dossier

Macchine più intelligenti dell'uomo?

Il cammino verso la superintellingenza artificiale

Andrea Carobene
Giornalista, Head of data and digital management di United Risk Management, <a.carobene@unitedrisk.eu>

La possibilità di creare un'intelligenza artificiale talmente sviluppata da essere in grado di superare un essere umano nei vari campi non costituisce più un'ipotesi remota, da romanzo di fantascienza. Quali passi sono stati fin qui compiuti in questa direzione? Quali potranno essere i tempi per raggiungere questo obiettivo? Quali potrebbero essere le conseguenze di una tale invenzione?

Lo sviluppo dell'intelligenza artificiale (AI) negli ultimi decenni apre lo spazio per una domanda radicale: esiste la possibilità che questa tecnologia possa arrivare a sostituire l'essere umano non solamente nel lavoro, nelle tecniche di videosorveglianza, nella ricerca scientifica, ma l'essere umano in quanto tale, ossia in quanto essere intelligente?

Il primo a porsi questa questione, da un punto di vista tecnologico, fu il matematico britannico Irving John Good nel 1965. Così scriveva: «Una macchina ultra intelligente può esser definita come una macchina che può superare di gran lunga tutte le attività di ogni uomo di qualunque intelligenza. [...] Una macchina superintelligente potrebbe progettare macchine ancora migliori; ci

sarebbe quindi sicuramente un'esplosione di intelligenza, e l'intelligenza dell'uomo verrebbe lasciata molto indietro. Così, la prima macchina superintelligente sarebbe l'ultima invenzione che l'uomo ha bisogno di realizzare, purché sia sufficientemente docile da dirci come tenerla sotto controllo» (Good 1965, nostra trad.). In questo testo incontriamo tutti i temi fondamentali relativi alla superintelligenza. Accanto alla sua definizione – un'intelligenza superiore a quella umana –, troviamo già preconizzato **il timore che sia l'ultima invenzione umana, timore legato alla capacità di questa macchina di autoinventarsi e di automigliorarsi, nonché la paura di non riuscire a tenerla sotto controllo**.

Nel 1993, quasi 30 anni dopo, un saggio di Vernor Vinge, professore del Dipartimento di Scienze matematiche dell'Università della California a San Diego, porta nuovamente alla ribalta la questione. Il suo testo, redatto per il Simposio sponsorizzato dalla NASA e dall'Istituto aerospaziale dell'Ohio "VISION-21" aveva il titolo evocativo *The Coming Technological Singularity: How to Survive in the Post-Human Era*. Alla domanda di esordio del testo, «Che cos'è la singolarità?», Vinge, autore di racconti di fantascienza oltre che docente universitario, rispondeva: «L'accelerazione del progresso tecnologico è stata la caratteristica centrale di questo secolo. Sostengo in questo articolo che siamo sulla soglia di un cambiamento paragonabile all'avvento della vita umana sulla terra. La causa precisa di questo cambiamento è nella creazione imminente da parte della tecnologia di entità con un'intelligenza maggiore rispetto a quella umana» (*ivi*, nostra trad.). L'avvento della superintelligenza è considerato quindi da Vinge una singolarità nella storia umana, esattamente come l'apparizione della vita rispetto ai fenomeni chimici della materia inorganica, o l'avvento del pensiero e della coscienza che caratterizzano l'essere umano.

Superintelligenza forte e superintelligenza debole

Oggi non esiste una definizione univoca di superintelligenza, chiamata anche ultraintelligenza. In genere, si utilizza quella data da Irving Good nel 1965, secondo cui **la superintelligenza è «la capacità di una macchina di superare tutte le abilità di un uomo, per quanto intelligente»** (Good 1965, nostra trad.). Vi è invece una discussione accesa sulla possibilità che una tale intelligenza possa essere creativa, o ancora se possa avere una consapevolezza di sé (cfr Searle 1980).

