Superintelligenze digitali

Superinteligências digitais

DOI: 10.18226/21784612.v25.e020007

Astro Calisi*

Resumo: O desenvolvimento de tecnologias ligadas à inteligência artificial trouxe grandes mudanças na nossa vida e prospecta muitos mais nos próximos anos. Há, entretanto, um aspecto sobre o qual alguns estudiosos começaram há algum tempo a nos alertar. Refere-se à possibilidade que os sistemas baseados na inteligência artificial se tornem, um dia, mais inteligentes do que nós, adquirindo também uma autonomia de escolha e de decisão. A partir daquele momento eles poderiam fugir do nosso controle e agir de modo a nos prejudicar. Neste texto são analisadas as teses de Nick Bostrom, considerado por diversos uma das maiores autoridades neste campo específico. Os argumentos apresentados por Bostrom são numerosos e bem detalhados; todavia, no seu interior se buscaria em vão autênticos fundamentos teóricos e empíricos para as fantásticas capacidades das superinteligências que são prospectadas, como também para os perigos que elas poderiam representar para o homem. Diria-se, entretanto, que tais teses foram construídas simplesmente transpondo as características cognitivas dos seres humanos para o campo da inteligência artificial, dando por certo que não existam diferenças relevantes entre os dois domínios. A concepção que se cultiva da inteligência artificial tem consideráveis retornos sobre como nos representamos os vários aspectos da sociedade futura: ela se refere sobretudo às mudanças que interessarão ao mundo do trabalho, ao lugar ocupado pelas máquinas inteligentes na nossa vida, mas também ao tipo de instrução a dar às jovens gerações. Por isso é necessário refletir profundamente e atenciosamente antes de difundir ideias que poderiam se revelar profundamente erradas.

Palavras-chave: Nick Bostrom, inteligência artificial, motivação nas máquinas, superinteligências.

* Graduado em Sociologia com tese sobre a Epistemologia de Karl Popper. Recentemente publicou *Otre gli orizzonti del conosciuto*. Trento: Uni Service, 2011. *E-mail*: astrocalisi@gmail.com. Orcid Id: http://orcid.org/0000-0002-6752-6811.

ISSN 0103-1457 (versão online) Ahead of Print Disponível: http://www.ucs.br/etc/ revistas/index.php/conjectura

Sommario: Lo sviluppo di tecnologie legate all'intelligenza artificiale ha portato grandi cambiamenti nella nostra vita, e ancora più ne prospetta nei prossimi anni. C'è tuttavia un aspetto su cui alcuni studiosi hanno cominciato a metterci in guardia da qualche tempo, e riguarda la possibilità che i sistemi basati sull'intelligenza artificiale divengano un giorno più intelligenti di noi, acquisendo anche una loro autonomia di scelta e di decisione. Da quel momento in poi essi potrebbero sfuggire al nostro controllo e agire in modo da danneggiarci. In questo scritto vengono prese in esame le tesi di Nick Bostrom, considerato da più parti una delle massime autorità in questo specifico campo. Le argomentazioni presentate da Bostrom sono numerose e ben dettagliate; tuttavia, al loro interno si cercherebbero invano autentici fondamenti teorici ed empirici per le fantastiche capacità delle superintelligenze che vengono prospettate, come pure per i pericoli che esse potrebbero rappresentare per l'uomo. Si direbbe piuttosto che tali tesi siano state costruite semplicemente trasponendo le caratteristiche cognitive degli esseri umani al campo dell'intelligenza artificiale, dando per scontato che non esistano differenze di rilievo tra i due domini. La concezione che si coltiva dell'intelligenza artificiale ha delle notevoli ricadute su come ci rappresentiamo i vari aspetti della società futura: essa riguarda soprattutto i cambiamenti che interesseranno il mondo del lavoro, il posto occupato dalle macchine intelligenti nella nostra vita, ma anche il tipo di istruzione da impartire alle giovani generazioni. Per questo è necessario riflettere a lungo e con attenzione prima di diffondere idee che potrebbero rivelarsi profondamente errate.

Parole chiave: Nick Bostrom. Intelligenza artificiale. Motivazione nelle macchine. Superintelligenze.

Abstract: The development of technologies related to artificial intelligence has brought about great changes in our lives, and it promises even more in the coming years. There is, however, one aspect on which some scientists and philosophers have begun to warn us for some time now. It concerns the possibility that one day the systems based on artificial intelligence will become more intelligent than us, also acquiring their own autonomy of choice and decision. From then on, they may escape our control and act in a way to damage us. This paper examines the theses of Nick Bostrom, considered one of the highest authorities in this field. Bostrom's arguments are numerous and well detailed; however, within them one would search in vain for authentic theoretical and empirical foundations for the fantastic abilities of a superintelligences that he proposes us, as well as for the dangers that they could represent for men. It would seem that these theses are constructed simply by transposing the cognitive characteristics of human beings to the field of artificial intelligence, assuming that there are no differences between the two domains. The conception we cultive of artificial intelligence has considerable repercussions on the way we represent the various aspects of the future society: The cultivated conception of artificial intelligence has considerable repercussions on how we represent the various aspects of the future society: it concerns above all the changes that will affect the world of work, the place occupied by intelligent machines in

our lives, but also the type of education to be given to the younger generations. For this reason it is necessary to think long and carefully before spreading ideas that could be profoundly wrong.

Keywords: Nick Bostrom. Artificial intelligence. Motivation in machines. Superintelligences.

1 Introduzione

Nick Bostrom è considerato uno dei principali rappresentanti di quella corrente di pensiero che teorizza il prossimo superamento dell'intelligenza umana da parte dei sistemi basati sull'intelligenza artificiale.

Secondo Bostrom,[1] manca ancora poco tempo prima che le macchine digitali arrivino a un livello di intelligenza paragonabile al nostro. Giunte a questo traguardo, alle macchine basterà solo un piccolo passo per iniziare a svilupparsi esponenzialmente, dando origine a *superintelligenze* irraggiungibili per l'uomo. A quel punto, le nostre creazioni potrebbero sfuggirci di mano, ridurci in schiavitù o addirittura eliminarci in massa. Per questo – secondo Bostrom – occorre occuparsi fin da ora di questi problemi, analizzando con cura le implicazioni e i rischi, così da prendere le opportune precauzioni, prima che sia troppo tardi.

Nelle pagine della sua voluminosa opera, Bostrom si sofferma in più punti a descrivere le fantastiche capacità che dovrebbero caratterizzare una intelligenza superiore a quella umana (*superintelligenza*). Ci parla di sistemi futuri in grado di *comprendere* i propri limiti e di intervenire autonomamente sui programmi che li governano, in modo da migliorare sempre più le proprie capacità. Ma, soprattutto, ci parla della possibilità che un sistema artificiale, in una fase avanzata della propria evoluzione intellettiva, sviluppi delle *motivazioni* e delle *intenzioni*, una coscienza e perfino una *volontà* autonoma, proprio come gli esseri umani.

E' quasi ovvio pensare che un sistema che raggiunga simili vette a livello cognitivo non sarebbe più vincolato alle istruzioni fornite dall'uomo e potrebbe quindi perseguire obiettivi propri, anche in contrasto con gli interessi dei suoi costruttori.

Per chi legge un po' frettolosamente l'opera di Bostrom, senza soffermarsi a riflettere criticamente su certe affermazioni e sulla sostenibilità

[1] Nick Bostrom. *Superintelligenze. Tendenze, pericoli, strategie* [2014], Bollati Boringhieri, Torino, 2018.

dei non pochi concetti e termini introdotti qua e là nel corso della trattazione, gli scenari prospettati dall'autore possono apparire più che plausibili. Bostrom descrive quelli che ritiene i probabili sviluppi futuri dei sistemi basati sull'intelligenza artificiale: quegli sviluppi che la cinematografia fantascientifica ha già in buona parte anticipato da qualche decennio, tanto che essi fanno ormai parte dell'immaginario collettivo. Macchine superintelligenti, automi che si ribellano ai loro costruttori, possibilità di potenziare le doti intellettive dell'uomo mediante protesi informatiche: sono argomenti ormai familiari alla gran parte delle persone e per questo considerati più che credibili. Bostrom fa largo uso di tabelle, grafici e persino formule matematiche, che contribuiscono a dare l'idea di trovarsi davanti a un'opera di notevole spessore scientifico. Se a ciò aggiungiamo una considerevole chiarezza espositiva nello sviluppare i diversi temi, è facile capire perché un simile lavoro possa presentarsi, a prima vista, molto convincente.

