sull'intelligenza artificiale che si concentra su quei risultati finali fittizi senza fornire solide giustificazioni per un tale salto. Libri come "Superintelligence"[24] entrano nei dettagli di come un computer super intelligente che diventa senziente potrebbe elaborare una strategia per dominare il mondo senza fornire alcuna indicazione su come tale superintelligenza potrebbe arrivare ad esistere in primo luogo. Un'altra narrazione in intelligenza artificiale che è indistinguibile dalla fantascienza è il concetto di "singolarità", proposto da Ray Kurzweil[25], secondo il quale il ritmo del cambiamento tecnologico aumenterà a tal punto che l'intelligenza biologica e quella delle macchine si fonderanno

22 Mark Brozel, (regista), *Humans*, episodio #2.8 [Serie TV]. Channel 4, 2016.
23 James Cameron, (regista), *Terminator 2: Judgement Day* [Film], Carolco Pictures, 1991.
24 Nick Boström, *Superintelligence: Paths, dangers, strategies*, Oxford 2014.
25 Ray Kurzweil, *The singularity is near*, in Ronald L. Sandler (a cura di), "Ethics and emerging technologies", London 2014, pp. 393-406.

nella fase successiva della coevoluzione uomo-macchina, dove la vita umana sarà trasformata in modo irreversibile. Molte di queste fantasie sono estrapolazioni dello straordinario sviluppo delle tecnologie informatiche durante il XX secolo. Uno degli esempi più famosi è la legge di Moore, che prende il nome dal co-fondatore di Intel Gordon Moore, che osservò nel 1965 che il numero di transistor in un chip raddoppiava ogni anno grazie ai miglioramenti nella tecnologia di miniaturizzazione[26].

Nonostante alcuni aggiustamenti nei decenni successivi, il modello individuato da Moore sembra essere tuttora veritiero. Cosa possiamo dedurre da questo? Ci sono almeno tre osservazioni che dovrebbero impedirci di trarre conclusioni apocalittiche o utopistiche sul futuro dell'IA. In primo luogo, ci sono limiti fisici all'hardware dati dalle leggi della fisica a cui sono soggette tutte le entità materiali. È vero che i transistor possono diventare sempre più piccoli, ma non possono essere più piccoli di un atomo. La curva del numero di transistor per chip può essere modellata come un esponenziale, ma c'è un limite[27]. Più in generale, non dobbiamo prendere un modello matematico come una rappresentazione realistica in tutte le sue parti, quindi anche se la tecnologia AI e la tecnologia digitale hanno mostrato una crescita eccezionale negli ultimi decenni, questo potrebbe non essere più vero in futuro. In secondo luogo, non dobbiamo dimenticare la dicotomia quantità/qualità: un aumento della densità di transistor porta sicuramente a macchine di calcolo più veloci, il che a sua volta significa che una maggiore quantità di operazioni computazionali può essere eseguita per unità di tempo, ma ciò non implica che determinati compiti diventeranno adatti alla simulazione della macchina. C'è una distinzione tra compiti irrealizzabili e impossibili: un compito irrealizzabile è un compito per il quale esiste una soluzione computazionale, ma essa richiede così tante risorse computazionali che non è ragionevole affrontarlo; un compito impossibile è un compito per il quale non esiste una soluzione computazionale (nota). Rompere una protezione basata sulla crittografia è attualmente irrealizzabile, ma potrebbe diventare molto più facile una volta che il calcolo quantistico, ovvero il calcolo che sfrutta i fenomeni della meccanica quantistica, diventerà disponibile grazie a una svolta tecnologica[28]. Computare la coscienza è, invece, un compito impossibile, poiché non sappiamo come la coscienza venga prodotta o emerga nel cervello, né se quel meccanismo può essere simulato tramite calcolo. Aumentare il numero di operazioni che una macchina di calcolo esegue in un'unità

26 Gordon E. Moore, *Cramming more components onto integrated circuits*, "Electronics", 38, 1965 (8).

27 Laszlo Kish, *End of Moore's law: thermal (noise) death of integration in micro and nano electronics*, in "Physics Letters A", 305, 2002 (3-4), pp.144-149.

28 Dorothy E. Denning, *Is Quantum Computing a Cybersecurity Threat? Although quantum computers currently don't have enough processing power to break encryption keys, future versions might*, in "American Scientist" 107, 2019 (2), pp. 83-86.

di tempo non cambierà questa realtà.