La superintelligenza è chiamata anche AGI – Artificial general intelligence (Intelligenza artificiale generale), ma alcuni autori preferiscono distinguere questi due termini riservando l'acronimo AGI

unicamente all'intelligenza artificiale che eguaglia, ma non supera, tutte le abilità umane (cfr Hassani, Sirimal Silva *et al.* 2020). Questa discussione è resa ancora più complicata dalla difficoltà di stabilire che cosa si intenda con questa formula, considerato anche – come sottolineano altri ricercatori – che oggi non esiste una definizione univoca di intelligenza umana (cfr Drexler 2019; Leggs e Hutter 2007, che elencano 70 diverse definizioni di intelligenza).

Nel seguito di questo articolo definiremo la superintelligenza come la capacità di superare l'intelligenza umana in tutti i campi, ma per avere un quadro completo **va tenuta in mente un'ulteriore distinzione**, comunemente accettata, **tra la superintelligenza tout court e la superintelligenza debole** (o *narrow*, ossia ristretta), **intesa come la capacità della macchina di superare l'abilità umana in compiti specifici**.

Il percorso verso la superintelligenza artificiale debole ha attraversato tre tappe specifiche, che hanno un valore paradigmatico. La prima è relativa alle famose sei partite con le quali nel 1997 il programma informatico Deep Blue della IBM sconfisse il campione mondiale di scacchi Garry Kasparov. Nel 2011, invece, un altro software IBM, chiamato Watson, fu utilizzato per battere Ken Jennings e Brad Rutter, due campioni del gioco televisivo statunitense *Jeopardy!*, un quiz televisivo nel quale il nozionismo ha un ruolo molto limitato. La terza tappa significativa è stata la sconfitta nel gioco Go subita dal campione Lee Se-dol nel 2016, risultato ottenuto con il computer Alpha Go di Google. Go è un gioco da scacchiera che, nonostante regole semplici, conduce a un numero di possibili posizioni delle pedine di gran lunga superiore al numero di atomi dell'universo. Questo significa che non è sufficiente la potenza di calcolo della macchina per la vittoria, ma sono necessarie abilità come intuizione e creatività. La superintelligenza debole costituisce, in prospettiva, una tappa verso la superintelligenza forte, ossia verso la realizzazione di un'entità capace di svolgere qualsiasi compito cognitivo meglio degli esseri umani.

Il percorso verso la superintelligenza

Secondo Vinge, **sono quattro le strade che potrebbero condurre all'avvento di una superintelligenza**: lo sviluppo di supercomputer "smart" e con un'elevatissima potenza di calcolo; lo sviluppo di grandi network tra computer; la realizzazione di interfacce computer/essere umano capaci di rendere gli umani che le utilizzano superumanamente intelligenti e, infine, l'utilizzo di tecniche biologiche che arrivino a selezionare progressivamente intelletti migliori fino a creare esseri umani superintelligenti. Mentre le prime due strade si riferiscono esclusivamente a tecniche di tipo infor-

matico, le ultime due coinvolgono anche il mondo della biologia, pur prevedendo sempre l'utilizzo di tecniche di AI. Anche la stessa strada della selezione genetica, infatti, comporta l'uso di tecnologie informatiche per arrivare a individuare le migliori metodologie di selezione da impiegare.

Queste quattro possibili strade sono analizzate nel dettaglio nel 2014 da Nick Bostrom, direttore dello Strategic Artificial Intelligence Research Center a Oxford, nel suo testo *Superintelligenza. Tendenze, pericoli, strategie*, considerato uno dei saggi più completi sul tema. A parere di Bostrom, che ha fondato e dirige il Future of Humanity Institute, un centro di ricerca interdisciplinare dedicato a questioni di ampio respiro che interessano tutta l'umanità, «**Le interfacce cervello-computer sembrano essere una fonte improbabile di superintelligenza**»; al contrario, «il potenziamento cognitivo biologico è chiaramente realizzabile specie se basato sulla selezione genetica». I progressi ottenibili in questo modo sarebbero tuttavia «relativamente lenti e graduali», producendo al massimo «forme relativamente deboli di superintelligenza» (Bostrom 2014, 90-91 *passim*).

Secondo Bostrom, le vie biologiche alla superintelligenza sono da considerare ausiliarie e indirette, in quanto permetterebbero la formazione di scienziati "potenziati" che a loro volta sarebbero «in grado di progredire in misura maggiore e in modo più rapido rispetto agli scienziati e agli ingegneri *au naturel*» (*ivi*, 91) nella costruzione di una superintelligenza artificiale. Queste prospettive, tuttavia, fanno sorgere molti interrogativi etici circa la liceità di selezionare biologicamente degli esseri umani.