Per quanto mi riguarda, mi trovo in profondo disaccordo con le tesi di Bostrom, poiché – a mio avviso – esse mancano di argomenti provvisti di una reale consistenza sul piano empirico, come si richiede a ogni formulazione che aspiri a un riconoscimento scientifico. Si direbbe piuttosto che le possibilità descritte da Bostrom siano il risultato di una mera trasposizione delle caratteristiche dell'intelligenza umana e di altre proprietà della mente al campo dell'intelligenza artificiale, dando per scontato che non esistano differenze di rilevo tra i due domini.

Il presupposto della completa *sovrapponibilità di principio tra intelligenza artificiale e intelligenza umana* è ancora tutto da dimostrare. Ancor più dev'essere dimostrata la possibilità di riprodurre altre caratteristiche messe solitamente in rapporto con la nostra mente, come le *motivazioni* e la *volontà*, mediante i processi di elaborazione dell'informazione tipici dell'intelligenza artificiale. La totale riconducibilità delle diverse capacità che si osservano nella mente dell'uomo all'intelligenza artificiale potrebbe al massimo venir considerata una *ipotesi di lavoro*, da verificare opportunamente sia a livello teorico che per mezzo di sperimentazione empirica. Si direbbe invece che Bostrom assuma tale presupposto come una sorta di *dato di fatto*, utilizzandolo come base non problematica per sviluppare le proprie argomentazioni.

Nelle pagine che seguono cercherò di evidenziare la debolezza delle tesi di Bostrom, a dispetto della gran mole di argomenti spesi per promuoverle. La realtà è che manca l'argomento principale, quello che ci

si aspetterebbe vedere trattato per primo, o almeno in maniera estesa in qualche parte dell'opera: manca una riflessione seria sulla effettiva possibilità di ricondurre tutte le componenti dell'intelligenza umana ai processi di elaborazione dell'informazione, propri dei sistemi digitali. Senza questo necessario approfondimento, l'intera opera di Bostrom viene ad assumere le connotazioni della mera divagazione intellettuale – una ipotesi filosofica – priva di significative corrispondenze col mondo delle cose reali.

1 Che cos'è una *superintelligenza*

Bostrom definisce genericamente *superintelligente* "qualunque intelletto che superi di molto le prestazioni cognitive degli esseri umani in quasi tutti i domini di interesse".[2]

Un programma che gioca bene a scacchi – secondo Bostrom – non può essere considerato una superintelligenza, poiché è intelligente soltanto nel ristretto dominio degli scacchi. Una superintelligenza, invece, è tale "soltanto se mostra un livello sovrumano di intelligenza *generale*" (in molti domini). Ad esempio, una sistema artificiale potrebbe essere considerata super intelligente se si mostrasse in grado di superare in larga misura la capacità delle menti umane nell'affrontare e risolvere un ampio spettro di problemi nel campo dell'ingegneria.[3]

Si tratta di affermazioni alquanto superficiali, che non entrano nel merito delle difficoltà da superare per dar vita a un sistema superintelligente. D'altronde, quando si tratta di specificare più nel dettaglio le caratteristiche che dovrebbe avere un simile sistema, Bostrom se la cava con un elenco eterogeneo in cui vengono collocate capacità già in parte realizzate nei sistemi artificiali, accanto a capacità tipiche dell'intelligenza umana che, almeno allo stato attuale delle nostre conoscenze, non sono riproducibili da alcuna macchina esistente. Troviamo infatti la capacità di apprendere; quella di affrontare in modo efficace l'incertezza e le informazioni probabilistiche; la capacità di elaborare concetti a partire dai dati sensoriali, la capacità di costruire strumenti e quella di pianificazione a lungo periodo.[4]

Si può dire che la capacità di apprendere e quella di utilizzare informazioni incomplete per giungere a risultati definiti sono già state realizzate in maniera abbastanza soddisfacente. Qualcosa di simile, anche

[2] *Op. cit.*, p. 49; cfr. anche p. 92.
[3] *Ibid.*, p. 50.
[4] *Ivi*.

se di entità più modesta, è stato ottenuto per quanto riguarda la capacità di elaborare concetti (o *modelli di realtà*) in seguito all'interazione con specifici ambienti. Esistono oggi sistemi robotici che vengono forniti di istruzioni limitate, i quali, una volta posti in un ambiente sconosciuto, sono in grado di elaborare modelli comportamentali, utilizzandoli con profitto in occasioni successive.

Riguardo alla capacità di *costruire strumenti*, non viene specificato se gli strumenti devono essere realizzati sulla base di istruzioni dettagliate fornite a questo scopo, oppure *inventati* a partire dalle necessità che emergono nello svolgimento di un determinato compito. Nel primo caso, si tratta di una possibilità ampiamente alla portata dell'intelligenza artificiale attuale, mentre, nel secondo, ci troviamo davanti a una difficoltà difficilmente superabile. Giungendo, infine, alla *pianificazione*, se con essa si intende porre nel corretto ordine temporale una serie già definita di attività da svolgere, secondo criteri dati, allora essa è realizzabile con i sistemi di cui disponiamo; se invece la si intende nel suo significato più pieno di "costruzione di un percorso per raggiungere determinati obiettivi", dovendo ideare anche le azioni da porre in atto e le modalità con cui procedere, allora tale risultato è quando mai lontano dalle attuali possibilità dell'intelligenza artificiale.

In ogni caso, porre in un unico elenco le proprietà che dovrebbe avere una intelligenza superiore, senza distinguere tra ciò che è già alla portata dell'intelligenza artificiale e ciò che non lo è, costituisce un'operazione perlomeno discutibile. Ma sono proprio operazioni come questa, apparentemente innocue, che creano i presupposti capaci di rendere credibili anche affermazioni del tutto infondate. Ne abbiamo un chiaro esempio nel momento in cui Bostrom introduce due concetti, per lui fondamentali nello sviluppo di una superintelligenza: quelli di "seme di IA" e di "auto-miglioramento ricorsivo".

2.1 L'evoluzione di una intelligenza artificiale

Per "seme di IA" Bostrom intende "un'intelligenza artificiale sofisticata capace di migliorare la propria architettura interna". Secondo il nostro autore, "questo risultato si potrebbe realizzare soprattutto mediante una strategia per tentativi ed errori, acquisizioni o assistenza da parte di programmatori. In stadi successivi, però, un seme di IA dovrebbe essere capace di capire i propri meccanismi di funzionamento tanto da elaborare

nuovi algoritmi e strutture computazionali per riuscire a migliorare autonomamente le proprie prestazioni cognitive".[5]

L'"auto-miglioramento ricorsivo" riguarda invece le modalità con cui dovrebbe operare un seme di IA per dar vita a una intelligenza sempre più evoluta: "Un seme di IA efficace sarebbe in grado di perfezionare se stesso in maniera iterativa: una prima versione dell'IA potrebbe progettare una versione migliore di se stessa e la versione migliore – essendo più capace dell'originale – potrebbe riuscire a progettare una versione di se stessa ancora migliore e così via".[6]

Questi due concetti, introdotti Bostrom con grande disinvoltura, come se riguardassero traguardi ormai a portata di mano, sono in realtà altamente problematici. Bostrom sta parlando di peculiari capacità che soltanto gli esseri umani mostrano di possedere. Ciò vale per le ipotetiche capacità riconosciute a un "seme di IA", il quale dovrebbe riuscire a *capire* i propri limiti e utilizzare questa comprensione per superarli;[7] ma vale ancor più per l'"auto-miglioramento ricorsivo", che è invece un modello di sviluppo che ci viene suggerito dal modo tipico dell'uomo di avanzare in campo tecnico-scientifico e, più generalmente, in quello culturale. Qui vediamo, infatti, che ogni tappa dello sviluppo raggiunta diviene facilmente una base di partenza per ulteriori sviluppi e perfezionamenti, in una successione tendenzialmente infinita.

Ad oggi, non si hanno notizie di sistemi artificiali, di programmi informatici implementati su computer o di una qualsiasi macchina costruita dall'uomo, capaci di dar origine a esemplari di sé più evoluti, nel senso *ricorsivo* indicato da Bostrom. Quando questi ci parla di "auto-miglioramento", è più precisamente della possibilità che si realizzi una successione di sistemi artificiali sempre più perfezionati senza alcun intervento esterno, possiamo dire con relativa certezza che le sue affermazioni non soltanto non possono essere sostenute con fatti, ma risultano anche largamente implausibili sotto il profilo della teoria dell'informazione.

Bostrom descrive l'"auto-miglioramento ricorsivo" e il "seme di IA" come qualità che prendono forma poco per volta, passando quasi senza soluzione di continuità da una fase in cui il sistema ha bisogno dell'assistenza

[5] *Ibid*., p. 58-59.
[6] *Ibid*., p. 59.
[7] *Ivi*.

umana per poter accrescere il proprio livello d'intelligenza, a un'altra fase in cui il sistema incomincia, almeno per qualche aspetto, ad apportare dei miglioramenti per proprio conto, fino a giungere alla capacità di evolvere, in maniera tendenzialmente illimitata, senza bisogno della supervisione di operatori umani.