In terzo luogo, non dobbiamo dimenticare che i computer e i sistemi di intelligenza artificiale sono, come qualsiasi altra impresa tecnologica, un prodotto industriale, impigliato in una rete mondiale di catene di fornitura, interessi economici, strategie politiche e, in ultima analisi, persone[29]. Le macchine di calcolo potrebbero diventare sempre più sofisticate e in grado di servirci in modi che oggi sono solo nel regno della fantascienza, ma chi farà davvero parte di quel "noi"? Ci sono futurologi che immaginano un futuro in cui umani e robot coesistono, questi ultimi assumendo tutto il lavoro pesante[30] o addirittura sostituendo altri umani come perfetti partner sentimentali e sessuali[31]. A parte la solita mancanza di qualsiasi spiegazione scientifica su come tali risultati potrebbero essere ottenuti, questi autori non riescono a dirci chi finanzierà gli enormi sforzi tecnologici necessari per costruire tali macchine e chi sarà in grado di permettersi di godersele, se si può anche solo immaginare di godersi la vita in un mondo così particolare. Per evitare di sconfinare nella fantascienza, un approccio più concreto per immaginare il futuro dell'IA potrebbe essere quello di osservare il presente dell'IA che, nonostante un'apparente attenzione all'apprendimento, è profondamente diverso da ciò che McCarthy aveva immaginato nel 1955. L'IA delle origini, ora nota come GOFAI, la cara vecchia IA (in inglese, Good Old Fashioned AI)[32], era caratterizzata da un approccio top-down basato su regole che mirava a codificare la conoscenza in un computer sotto forma di assiomi e regole di inferenza che simulavano il ragionamento deduttivo negli esseri umani. Oggigiorno il paradigma dominante nell'IA è l'approccio bottom-up basato sui dati del Machine Learning (ML)[33]. In ML, i computer sono programmati per cercare modelli, schemi e leggi generali tra grandi quantità di dati mediante processi statistici induttivi. Questi processi sono implementati sotto forma di funzioni matematiche complesse i cui parametri vengono modificati in base a quanto bene i loro output soddisfano gli obiettivi per cui il sistema è stato creato in primo luogo. Questi obiettivi sono solitamente il completamento di attività di classificazione dei dati (ad esempio, di immagini mediche digitali), clustering (ossia raggruppamento, ad esempio, di spettatori di un servizio di streaming) e rilevamento di valori anomali (ad esempio, di acquisti sospetti con carta di credito).

29 Kate Crawford, Vladan Joler, *Anatomy of an AI System: An anatomical case study of the Amazon echo as an artificial intelligence system made of human labor*, 2018, disponibile online: https://anatomyof.ai/img/ai-anatomy-map.pdf (ultima consultazione, agosto 2024).

30 Aaron Bastani, *Fully automated luxury communism*, London 2019.

31 Michael Hauskeller, *Mythologies of transhumanism*, Berlin 2016.

32 John Haugeland, *Artificial Intelligence: The Very Idea*, Cambridge 1989.

33 Michael Jordan, Tom Mitchell, *Machine learning: Trends, perspectives, and prospect*, in "Science", 349, 2015 (6245), pp. 255-260.

Il ruolo dei ricercatori di IA è cambiato radicalmente in questo cambio di paradigma da GOFAI a ML: non programmano dati e operazioni in macchine di calcolo, ma forniscono dati a funzioni matematiche per plasmarle finché esse non diventano in grado di elaborare nuovi dati come prescritto dagli obiettivi preposti. In GOFAI, gli esseri umani stabiliscono le regole per raggiungere gli obiettivi, mentre in ML gli esseri umani stabiliscono solo gli obiettivi, mentre le regole vengono sviluppate automaticamente all'interno di funzioni matematiche eseguite dalle macchine di calcolo. Le operazioni in un sistema ML sono troppo complesse perché i programmatori umani possano tenerle sotto controllo. L'unico aspetto che gli esseri umani possono controllare è se il sistema ML ha raggiunto l'obiettivo o meno. Pertanto, i sistemi ML sono chiamati "scatole nere" ("black box" in inglese): gli esseri umani possono solo vedere cosa entra e cosa esce, ma non cosa succede dentro. Quando si tratta di ML, c'è una significativa diminuzione del coinvolgimento diretto dei ricercatori di IA, che ha implicazioni significative sulla responsabilità nell'IA: chi è responsabile quando un sistema ML completamente automatizzato classifica erroneamente i dati e persone subiscono danni? Finora, gli incidenti causati dai sistemi ML sono stati casi isolati di razzismo basato sulle macchine[34] e di eccessiva dipendenza dalla guida autonoma, con conseguenze mortali[35]. Tuttavia, se ML è il (prossimo) futuro dell'IA con un'adozione più diffusa in diversi contesti e campi, si potrebbe pensare che anche tali casi negativi aumenteranno.

CONCLUSIONI

L'IA mira a simulare l'intelligenza umana mediante modelli computazionali eseguiti su computer digitali. Questa impresa è piena di sfocature e imprecisioni: non abbiamo una definizione chiara di intelligenza, non siamo d'accordo su quali aspetti dell'intelligenza siano realmente suscettibili di modellazione computazionale, non abbiamo obiettivi specifici che distinguano l'IA da altre attività basate sui computer, l'IA sembra essere un bersaglio mobile che continua a cambiare ciò che consideriamo intrinsecamente umano e, infine, non abbiamo un'idea chiara su dove si stia muovendo l'IA, sebbene vi siano fondati sospetti che sempre più persone ne saranno danneggiate. Questo non vuole essere un appello antitecnologico, ma la raccomandazione di un atteggiamento critico nei confronti di una tecnologia ancora poco matura, ma su cui sono stati fatti ingenti investimenti e attorno cui, quindi, gravitano numerosi e forti interessi, non sempre orientati verso il benessere degli

34 Loren Grush, *Google engineer apologizes after Photos app tags two black people as gorillas, in "The Verge", 1,* 2015.

35 John Baruch, *Steer driverless cars towards full automation*, in "Nature", 536, 2016 (7615), pp. 127-127.

utenti finali. Un costante dialogo tra chi concepisce e costruisce l'IA, chi emana leggi per regolarne l'uso, e chi la usa e ne trae beneficio o ne subisce danno è l'ingrediente fondamentale per porre le basi di un futuro lontano da certi immaginari distopici.