Seguendo Bostrom, **è però probabile che la via che permetterà di arrivare a una superintelligenza o a una AI generale passerà comunque per l'utilizzo di computer superveloci o di reti informatiche**. In questo senso, l'avvento dei quantum computer, ossia della computazione quantistica con l'aumento esponenziale della capacità computazionale, potrebbe costituire un passaggio decisivo. Secondo Cem Dilmegani, fondatore della società AIMultiple, «gli algoritmi di intelligenza artificiale che operano su computer quantistici stabili hanno una possibilità di aprire alla singolarità» (2020, nostra trad.).

I **quantum computer** o "computer quantistici" lavorano utilizzando una logica diversa da quella binaria dei bit, basata unicamente sui valori 0 e 1 che si escludono mutualmente. Si basano invece sui qubit (quantum bit) che permettono attraverso il fenomeno quantistico della sovrapposizione degli stati, di tenere in memoria contemporaneamente combinazioni con differenti probabilità di 0 ed 1, aumentando esponenzialmente la potenza di calcolo disponibile

L'ultima strada per la realizzazione della superintelligenza passa per la realizzazione di reti informatiche, che potrebbero coin-

volgere cervelli potenziati, ricadendo quindi nelle ipotesi di tipo biologico, o calcolatori. Indipendentemente dalla computazione quantistica, oggi sono due le tecniche più utilizzate per aumentare la velocità di calcolo dei computer: l'aumento di potenza dei singoli processori o il loro utilizzo in parallelo. Col tempo, sono stati infatti realizzati processori sempre più performanti, adattati espressamente per l'AI, come le TPU (Tensor Processing Unit), che aumentano ulteriormente la loro potenza se sono messe in grado di lavorare in parallelo. Attraverso i cosiddetti servizi cloud offerti da diverse società informatiche, chiunque può noleggiare "istanze" parallele di TPU, creando così reti di computer che eseguono i programmi richiesti. La computazione in parallelo oggi è una possibilità aperta a tutti e queste reti, distribuite in centri di calcolo sparsi in tutto il mondo, sono uno dei luoghi dove potrebbe un giorno nascere la superintelligenza.

I tempi della superintelligenza

Non è detto che si debba aspettare la diffusione dei computer quantistici per arrivare alla singolarità predetta da Vinge, il quale affermava che l'umanità avrebbe avuto i mezzi tecnologici per creare la superintelligenza «entro 30 anni» (Vinge 1993), ponendo così la data del salto verso la singolarità al 2023. Alcuni sondaggi realizzati con persone che operano a diverso titolo in ambito informatico mostrano che il 50% degli interpellati ritiene che il salto avverrà entro il 2040, percentuale che sale al 90% se si pone come data il 2075. In altre parole, **vi è la percezione di un salto imminente, non legato necessariamente ancora a innovazioni radicali nella tecnologia** (cfr Baum, Goertzel e Goertzel 2011).

È interessante riflettere anche sui possibili luoghi di partenza della superintelligenza, che potrebbero anche non essere i grandi laboratori di un'istituzione di ricerca militare come la statunitense DARPA (Defense Advanced Research Projects Agency, Agenzia militare per progetti di ricerca avanzati), o quelli di qualche prestigiosa università, o ancora di colossi come Microsoft, Google o Amazon. **Oggi il codice per realizzare AI è a disposizione di chiunque**, perché nella maggior parte dei casi è libero e open source, **così come è a disposizione la possibilità**

Il **cloud computing** indica l'erogazione di servizi come archiviazione, elaborazione o trasmissione di dati, attraverso la Rete. Questi servizi si basano su un'architettura distribuita, cioè su risorse hardware dislocate in luoghi diversi.