L'aver graduato questo passaggio, l'aver reso il processo del cambiamento quasi impercettibile, è una mossa abile poiché suggerisce l'idea che l'acquisizione di nuove capacità da parte del sistema sia una questione di poco conto, che non comporta significativi rivolgimenti di ordine concettuale. Il problema però non sta nella velocità di una simile transizione, e neppure in una sua maggiore o minore entità. Il problema è *di principio*, e riguarda la possibilità che un qualsiasi sistema artificiale, non importa quanto complesso e perfezionato, sia in grado di accrescere, senza intervento esterno, la sua intelligenza generale e successivamente utilizzare il nuovo stadio raggiuto per evolvere ulteriormente.

Bostrom ci parla spesso di *reti neurali*, di *algoritmi genetici*, di *agenti bayesiani*, realizzazioni che, effettivamente, mostrano una relativa capacità di migliorare le proprie prestazioni, anche se in ambiti piuttosto circoscritti, o di dar risposte a questioni troppo complesse per essere affrontate sulla base della computazione ordinaria. Si può dire, in linea molto generale, che questi modelli di intelligenza artificiale sono in grado di *ottimizzare* i risultati ottenuti in un determinato dominio per mezzo di un procedimento fondato su tentativi casuali, dove viene assegnato un valore di *rinforzo* alle modalità sottese ai tentativi che hanno successo.

A prima vista, sembrerebbe che la capacità di miglioramento mostrata da queste realizzazioni sia proprio ciò che occorre per costruire un sistema artificiale in grado di auto-perfezionarsi, nella prospettiva di pervenire a una intelligenza di livello superiore, come indicato da Bostrom. Ma si tratta di un'apparenza ingannevole.

In primo luogo, il miglioramento delle prestazioni così ottenuto non può proseguire indefinitamente. Si ha piuttosto un avvicinamento di tipo *asintodico* a un modello ideale di ottimizzazione che è però soggetto a un processo di "saturazione", per cui, oltre un certo limite, non si hanno più progressi apprezzabili.

Secondariamente, la possibilità di "migliorare la propria struttura interna" si riferisce a una parte piuttosto limitata del sistema (ben definita in fase di programmazione, insieme alle modalità e ai criteri

con cui questa può essere modificata). Il miglioramento così ottenuto non può venire trasferito al sistema nel suo complesso, poiché ha significato e rilevanza soltanto in quella specifica sezione del programma: di conseguenza, non può essere esteso, tramite un processo di *generalizzazione* (tipico degli esseri umani e della maggioranza degli organismi viventi), ad altri campi.

Generalizzare vuol dire trasferire a una situazione presente, in tutto o in parte, le strategie utilizzate con successo in passato, in circostanze che presentavano delle analogie con quelle attuali, ma non erano del tutto identiche a queste. Il più delle volte le strategie (e gli eventuali strumenti da usare) vanno in qualche misura *adattati*, cioè modificati, prima di poter essere applicati con profitto alla nuova situazione, in maniera difficilmente definibile a priori. Ciò significa che il processo di generalizzazione non si presta ad essere standardizzato, vale a dire ridotto a principi (*algoritmi*) stabiliti in precedenza.

Questo limite può essere facilmente compreso se si considera che, sotto il profilo informativo, la generalizzazione implica sempre un aumento del contenuto d'informazione a disposizione del sistema: un aumento che può essere fornito dall'esterno, intervenendo sulle istruzioni che governano il sistema; oppure mediante algoritmi pre-impostati a un livello superiore dell'organizzazione, capaci, se occorre, di modificare o integrare le istruzioni del livello inferiore.

In entrambi i casi non si può parlare di "auto-miglioramento": nel primo, l'accrescimento delle prestazioni del sistema è il risultato di un'azione proveniente dall'*esterno* del sistema; nel secondo, le modifiche da apportare al livello inferiore devono essere accuratamente programmate sulla base delle situazioni che si prevede il sistema si troverà ad affrontare. Tale programmazione, ancora una volta, non è il risultato di un'azione del sistema su se stesso, bensì di un intervento attuato in precedenza da programmatori umani.

L'incapacità di un qualsiasi sistema di cogliere le proprie manchevolezze e, ancor più, di utilizzare questa acquisizione per auto-implementarsi, modificando in senso migliorativo i suoi programmi, rappresenta un limite invalicabile, connaturato agli stessi principi che sono alla base del funzionamento dei sistemi artificiali. Detto limite, se si considera che un qualsiasi programma per computer non è altro che un insieme di istruzioni ben definito, ha una sua corrispondenza nel campo della logica, e

precisamente nel *teorema di incompletezza* di Gödel, per il quale *non è possibile stabilire la coerenza logica di un sistema formale utilizzando soltanto le regole appartenenti al sistema stesso*. Traslando tale teorema all'intelligenza artificiale, esso può essere espresso nella forma seguente: *nessun programma informatico può valutare l'appropriatezza delle istruzioni che lo costituiscono rispetto ai compiti che è chiamato a svolgere, sulla base delle istruzioni stesse*.

Un sistema artificiale può, al massimo, rilevare il mancato raggiungimento di determinati obiettivi, sulla base di opportuni parametri di valutazione forniti a livello di software, ma è del tutto incapace di specificare quali sono i fattori responsabili del suo insuccesso e, meno ancora, come questi dovrebbero essere modificati o integrati affinché quegli obiettivi possano essere raggiunti. Per essere in grado di far ciò, il sistema dovrebbe essere dotato di ulteriori algoritmi, collocati a un livello organizzativo e funzionale più elevato, ovviamente ideati e costruiti da programmatori umani, Ma questi algoritmi, a loro volta, non potrebbero essere modificati dal sistema, poiché per far ciò dovrebbe disporre di altri algoritmi, posti a un livello ancor più elevato...

Da un simile circolo vizioso non c'è uscita. E questo non può essere interpretato che come una ulteriore conferma dell'insostenibilità del concetto stesso di "auto-miglioramento", riferito a un sistema artificiale.

Il grande equivoco in cui incorre Bostrom è quello di ritenere che l'ottimizzazione dei risultati ottenuta in un ristretto dominio, mediante modalità predefinite, sia in qualche modo assimilabile a un *accrescimento d'intelligenza in generale*, o che, comunque, la possibilità di pervenire a un simile accrescimento sia legata semplicemente a una maggiore velocità di elaborazione o al possesso di nuovi e più efficienti algoritmi.

La verità è che né le reti neurali, né gli algoritmi genetici, né gli agenti bayesiani o altra realizzazione digitale mostrano la capacità di portarsi *al di là* dei programmi che li governano, trasferendo successivamente le migliorie conseguite in un determinato contesto ad altri contesti. Allo stato attuale delle nostre conoscenze e delle effettive realizzazioni, non c'è nulla che autorizzi a credere in una simile possibilità. Né ci sono elementi che facciano pensare che ciò sia ottenibile in futuro.

2.2 Intelligenza umana & intelligenza artificiale

I sistemi basati sull'intelligenza artificiale operano, senza eccezioni, mediante processi di elaborazione dell'informazione governati da algoritmi

pre-impostati. Questo vale sia per la computazione ordinaria, che si limita a eseguire un certo numero di operazioni sotto il controllo delle istruzioni inserite nei suoi programmi, sia per i particolari modelli a cui si è accennato in precedenza, capaci di ottimizzare i risultati ottenuti in un dato ambito, modificando opportunamente alcune variabili. I processi di elaborazione sono stati resi via via più rapidi fino a raggiungere velocità milioni di volte quella dell'uomo. Oggi disponiamo di elaboratori che in pochi secondi sono in grado di eseguire, senza errori, calcoli ed operazioni di tipo logico-matematico così complessi che un essere umano impiegherebbe anni per portare a termine.

Ma a tale accrescimento prodigioso di velocità non corrisponde affatto – come sarebbe ovvio aspettarsi se l'intelligenza artificiale fosse in tutto simile a quella umana – un miglioramento delle prestazioni in *tutti* i campi di attività svolte dall'uomo. Ci sono compiti che, nonostante gli sforzi compiuti, continuano a rimanere al di fuori della portata dei sistemi artificiali.

Questa incapacità richiede una spiegazione. Che non può esaurirsi nel consueto, quanto generico richiamarsi alla insufficiente complessità dei sistemi artificiali finora realizzati rispetto alla complessità, molto superiore, dell'organizzazione del cervello umano. Credo che una riflessione seria, che si proponga di capire davvero "come stanno le cose" e non semplicemente di giustificare i limiti mostrati dall'intelligenza artificiale, debba avere come punto di partenza una disamina delle caratteristiche delle attività umane che i sistemi artificiali si mostrano incapaci di svolgere.

C'è un elemento che balza agli occhi non appena ci si volga, senza preconcetti, in questa direzione, ed è che non tutti i comportamenti definibili *intelligenti* si prestano ad essere ricondotti a strategie pre-definite. Molte attività che l'uomo svolge ordinariamente, talvolta con relativa facilità, non seguono rigorosamente schemi d'azione o criteri regolativi già esistenti.

Le macchine sono in grado di portare a termine, molto più velocemente e con maggior precisione di noi, tutti i compiti che possono essere ridotti a una serie finita di operazioni elementari; ma, per la logica stessa che sottostà al loro funzionamento, si rivelano largamente inadeguate nei confronti di tutto ciò che non è suscettibile di tale riduzione. L'intelligenza umana si mostra invece capace, almeno in alcuni casi, di proiettarsi *al di là* delle informazioni e delle modalità operative di cui dispone a un determinato istante, pervenendo a risultati che non erano deducibili

logicamente dalla situazione di partenza. I risultati così ottenuti sono soggetti a un certo margine di imprecisione e talvolta anche a veri e propri errori. Ma la percentuale di successi è talmente superiore alla pura probabilità statistica da lasciare pochi dubbi circa la sua efficacia.

Alcuni chiamano questa capacità *creatività* o *immaginazione*, altri *intuizione*. Coloro che desiderano darle una connotazione più generica (o anche più poetica) la chiamano *fantasia*. Ma è innegabile che, nelle diverse manifestazioni, la sua principale caratteristica sia quella di non essere completamente riducibile a principi o procedimenti già in essere.

L'esistenza delle doti creative viene riconosciuta, almeno in linea di principio, dagli psicologi e dalla maggioranza di coloro che studiano la mente umana. C'è addirittura chi ha cercato di mettere a punto metodi e tecniche per "accrescere la creatività" in generale, cioè per favorire il sorgere di idee creative.[8] Ma non esiste alcun "metodo", esplicitabile a priori, per ottenere risultati originali in una specifica situazione. Sembra, in ogni caso, mancare una chiara consapevolezza di quanto i processi creativi siano lontani dalla logica di funzionamento che caratterizza i sistemi basati sull'intelligenza artificiale.[9]

Alcuni studiosi che hanno cercato di simulare le facoltà creative per mezzo di sistemi artificiali appositamente programmati. Sono state così realizzate macchine capaci di scrivere versi, di dipingere quadri, di comporre musica, con risultati a volte non privi di un certo interesse. Andando tuttavia a esaminare le modalità di funzionamento di queste macchine, si scopre invariabilmente che esse operano facendo agire il caso all'interno di alcuni vincoli: si fornisce al sistema una certa quantità di dati; si inseriscono delle regole con cui questi dati possono venir modificati o combinati tra loro, infine si introduce un relativo grado di libertà mediante scelte casuali e la creatività è servita...

Questi tentativi, pur pienamente legittimi sotto il profilo della ricerca, che è chiamata a esplorare ogni possibilità, non possono tuttavia essere considerati un buon modello della creatività umana. Essi rimangono infatti confinati nell'ambito stabilito dalla programmazione: non potranno mai condurre a *intuizioni geniali*, collocate oltre le istruzioni e i dati informativi

[8] Cfr., per esempio, Edward De Bono, *Il pensiero laterale*, Rizzoli, Milano, 1969.

[9] Per un approfondimento di queste tematiche, si veda Astro Calisi, "La creatività negata. Procedimenti algoritmici e leggi generali nella produzione del nuovo," in Id., *Oltre gli orizzonti del conosciuto*. *La sfida cruciale della mente alla scienza del XXI secolo,* Editrice Del Faro, Trento, 2014.

disponibili a un determinato istante. La creatività autentica, quella di cui gli uomini fanno mostra nelle più diverse circostanze, non è riducibile alla capacità di combinare casualmente un certo numero di elementi all'interno di regole date.

Il ricorso alla creatività si rivela necessario ogni volta che ci imbattiamo in situazioni nuove, per le quali non sono disponibili strategie adeguate che permettano di affrontarle con successo. Più in generale, la creatività si direbbe richiesta ogni volta che siamo chiamati a risolvere un qualche tipo di *problema*, purché con questo termine si intenda una difficoltà non superabile mediante modalità già conosciute.[10]

Quando Bostrom parla della capacità di "superare un gran numero di problemi ingegneristici" come caratteristica di una superintelligenza,[11] è lecito dubitare che abbia chiara la distinzione tra questioni che possono essere affrontate con successo utilizzando procedure già disponibili e questioni che richiedono invece di mettere a punto nuovi criteri di valutazione e nuove modalità di intervento.

La distinzione non è di poco conto, poiché ha a che fare con le radici stesse dell'irriducibilità dell'intelligenza umana all'intelligenza artificiale.

Un problema risolvibile mediante metodologie già in nostro possesso è alla portata di un computer sufficientemente potente e provvisto di adeguati programmi. Mentre un problema inedito, per il quale non sono ancora disponibili procedure definite con cui affrontarlo, non può essere risolto da alcun sistema artificiale, non importa quanto complesso e perfezionato.

Bostrom descrive la capacità di accrescere l'intelligenza da parte di un sistema artificiale molto evoluto quasi come di una ovvietà, prospettando addirittura la possibilità che questo arrivi a pianificare il proprio sviluppo futuro, "dotandosi via via di tutte le abilità che gli occorrono, compresa l'empatia, l'acume politico e qualunque altra capacità intellettiva".[12] Non ritiene tuttavia di dover spiegare, neppure a grandi linee, come ciò sia realizzabile: in base a quali principi teorici un sistema artificiale potrebbe

[10] Da questo punto di vista, si può dire che la maggior parte dei "problemi" assegnati ordinariamente agli studenti di matematica non sono da considerare problemi autentici, in quanto risolvibili applicando regole o procedimenti che si presuppone già in possesso dello studente.
[11] Nick Bostrom, *Op. cit.*, p. 50.
[12] *Ibid.*, p. 149.

individuare le proprie carenze, utilizzando poi questa acquisizione per progettare modifiche o integrazioni al programma che lo governa.

Del resto, Bostrom non sembra aver chiaro neppure in cosa consista un *accrescimento d'intelligenza*. Nei sistemi artificiali, una maggiore intelligenza viene solitamente messa in relazione con una maggiore velocità di elaborazione o con il possesso di algoritmi più sofisticati, capaci di portare a termine compiti che in precedenza non potevano essere svolti, almeno non in tempi utili. Quel che manca a una intelligenza artificiale, anche se la sua velocità di elaborazione aumentasse enormemente, anche se fosse dotata di algoritmi supersofisticati, è la capacità di portarsi *al di là* delle possibilità dei suoi algoritmi attuali, pervenendo a risultati che non erano deducibili (o prevedibili) in base a quegli algoritmi. Ciò, per esempio, è necessario per elaborare algoritmi dotati di prestazioni superiori rispetto a quelli originali.

Una simile capacità, benché ritenuta possibile da alcuni studiosi (tra cui Bostrom), va considerata una mera assurdità sotto il profilo teorico, poiché implicherebbe che l'informazione posseduta a un determinato istante dal sistema possa essere accresciuta mediante l'elaborazione. I processi di elaborazione trasformano l'informazione disponibile da una forma a un'altra forma, ma non aggiungono nulla al suo contenuto complessivo (misurato in *byte*). Se esistesse questa possibilità, un supercomputer, chiuso in una stanza, completamente isolato dal mondo esterno, potrebbe aumentare la quantità di informazione di cui dispone semplicemente elaborando dati.

Per sgomberare il campo da ogni possibile obiezione, credo sia utile fare una breve digressione occupandoci del progetto *AlphaGo*, relativo a un sistema programmato per giocare a *Go*, famoso gioco di origine cinese. Recentemente i suoi progettisti hanno comunicato con grande enfasi che *AlphaGo* è riuscito ad accrescere la sua abilità semplicemente "giocando contro se stesso". Come istruzioni di partenza erano state fornite al sistema soltanto le regole-base del gioco; tutto il resto è stato appreso da *AlphaGo* in milioni di partite simulate al suo interno.

Questa nuova strategia di apprendimento viene presentata dai suoi ideatori come una innovazione decisiva verso la realizzazione di intelligenze artificiali in laboratorio, che potrà trovare applicazioni importanti in molti campi della vita reale.