Con il termine **open source** si intende il software il cui codice è disponibile e può essere liberamente modificabile e distribuibile secondo modalità specificate dalle corrispondenti licenze. Il software open source raccoglie generalmente il lavoro di collaboratori volontari. Questo termine è spesso accostato al «software libero», di cui condivide lo spirito, ma con differenze nelle rispettive licenze di utilizzo.

di affittare a costi relativamente contenuti la potenza di calcolo su cloud. In altre parole: nulla esclude che la scintilla che darà origine alla superintelligenza possa nascere per la prima volta a casa, nel garage di qualche gruppo di hacker, o nelle cantine di abili programmatori che utilizzano i computer disponibili attraverso le piattaforme cloud.

Il timore della superintelligenza

L'avvento di una superintelligenza è stato raccontato da diversi scrittori. Tra i primi, citiamo Harlan Ellison, vincitore nel 1965 del prestigioso premio Hugo per la fantascienza assegnato a il suo racconto *Non ho bocca e devo urlare*. Qui Ellison racconta di come la superintelligenza AM si imponga al genere umano. Celebre è poi il computer HAL 9000 del film *Odissea nello spazio* di Stanley Kubrick del 2001 e dell'omonimo romanzo di Arthur C. Clarke. Ancora, possiamo citare il libro *L'indice della paura* di Robert Harris del 2011, nel quale un supercomputer destinato a trarre il massimo profitto dai mercati finanziari diventa autonomo dai suoi creatori e difende la sua stessa esistenza a ogni costo.

In tutti questi esempi, il tema di fondo è la difficoltà da parte degli esseri umani di interagire con questo tipo di superintelligenza e di controllarla, ossia il timore di avere a che fare con un'intelligenza estranea, diversa da noi seppure costruita a partire dalla nostra tecnologia. Questi timori non appartengono solamente alla fantascienza, ma sono parte integrante della riflessione scientifica sulla superintelligenza. Lo stesso Vinge, nella sua presentazione sulla supersingolarità, affermava che «il suo avvento avrebbe posto fine all'era umana» (Vinge 1993). Lo stesso concetto è stato ripreso da James Barrat in un saggio del 2013, in cui parla dell'AI come dell'invenzione che porrà fine all'età dell'uomo.

Nel gennaio 2015 un gruppo di scienziati ha diffuso un appello pubblico per avviare un dibattito sull'impatto che l'AI può avere sulla società[1]. Tra i firmatari di questo documento vi sono personalità come il fisico di fama mondiale Stephen Hawking, il Nobel per fisica Frank Wilczek, Elon Musk, lo stesso Nick Bostrom, il Direttore delle ricerche di Google Peter Norvig o ancora l'italiana Francesca Rossi, global leader IBM per l'etica dell'AI. L'appello ricorda che i benefici dell'AI sono enormi e che questa tecnologia è davvero in grado di contribuire al benessere dell'umanità. «Noi non possiamo prevedere cosa si potrebbe ottenere con

[1] L'appello «Research Priorities for Robust and Beneficial Artificial Intelligence» è consultabile all'indirizzo <https://futureoflife.org/ai-open-letter/?cn-reloaded=1>.

questi strumenti» – scrivono i firmatari – tuttavia, «l'eliminazione delle malattie e della povertà non sono [traguardi] inimmaginabili». Contemporaneamente, però, si sottolinea la necessità di uno sforzo congiunto per evitare le controindicazioni insiti in questa tecnologia, riassunti nella necessità che «**i nostri sistemi di AI devono fare ciò che noi vogliamo che facciano**» (nostra trad.). È cioè necessario che le macchine dotate di AI non si diano autonomamente i propri obiettivi, ma che la loro determinazione resti sempre sotto il controllo di chi le programma e costruisce.

Il termine **reti neurali profonde** indica dei sistemi di AI composti da vari strati, ciascun dei quali costituito a sua volta da più unità di computazione, chiamate neuroni. Le informazioni vengono elaborate progressivamente da ciascuno strato, fino a produrre il risultato finale.