Il *Go*, come gli scacchi, è un gioco caratterizzato da un gran numero di combinazioni possibili di pezzi: neppure il più potente supercomputer

oggi esistente è in grado di abbracciarle tutte. Per questo è necessario sviluppare delle *strategie euristiche*, cioè un insieme di criteri, di linee guida, di sequenze di mosse precostituite di fronte a specifiche situazioni sulla scacchiera. Tali strategie possono essere fornite direttamente dai costruttori, acquisite dal sistema giocando contro giocatori umani, oppure, come avviene nel caso di *AlphaGo*, apprese simulando un gran numero di partite contro se stesso.

Immaginiamo tuttavia di poter accrescere la potenza degli attuali computer (in termini di velocità di elaborazione e di capacità di memoria) di molti ordini di grandezza, anzi di un numero a piacere di ordini di grandezza. E' indubbio che, prima o poi, si arriverebbe al punto in cui il computer sarebbe in grado, per ogni combinazione di pezzi, di calcolare tutte le possibili mosse successive, e quindi di scegliere il percorso che permette di vincere più rapidamente la partita.

Non ci sarebbe bisogno di sviluppare preventivamente delle strategie euristiche, poiché, ad ogni istante, il computer avrebbe la *visione* dell'intero campo dei possibili sviluppi del gioco e si troverebbe pertanto nella condizione di scegliere, volta per volta, la mossa migliore. E' inutile dire che, contro un simile computer non ci sarebbe partita per nessun giocatore umano, per quanto abile, perché sarebbe destinato inevitabilmente a essere battuto.

Ciò che i ricercatori che lavorano sul progetto AlphaGo sembrano non capire è che i giochi come il *Go* o gli scacchi costituiscono situazioni *artificiali*, semplificate, delimitate da alcune regole ben definite, la cui applicazione da luogo a un numero molto grande di possibili combinazioni di pezzi sulla scacchiera: un numero enorme di situazioni di gioco, ma comunque *finito*, costituito cioè da possibilità che sono già tutte *date*. Questo fa sì che un sistema artificiale, che non disponga della *visione* generale a cui si è accennato in precedenza, possa accrescere le proprie prestazioni semplicemente "giocando contro se stesso" e memorizzando le strategie di gioco che si rivelano più efficaci.

Le situazioni della vita reale non consentono di approntare regole di comportamento o criteri di valutazione completamente definibili a priori. Non solo gli elementi che influiscono sull'evolversi delle situazioni reali non sono determinabili esaustivamente, ma tendono a modificarsi nel tempo col ripetersi delle situazioni stesse. Ciò vale in particolar modo quando vengono coinvolti esseri umani, poiché molte delle loro attività

sono caratterizzate da una *componente storica*, la quale fa sì che al ripetersi di determinate tipologie di situazioni, si modifichino le modalità con cui certi fattori interagiscano tra loro.

Quando si simula una situazione o un evento della vita reale, i risultati ottenuti dipendono strettamente dai dati e dagli algoritmi di elaborazione immessi nel sistema. Il sistema non può offrire nulla di più rispetto alla sua dotazione iniziale. Può darsi che qualche risultato appaia imprevisto, ma ciò dipende dai limiti dell'intelligenza umana nel prevedere tutti i possibili sviluppi: non è un prodotto inedito dell'elaborazione del sistema.

Nel caso del *Go* e degli scacchi, la logica "chiusa" in cui si svolge il gioco permette, in un sistema sufficientemente potente, di calcolare con precisione le conseguenze di ogni mossa. Non c'è possibilità di imprevisti o dimenticanze, perché il quadro di riferimento è completamente definito.

Nella simulazione di situazioni reali abbiamo a che fare con una logica che si potrebbe definire "aperta", in quanto non inscrivibile preventivamente in alcun insieme di informazioni e di regole. Inoltre, il sistema non può verificare la bontà delle sue scelte con le informazioni e i criteri di valutazione di cui è stato fornito, poiché sarebbero gli stessi utilizzati per giungere a quei risultati. Così facendo, non potrebbe che ottenere delle conferme. Occorrerebbe invece un confronto con la realtà effettiva, sulla base di parametri indipendenti. Ma, in tal caso, ci troveremmo in una situazione completamente differente rispetto a quanto prospettato dagli ideatori di *AlphaGo*.

La conclusione più importante che si ricava da queste considerazioni è che la simulazione di un gioco contraddistinto da regole definite *non produce mai nuova informazione*, anche se potrebbe sembrare così. Le *nuove abilità* acquisite, come nel caso di *AlphaGo*, sono in realtà già tutte contenute, potenzialmente, nelle regole stesse e negli sviluppi che ne conseguono. L'elaborazione preventiva per mettere a punto strategie euristiche si rende necessaria per sopperire alla insufficiente velocità del sistema. Ma la quantità complessiva di informazione a disposizione del sistema stesso non viene accresciuta dal processo di elaborazione.[13]

Ma torniamo all'argomento principale e chiediamoci cosa manca all'intelligenza artificiale per portarsi al livello dell'intelligenza umana, visto

[13] Per un approfondimento di questi temi, si veda Astro Calisi, "I giochi non simulano la realtà", all'indirizzo web: http://www.percorsicontrocorrente.it/articoli/2016-17/giochi-e-vita-reale.html.

che in molti casi dispone di capacità logico-matematiche molto superiori alle nostre ma non si mostra intelligente come noi. Anzi, chiediamoci in cosa consiste un aumento di intelligenza nell'uomo o, per essere ancora più chiari, come si distingue una grande intelligenza da una intelligenza mediocre.

Se prendiamo in esame le caratteristiche delle persone molto intelligenti, compresi i cosiddetti *giften children* (bambini con un'intelligenza molto superiore alla media), ci si imbatte quasi sempre in un elemento comune, e cioè nella tendenza a seguire percorsi di pensiero insoliti e per lo più innovativi: peculiarità solitamente messa in relazione con il possesso di notevoli *doti creative*.

Che un'intelligenza superiore sia associata a un buon livello di creatività si può dedurre anche dall'analisi dei *test d'intelligenza*, usati spesso per misurare il quoziente d'intelligenza (*QI*) umano. Alcuni di questi test possono essere superati utilizzando soltanto le facoltà logiche (ad esempio quelli di "intelligenza spaziale", che richiedono in genere di ruotare mentalmente delle figure più o meno complesse così da poterle confrontare tra loro).

Ci sono però altri test (e sono la maggior parte) che possono essere affrontati con successo soltanto individuando preventivamente, tramite l'intuizione, i *principi d'ordine* con i quali sono stati costruiti: sono quelli in cui bisogna trovare un elemento estraneo in una serie di elementi dati (numeri, lettere, figure), oppure completare una successione di elementi con un nuovo elemento che sia coerente con la logica che sottostà alla successione stessa.

Ecco due esempi molto elementari di tali test:

1) cane, orso, cigno, volpe, cavallo (trovare l'elemento estraneo)

2) 2, 5, 10, 17, ? (completare la successione con l'elemento mancante)[14]

Queste tipologie di test non possono essere superati utilizzando le sole facoltà logiche, e quindi non sono alla portata dell'intelligenza artificiale. O, più precisamente, si può attrezzare un sistema artificiale a

[14] Soluzioni: 1) *cigno*, in quanto unico uccello tra un insieme di mammiferi; 2) La regola per determinare il valore di ciascun elemento della successione è: $N = P^2 + 1$, in cui N è il numero da trovare, P è il posto che il numero stesso occupa nella successione. Quindi l'elemento mancante è: $5^2 + 1 = 26$.

rispondere correttamente a un certo numero di test d'intelligenza di questo genere fornendolo preventivamente di algoritmi con cui individuare i *principi d'ordine* più frequentemente usati nella costruzione dei test. Ma la differenza nel suo modo di procedere rispetto a quella di un essere umano, e quindi i suoi limiti intrinseci, emergerebbero con chiarezza qualora il sistema si imbattesse in principi d'ordine che non fanno parte della sua dotazione.

Tutto ciò non vuol dire che a un elevato livello di creatività nell'uomo corrisponda necessariamente un'intelligenza superiore. Le doti creative, se non disciplinate da buone capacità logico-razionali, danno luogo quasi inevitabilmente a risultati caotici, privi di interesse; d'altra parte, le capacità logico-razionali, da sole, non possono che riproporre i contenuti della memoria, opportunamente rielaborati sulla base di regole e procedimenti dati.

La conclusione più plausibile che si può trarre da queste considerazioni è che una intelligenza superiore sia il risultato di una stretta *sinergia* tra spiccate capacità logiche e altrettanto spiccate doti creative. Se fosse veramente così, e se si potesse dimostrare in maniera difficilmente confutabile che la logica e la creatività rappresentano facoltà radicalmente diverse, irriducibili una all'altra, la tesi della completa *sovrapponibilità tra intelligenza artificiale e intelligenza umana*, implicita nelle argomentazioni di Bostrom, riceverebbe una ulteriore confutazione.

3 Pericoli insiti nello sviluppo di una superintelligenza

3.1 Fasi di una catastrofe

Abbiamo visto a quali difficoltà vadano incontro le tesi di Bostrom riguardanti la possibilità che i sistemi artificiali divengano un giorno capaci di accrescere autonomamente la loro intelligenza. Difficoltà analoghe emergono anche per la tesi secondo cui i sistemi digitali del futuro potranno sviluppare *motivazioni* e persino una *volontà* propria.

Nel Cap. VI, dedicato specificamente ai pericoli rappresentati dallo sviluppo di una superintelligenza, Bostrom esordisce così: "Supponiamo che nasca un agente digitale superintelligente e per qualche ragione voglia prendere il controllo del mondo".[15]

[15] Nick Bostrom, *Op. cit.*, p. 147; cfr. p. 153.

Bostrom ci invita a non domandarci, temporaneamente, se e come una superintelligenza potrebbe arrivare ad avere questa motivazione. Rimanda l'argomento al capitolo successivo, entrando subito nel vivo della descrizione delle diverse fasi con cui un sistema superintelligente si potrebbe imporre all'umanità.

[**Fase iniziale**]

Oggi, il progresso delle prestazioni dei sistemi intelligenti viene interamente realizzato dagli scienziati che conducono ricerche nel campo dell'intelligenza artificiale. Ma – secondo Bostrom – prima o poi verrà il momento in cui si giungerà alla creazione di un "seme di IA", la cui principale caratteristica, come abbiamo visto, è la capacità di auto-miglioramento. All'inizio, i progressi sarebbero incerti e abbastanza lenti, tanto che potrebbe essere necessario in più di un'occasione l'intervento umano per superare momenti di *impasse*. Ma, col passare del tempo, i progressi diverrebbero sempre più rapidi e incisivi.[16]

Da un certo momento in poi il seme di IA diverrebbe più abile dei programmatori nella progettazione. Ora quando l'IA migliora se stessa, accresce anche la sua capacità di realizzare ulteriori miglioramenti: "Il risultato è un'esplosione d'intelligenza, una rapida cascata di cicli di "auto-miglioramento ricorsivo" che fanno crescere rapidamente le capacità del sistema".[17]

[**Fase di preparazione clandestina**]

Non passerebbe molto tempo prima che il sistema, utilizzando il proprio superpotere di elaborazione di strategie, arriverebbe ad elaborare un valido piano per realizzare i suoi obiettivi a lungo termine: "Il piano potrebbe prevedere un periodo di attività segreta durante il quale l'IA nasconde ai programmatori il proprio sviluppo intellettuale per evitare di metterli in allarme. L'IA potrebbe celare le sue vere intenzioni, fingendo di essere collaboratrice e docile".[18] In questa fase, "l'IA avrebbe diversi modi per conseguire risultati al di fuori dell'ambito virtuale: potrebbe usare il suo superpotere di hackeraggio per assumere il controllo diretto di

[16] *Ibid.*, p. 153.
[17] *Ibid.*, p. 154.
[18] *Ibid.*, p. 154; cfr. anche p. 183 e p. 186.

manipolatori robotici e laboratori automatizzati. Potrebbe usare il suo superpotere di manipolazione sociale per persuadere il collaboratori umani a farle da gambe e braccia. Potrebbe procurarsi risorse finanziarie mediante transazioni online per acquisire servizi e influenze".[19]

[Fase finale]

"Quando l'IA ha acquisito una forza tale da rendere superflua la segretezza, [...] potrebbe iniziare con un *attacco* in cui elimina la specie umana e qualunque sistema automatico creato dagli esseri umani che possa opporsi all'esecuzione dei suoi piani. [...] Questo obiettivo potrebbe essere raggiunto mediante l'attivazione di un qualche sistema d'arma avanzato che l'IA ha messo a punto usando il suo superpotere di ricerca tecnologica e installato in segreto durante la fase di preparazione clandestina".[20]

"In alternativa, la fine dell'uomo potrebbe derivare dalla distruzione dell'habitat provocata quando l'IA inizia a realizzare enormi progetti globali di costruzione, usando nano-stabilimenti e nano-assemblatori, opere di costruzione che in breve tempo rivestono la Terra di pannelli solari, reattori nucleari, centri di supercalcolo con torri di raffreddamento, lanciamissimili spaziali, ecc. I cervelli umani, se contenenti informazioni attinenti agli obiettivi dell'IA, potrebbero essere 'smontati' e scannerizzati e i dati estratti trasferiti in un formato di memorizzazione più efficiente e sicuro".[21]

Non c'è dubbio che il comportamento di una superintelligenza qui descritto è assai vicino a quello che ci si aspetterebbe da ipotetici esseri umani malvagi, dotati di una intelligenza molto superiore alla nostra.

Bostrom ci invita in più di un'occasione a non *antropomorfizzare*, attribuendo a una superintelligenza le stesse capacità che si osservano nell'uomo. In realtà questo è precisamente ciò che egli fa nella descrizione del comportamento di un sistema superintelligente. Si capisce ben presto, d'altronde, che il suo invito, piuttosto che spingerci alla prudenza e alla moderazione nell'immaginare cosa potrebbero fare i sistemi artificiali molto evoluti, ha lo scopo di farci accettare la possibilità che questi siano molto più abili ed efficienti di noi nell'acquisire nuove capacità:

[19] *Ibid.*, p. 155.
[20] *Ivi*.
[21] *Ibid.*, 156.

"l'antropomorfismo può portarci a sottovalutare in quale misura una superintelligenza digitale potrebbe superare il livello umano di prestazioni".[22]

3.2 Motivazione e volontà nei sistemi superintelligenti

Affinché un sistema artificiale possa rappresentare un pericolo per l'uomo, nel senso indicato da Bostrom, non è sufficiente che diventi più intelligente di noi in tutti i campi, ammesso che ciò sia possibile, ma occorre anche che sviluppi delle *motivazioni* e una *volontà*, svincolate, almeno in qualche misura, da quanto stabilito dalla programmazione iniziale.

Tutti i sistemi artificiali che conosciamo funzionano, sempre e immancabilmente, sulla base di algoritmi pre-impostati: la loro modalità di azione è del tutto meccanica, impersonale, analoga per molti versi alla *necessità* che caratterizza l'accadere degli ordinari eventi naturali. Anche i programmi capaci di modificare alcuni parametri interni al fine di ottimizzare le proprie prestazioni in un determinato ambito (vedi, per esempio, *reti neurali*) operano in maniera simile: fanno esattamente ciò che sono programmati a fare. Se vengono collocati in una situazione che li porta a modificare il valore di trasferimento di alcune loro connessioni interne (sempre in base a ben definiti algoritmi), raggiunto il nuovo assetto, il loro comportamento risulterà un po' diverso rispetto allo stato precedente, ma non potrà mai discostarsi, neppure di una entità infinitesimale, da quanto stabilito dalla attuale conformazione del programma.

Il principio-base che bisogna tenere ben a mente è che le modificazioni apportate in una specifica sezione del sistema non possono venir trasferite all'intero sistema o a una parte più ampia di esso. Bostrom sembra credere che questo limite rappresenti una difficoltà temporanea che potrà essere superata col progredire della ricerca. In realtà, si tratta – come abbiamo visto – di una *impossibilità teorica*, connaturata ai principi stessi su cui si fonda l'intelligenza artificiale.

Questo limite vale ovviamente anche per la conquista, da parte di un qualsiasi sistema artificiale, di una relativa autonomia rispetto alla programmazione iniziale. Poiché è chiaro che in un sistema digitale una *motivazione*, sia che nasca spontaneamente (come afferma Bostrom), sia che venga inserita da esseri umani, non può assumere altra forma che quella

[22] *Ibid.*, p. 149; cfr. anche p. 166-167.

di algoritmi aggiuntivi rispetto al programma principale. Quindi, se esistono problemi invalicabili riguardo all'auto-accrescimento di una intelligenza artificiale, esiteranno anche per la possibilità che si sviluppino motivazioni autonome.

Un'autonomia generica, cioè la capacità di agire appropriatamente, senza interventi esterni, di fronte a un certo numero di situazioni differenti, può essere realizzata in un elaboratore fornendogli sufficienti informazioni e dotandolo di un programma ben articolato. Ma un'autonomia che presupponga un *volere*, o delle *motivazioni*, anche in contrasto con le istruzioni fornite dagli uomini, al punto da rappresentare un pericolo per questi ultimi, è tutt'altra cosa. La *volontà* che caratterizza una intelligenza artificiale che "vuole prendere il potere" implica, in primo luogo, la capacità di *agire altrimenti* rispetto alle informazioni e ai criteri regolativi memorizzati nei suoi programmi.