Va qui però ricordato che il *machine learning*, ossia la possibilità delle macchine di apprendere, è un aspetto proprio dell'AI. Se l'AI è così efficace, è proprio per la capacità delle macchine di apprendere, grazie alle reti neurali profonde, scoprendo pattern, cioè schemi e modelli, invisibili agli occhi umani, e alla loro potenza di calcolo che trae il massimo dai big data. Questa capacità di apprendere si traduce anche nell'eventualità di macchine in grado di autoprogrammarsi o di programmare altre macchine, ampliando in maniera autonoma le proprie possibilità. L'utilizzo delle reti neurali profonde introduce poi un ulteriore elemento di complessità, ossia la difficoltà di ricostruire i processi decisionali seguiti da tali reti. Oggi è nata una disciplina specifica, chiamata XAI (Explainable Artificial Intelligence), che mira a fornire spiegazioni comprensibili agli esseri umani sui risultati dei sistemi di AI. Tale disciplina si scontra con difficoltà oggettive, in quanto proprio la struttura a strati delle reti neurali e le funzioni matematiche che collegano i diversi strati ostacolano il processo di ricostruzione a posteriori di tali "ragionamenti".

Collegato a questo filone di pensiero **vi è poi il lavoro per la cosiddetta intelligenza artificiale centrata sull'essere umano** (Human-Centered Artificial Intelligence, HCAI), **ossia un'AI che eviti i pericoli dell'eccessivo controllo da parte del computer**, per cui sia sempre possibile per l'operatore umano intervenire e, nel caso, «staccare la spina» (Shneiderman 2020).

Accanto alle preoccupazioni e alle difficoltà di ordine tecnologico vi sono i **dubbi sui criteri e obiettivi che guideranno le macchine dotate di superintelligenza** (cfr Krienke 2019). Saranno assegnati dai programmatori, così come immaginava Asimov con le sue tre leggi della robotica[2], o saranno le macchine stesse a indivi-

[2] Le tre leggi della robotica enunciate da Isaac Asimov nella nota raccolta di racconti *Io, robot* (1950) affermano che: «1) Un robot non può recar danno a un

duarli, sulla base di una logica che non sarà sempre facile interpretare? E ancora, se questi criteri e obiettivi saranno in contrasto con quelli degli esseri umani, le macchine seguiranno i loro interessi o quelli dell'umanità? In altre parole, come ricorda Bostrom, dobbiamo diventare consapevoli di «che cosa vogliamo che voglia la superintelligenza» (Bostrom 2014, 314).

Una conclusione aperta

La questione della superintelligenza si lega a interrogativi di tipo antropologico, filosofico ed etico. Domandarci come sarà il nostro rapporto con una intelligenza artificiale generale, riflettere sui suoi pericoli o sulle sue possibilità, significa comunque riflettere sull'essere umano stesso, per definire quali sono le caratteristiche che vogliamo trasferire alla superintelligenza e quelle che non desideriamo che questo tipo di macchine abbiano.

A quest'ordine di problemi se ne aggiunge un altro di carattere più generale, legato alla definizione di intelligenza e di consapevolezza. Oggi l'informatica utilizza generalmente come definizione di intelligenza quella descritta dal test che Alan Turing formulò nel 1950, affermando che una macchina mostra un comportamento intelligente quando chi interagisce con essa non è in grado di distinguere se sta dialogando con un essere umano o un computer. Questa definizione non copre tutti i molteplici aspetti della intelligenza umana, ma soprattutto non risponde a **una domanda che dovremmo porci, forse tra qualche anno, ossia se la macchina che avremo di fronte avrà acquisito la capacità di pensare e se avrà consapevolezza di sé**. Un primo passaggio importante per affrontare queste questioni potrebbe essere un accordo generale su ciò che si intende per superintelligenza e sui criteri per misurarla, così pure come definire dei criteri per misurare l'eventuale grado di consapevolezza delle macchine. Da questo punto di vista è interessante il dibattito, riportato in un numero speciale del *Journal of Artificial General Intelligence*, che ha coinvolto decine di ricercatori sul tema della definizione di superintelligenza. Dibattiti di questo tipo devono essere favoriti, coinvolgendo non solo matematici e informatici, ma anche neuroscienziati, psicologi e filosofi (Monett, Lewis e Thorisson 2020).

essere umano né può permettere che, a causa del suo mancato intervento, un essere umano riceva danno. 2) Un robot deve obbedire agli ordini impartiti dagli esseri umani, purché tali ordini non vadano in contrasto alla Prima Legge. 3) Un robot deve proteggere la propria esistenza, purché la salvaguardia di essa non contrasti con la Prima o con la Seconda Legge».