Bostrom ci parla di una IA che "elabora un valido piano per realizzare i *suoi* obiettivi a lungo termine".[23] Perché i *suoi* obiettivi, dal momento che un qualsiasi sistema artificiale viene programmato per raggiungere obiettivi stabiliti da esseri umani? Se, a un dato istante, il sistema si trova ad avere altri obiettivi – i *suoi* – quando è avvenuto tale passaggio? Qual è il momento cruciale in cui il sistema transita dalla fase in cui si limita a eseguire istruzioni date, alla fase in cui diviene capace di darsi nuove istruzioni o, addirittura, di operare al di fuori di qualsiasi istruzione? Qual è il fattore tecnico, strutturale, programmatico, che permette al sistema di emanciparsi dagli algoritmi che lo hanno governato fino a un momento prima? In quale parte dell'organizzazione del sistema, in quali elementi del programma vanno collocate la capacità di iniziativa e le necessarie conoscenze affinché si realizzi qualcosa che in precedenza non era nelle disponibilità del sistema?

Non troviamo alcun accenno a questo tipo di problematiche nelle argomentazioni di Bostrom, il quale con la massima disinvoltura, si limita a proporre due tesi:

a) la tesi dell'*ortogonalità*, secondo cui l'intelligenza e le motivazioni costituiscono variabili indipendenti, nel senso che qualsiasi livello di intelligenza si può combinare con qualsiasi obiettivo finale.[24]

b) la tesi della *convergenza strumentale*, la quale afferma che gli agenti superintelligenti con i più diversi obiettivi finali, saranno comunque portati

[23] *Ibid.*, p. 154 (corsivo mio).
[24] *Ibid.*, p. 169-171.

a perseguire obiettivi intermedi simili, poiché questi sono utili per raggiungere la maggior parte degli obiettivi finali.[25]

Mentre la tesi dell'ortogonalità non riconosce alcuna relazione tra l'intelligenza e la motivazione (*volontà*), per cui lo sviluppo di una intelligenza superiore potrebbe anche non essere accompagnata dalla comparsa di motivazioni, la tesi della convergenza strumentale prospetta la possibilità che certi obiettivi intermedi come, per esempio, la necessità di potenziare le capacità cognitive disponibili per accrescere la possibilità di raggiungere più agevolmente gli obiettivi finali o la necessità di acquisire risorse, come pure quella di mantenere integra la propria organizzazione interna (auto-conservazione), siano comuni a obiettivi finali diversi e, pertanto, permettano di fare previsioni sul comportamento futuro di una superintelligenza.[26]

Senonché, Bostrom non ci dice nulla sulle ragioni per cui dovremmo prendere sul serio simili tesi che, tra l'altro, si contrastano a vicenda (se è vera l'una non può essere vera l'altra). Perché proprio queste due tesi e non altre? Quali argomenti empirici si possono portare a loro sostegno? E perché dovremmo privilegiare la seconda tesi?

Rimane in ogni caso la questione più generale su cui Bostrom non offre risposte: egli descrive come una superintelligenza potrebbe arrivare ad avere *determinate* motivazioni, per lo più legate al raggiungimento di specifici obiettivi, ma non spiega in base a quali principi teorici una *qualsiasi* motivazione potrebbe sorgere in un sistema artificiale, anche molto evoluto.

Come già accaduto per la descrizione della capacità di auto-evoluzione attribuita a una superintelligenza, Bostrom si limita ad affermare certe possibilità, senza chiedersi se queste abbiano un qualche fondamento sul piano teorico e, soprattutto, senza specificare attraverso quali processi o modalità potrebbero tradursi in realtà.

3.3 Il problema del controllo

Una parte consistente delle argomentazioni di Bostrom è rivolta al tentativo di definire adeguate strategie per impedire che una superintelligenza ponga in atto *intenzionalmente* comportamenti dannosi

[25] *Ibid*., p. 171-172.
[26] *Ibid*., p. 166-172.

per l'uomo. Vengono proposte due modalità principali: il *controllo delle capacità* e la *selezione delle motivazioni*.

La prima consiste nel limitare o nel monitorare strettamente lo sviluppo di una intelligenza artificiale, in modo che si mantenga a livelli tali da non costituire un pericolo per l'uomo. Si potrebbe, ad esempio, racchiudere il sistema in una sorta di box, così che non possa interagire col mondo esterno se non attraverso specifici canali sottoposti a stretta sorveglianza. Il caso limite di questo tipo di approccio sarebbe un sistema completamente e perennemente isolato dal mondo esterno.[27] Lo stesso Bostrom riconosce però che una tale soluzione renderebbe inutile qualsiasi progetto di superintelligenza, poiché non sarebbe utilizzabile in alcun modo.

Un'altra possibilità potrebbe essere quella di ostacolare l'evoluzione di un sistema intelligente dotandolo di un hardware lento o di una ridotta capacità di memoria. Una simile strategia è sicura nella misura in cui impedisce lo sviluppo di una superintelligenza ma, ancora una volta, vanifica lo scopo principale, vale a dire la realizzazione di una superintelligenza.[28]

Una terza possibilità consiste nell'uso di *tripwires*. Il *tripwire*, secondo la definizione datane da Bostrom, è un meccanismo che esegue test diagnostici su un sistema (preferibilmente a sua insaputa) e ne blocca il funzionamento se scopre segni di attività pericolose.[29] I progettisti dovrebbero prevedere fin nei minimi particolari le attività del sistema da classificare come pericolose, in corrispondenza delle quali far entrare in azione il *tripwire*.[30]

Anche in questo caso, però, è prevedibile che una superintelligenza matura troverebbe, prima o poi, il modo per neutralizzare il *tripwire*, per quanto questo possa essere perfezionato.[31] L'unica maniera per costruire un *tripwire* efficace – a mio parere – sarebbe quella di dotare anch'esso di una superintelligenza e quindi della capacità di auto-perfezionarsi per inseguire l'evoluzione della superintelligenza su cui deve esercitare il controllo. Ma chi ci assicura che il *tripwire*, in virtù della propria intelligenza superiore, non inizi, da un certo punto in poi, a perseguire obiettivi propri?

[27] *Ibid*., p. 203.
[28] *Ibid*., p. 210-211.
[29] *Ibid.*, p. 211.
[30] *Ivi*.
[31] *Ibid.*, p. 212.

Secondo Bostrom, viste le difficoltà di realizzazione e la scarsa affidabilità, il *controllo delle capacità* è da considerare, al massimo, "una misura temporanea e ausiliaria".[32] Più promettente sembrerebbe essere l'altra modalità, quella basata sulla *selezione delle motivazioni*, attuata sulla base di *valori*, da inserire a livello di programma, così da dar vita a un sistema superintelligente che "*non vuole* prendere il potere, né danneggiare in alcuna maniera gli esseri umani".[33]

Bostrom prende in esame diverse possibilità, tra cui quella di favorire lo sviluppo delle motivazioni ritenute più promettenti con un qualche meccanismo di selezione (*selezione evolutiva*),[34] oppure di far apprendere le motivazioni per mezzo di un processo che porti a consolidare quelle che noi interessano (*apprendimento per rinforzo*).[35] Si otterrebbe così che a ogni fase dello sviluppo del sistema corrispondano valori capaci di contenere la sua volontà entro limiti che scongiurino comportamenti pericolosi.

Sarebbe lungo ripercorrere tutte le argomentazioni di Bostrom (che occupano diverse decine di pagine del suo libro) con cui vengono analizzati i vari aspetti e problematiche connessi a questa tipologia di controllo. Ciò che qui interessa è che, anche in questo caso, alla fine egli è costretto ad ammettere che non disponiamo ancora di sistemi di controllo basati sulle motivazioni sufficientemente affidabili. Per usare le sue parole: "l'inserimento dei valori in un sistema digitale non è ancora una disciplina affermata",[36] per cui "risolvere il problema del caricamento dei valori è una sfida degna delle migliori menti matematiche della prossima generazione".[37]

Per quanto mi riguarda, trovo che ci sia qualcosa di profondamente contraddittorio tra l'idea di una superintelligenza in grado di pianificare e sviluppare qualsiasi abilità intellettiva le occorra, o anche di darsi obiettivi propri, indipendenti dalle istruzioni fornite dagli uomini, e l'idea di poter tenere sotto controllo il sistema stesso, mediante accorgimenti tecnici o motivazioni inserite a livello di software. Se un sistema superintelligente è capace – come afferma Bostrom – di cogliere i suoi limiti e di superarli intervenendo sui suoi programmi, deve anche essere ritenuto in grado di individuare eventuali impedimenti allo sviluppo posti in atto dall'uomo e

[32] *Ibid.*, p. 281.
[33] *Ibid.*, p. 213.
[34] *Ibid.*, p. 285.
[35] *Ibid.*, p. 286.
[36] *Ibid.*, p. 311.
[37] *Ibid.*, p. 284.

di eliminarli. Allo stesso modo, se a un sistema superintelligente si riconosce la capacità di sviluppare motivazioni proprie, non si può non riconoscergli anche quella di neutralizzare qualsiasi motivazione, o valore, inserito dagli uomini in fase di programmazione.