La riflessione sulla superintelligenza deve così giocarsi su due piani: da una parte **un approfondimento sui rischi che questa tecnologia pone**, dall'altra **un'analisi su ciò che può significare per noi esseri umani il confronto con questo tipo di intelligenza**. Si tratta di riflessioni che oggi non possono essere eluse, se non altro perché la possibilità di tematizzarle ne dimostra l'attualità. Il possibile non coincide con il reale, ma la storia della tecnologia ha mostrato più volte che ciò che era un tempo solamente ipotizzato è diventato prima possibile e poi reale. La superintelligenza oggi non è solamente un argomento di fantascienza, ma è oggetto di convegni scientifici e rappresenta una possibilità considerata concreta dalla maggior parte di chi lavora nel campo del *machine learning*: una possibilità, quindi, che dobbiamo cominciare a discutere.

risorse

Asimov I. (1950 ed. or.), *Io, robot*, Bompiani, Milano 1963.

Barrat J. (2013 ed. or.), *La nostra invenzione finale. L'intelligenza artificiale e la fine dell'età dell'uomo*, Nutrimenti, Roma 2019.

Baum S.D. – Goertzel B. – Goertzel T.G. (2011), «How Long Until Human-Level AI? Results from an Expert Assessment», in *Technological Forecasting & Social Change*, 1, 185-195.

Bostrom N. (2014), *Superintelligenza. Tendenze, pericoli, strategie*, Bollati Boringhieri, Torino.

Clarke A.C. (1968 ed. or.), *2001: A Space Odyssey*, TEA, Milano 1988.

Dilmegani C. (2020), «995 experts opinion: AGI / singularity by 2060 [2020 update]», in <https://research.aimultiple.com/artificial-general-intelligence-singularity-timing>, 13 settembre.

Drexler K.E. (2019): *"Reframing Superintelligence: Comprehensive AI Services as General Intelligence", Technical Report #2019-1*, Future of Humanity Institute, University of Oxford, Oxford.

Ellison H. (1965), «Non ho bocca e devo urlare», in Asimov I. (ed.), *Asimov presenta i premi Hugo 1955-1975*, Editrice Nord, Milano 1978.

Ford M. (2017), *Il futuro senza lavoro*, Il Saggiatore, Milano.

Good J., (1965), «Speculations concerning the first ultraintelligent machine», in *Advanced in Computers*, 6, 31-88.

Harris R. (2011), *L'indice della paura*, Mondadori, Milano.

Hassani H. – Sirimal Silva E. *et. al*. (2020), «Artificial Intelligence (AI) or Intelligence Augmentation (IA): What Is the Future?», in *AI*, 1, 143-155.

Krienke M. (2020), «I robot distinguono tra bene e male? Aspetti etici dell'intelligenza artificiale», in *Aggiornamenti Sociali*, 4, 315-321.

Legg S. – Hutter M. (2007), «A collection of definitions of intelligence», in *Frontiers in Artificial Intelligence and Applications*, 157, 17-24.

Monett D. – Lewis C.W.P. – Thorisson K.R. (edd.) (2020), *Journal of Artificial General Intelligence Special Issue "On Defining Artificial Intelligence" – Commentaries and Author's Response*, 11, 2.

Rossi F. (2019), *Il confine dell'intelligenza artificiale. Possiamo fidarci dell'intelligenza artificiale?*, Feltrinelli, Milano.

Searle J.R. (1980), «Minds, brains, and programs», in *Behavioral and Brain Sciences*, 3, 417-457.

Shneiderman B. (2020), «Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy», in *International Journal of Human–Computer Interaction*, 6, 495-504.

Vinge V. (1993), *The Coming Technological Singularity: How to Survive in the Post-Human Era*, in <https://edoras.sdsu.edu/~vinge/misc/singularity.html>.

Sitografia

Intelligence Explosion FAQ, in <https://intelligence.org/ie-faq>.

Future of Humanity Institute, <www.fhi.ox.ac.uk>.

Future of Life Institute, <https://futureoflife.org/ai-open-letter/?cn-reloaded=1>.