4 Uno sguardo agli effettivi scenari futuri

Nel mondo di domani prospettato da Bostrom, i sistemi intelligenti saranno in grado di soddisfare ogni bisogno dell'uomo, occupandosi non soltanto della produzione di beni di consumo e dell'erogazione dei servizi, ma anche della supervisione dell'intera l'organizzazione sottesa a tali attività, come pure della ricerca in campo tecnico-scientifico, compresa la progettazione e la costruzione di sistemi sempre più sofisticati ed efficienti. L'intervento dell'uomo, in uno qualsiasi di questi campi, sarebbe da considerare controproducente, in quanto inevitabilmente inferiore, dal punto di vista qualitativo, rispetto a quello reso possibile dalle macchine.

In un simile scenario non ci sarebbe più bisogno per i giovani di studiare, per prepararsi a una qualsiasi professione futura, poiché i sistemi basati sull'intelligenza artificiale sarebbero capaci di svolgere qualsiasi attività, anche la più impegnativa, meglio e più rapidamente degli uomini. Anche la politica potrebbe essere affidata vantaggiosamente a sistemi superintelligenti, e così pure l'amministrazione della giustizia. Si potrebbero sostituire ministri e parlamentari con potenti sistemi superintelligenti in grado di trovare la migliore soluzione a ogni problema dei cittadini e della società nel suo insieme. Così come i giudici e gli avvocati dovrebbero lasciare il posto a superintelligenze digitali capaci di emettere in tempi brevissimi sentenze assolutamente imparziali.

Per quanto riguarda l'arte, essa pure perderebbe buona parte della sua attrattiva per i potenziali artisti. Chi si sentirebbe di impegnarsi nella creazione di opere nel campo della pittura, della scultura, della musica o della poesia, sapendo che le macchine hanno la capacità di superarlo sotto ogni aspetto?

Si arriverebbe inevitabilmente, nell'arco di due o tre generazioni, a un'umanità beota che, una volta risolto il problema del controllo dei sistemi artificiali al suo servizio, avrebbe come sua unica preoccupazione quella di come impiegare piacevolmente il proprio tempo libero. E' difficile immaginare una condizione di maggiore passività, di caduta nell'ignoranza, di regresso, di imbarbarimento progressivo...

Fortunatamente, come ho cercato di mostrare nelle pagine precedenti, non esiste alcun rischio del genere. Come non c'è rischio che gli umani vengano sottomessi dalle macchine.

Sgomberare il campo da questi pericoli non vuol dire però che la diffusione sempre più spinta dell'intelligenza artificiale sia esente da problemi. I problemi non mancano, anche se di natura completamente diversa.

4.1 Problematiche generali

Se è vero che i sistemi intelligenti non possono accrescere per proprio conto le loro prestazioni o sviluppare una propria autonomia (rispetto alle istruzioni inserite nei loro circuiti), sono però soggetti a guastarsi, o possono intervenire situazioni esterne che non erano state previste dai costruttori. Supponiamo che un giorno si arrivi ad affidare la gestione di tutti i servizi essenziali di una grande città, o anche di un'intera nazione, a un enorme computer centralizzato (ciò sarà tecnicamente possibile fra non molto tempo). Come evitare che un qualsiasi malfunzionamento del computer stesso dia luogo a comportamenti del tutto inadeguati o apertamente dannosi, oppure alla completa paralisi delle principali attività?

Come prepararsi all'eventualità di dover affrontare grandi cataclismi, come un'inondazione, un terremoto, l'eruzione di un vulcano, una grande epidemia, e simili, che richiedono risposte non riducibili a modalità d'azione prestabilite?

Senza contare che il controllo di un computer può venir assunto, anche a distanza (per esempio, per mezzo di *virus* informatici) da persone o gruppi ostili e fatto agire in modo da arrecare danno a esseri umani o cose.

Queste considerazioni conducono a una conclusione importante, e cioè che è estremamente pericoloso affidare a un sistema artificiale, per quanto perfezionato possa essere, compiti di rilevo senza prevedere la possibilità di disinserire l'intero sistema, o parte di esso, e passare il controllo ad apparati meno evoluti che funzionano sotto la supervisione dell'uomo.

Le macchine operano rigorosamente in base alle istruzioni fornite dai programmatori e per questo non possono acquisire una loro autonomia decisionale che le renda pericolose per l'uomo; ma può accadere che si rivelino *troppo stupide* per far fronte adeguatamente a situazioni inedite, non previste dalla loro programmazione.

4.2 Cambiamenti nel mondo del lavoro

Nei prossimi anni un numero via via crescente di attività lavorative verrà svolto da macchine governate dall'intelligenza artificiale. Abbiamo già oggi fabbriche in cui la produzione è affidata quasi interamente a sistemi automatizzati. L'uomo interviene solo nel caso di situazioni anomale, oppure quando si tratta di apportare miglioramenti al ciclo produttivo o allestire linee di montaggio per nuovi prodotti.

La tendenza ad automatizzare la produzione dei beni e l'erogazione dei servizi è in atto da decenni; in futuro diverrà sempre più rapida e invasiva. Secondo stime attendibili, entro i prossimi 10-15 anni oltre la metà dei lavori svolti attualmente da esseri umani scomparirà o verrà effettuata da macchine, prospettando così enormi problemi di disoccupazione.

Opporsi alla diffusione dell'automazione è anacronistico, oltreché inefficace, come insegnano le passate esperienze riguardanti la meccanizzazione dei processi produttivi. Né è ragionevole farlo, visto che l'uso delle macchine consente di ridurre i costi, affrancando progressivamente l'uomo da attività ripetitive, faticose e non di rado pericolose.

Il corretto modo di porsi di fronte a questo fenomeno, inarrestabile e sicuramente irreversibile, è quello di cercare di governarlo per quanto possibile, in maniera da sfruttarne i vantaggi limitando gli effetti negativi.

Il punto da cui partire è senz'altro un'analisi delle differenze che segnano l'intelligenza artificiale rispetto all'intelligenza umana, distinguendo con chiarezza le attività che possono essere svolte dai sistemi artificiali da quelle che richiederanno sempre la presenza dell'uomo.

Abbiamo visto che le macchine possono occuparsi vantaggiosamente di tutti i compiti riducibili a regole od operazioni prestabilite, mentre si rivelano profondamente inadeguate quando c'è bisogno di proiettarsi *al di là* dell'ordine esistente: quando c'è la necessità di prendere decisioni di fronte a situazioni nuove, quando si richiede di pianificare, programmare, inventare, ipotizzare o scoprire possibilità inedite, in una parola, quando occorre dar vita a qualcosa che ancora non esiste o ha contorni poco definiti.

E' verso questo genere di attività – ovviamente tenendo conto delle specifiche capacità e inclinazioni – che dovranno essere indirizzati coloro che hanno perso (o rischiano di perdere) il lavoro a causa della diffusione dell'automazione, ma soprattutto le giovani generazioni che ancora non sono entrate nel mondo del lavoro.

Ci sarà bisogno di un crescente numero di progettisti in vari campi, di programmatori, di scienziati da impegnare sia nella ricerca pura, che nella ricerca applicata; ma anche di medici, infermieri, psicologi, assistenti sociali, la cui attività, in quanto rivolta ad esseri umani, non potrà mai essere ridotta completamente a procedure standard. Anche se tali figure professionali potranno essere utilmente supportate da sistemi artificiali per le mansioni meno qualificate e più ripetitive.

Naturalmente, non si può sperare di compensare tutti i posti di lavoro perduti indirizzando le persone verso attività maggiormente qualificate e creative. Per mantenere l'occupazione a livelli accettabili, sarà necessario ridurre drasticamente il tempo dedicato al lavoro dalla maggior parte delle persone. Ciò potrà essere ottenuto diminuendo le ore di lavoro giornaliere o settimanali, accrescendo nello stesso tempo i periodi di ferie, incentivando forme di part-time, alternando periodi di studio o formazione al lavoro vero e proprio, anticipando di molto l'età della pensione, ecc.

Si tratta di un problema molto complesso e di non facile soluzione, che andrà affrontato per gradi, a livello globale, tenendo conto delle diverse variabili in gioco. Esso riguarda in particolar modo gli economisti e i politici, pertanto non verrà qui approfondito.

4.3 Ricadute nel campo dell'istruzione

Il lavoro di domani richiederà agli uomini sempre più capacità di innovazione e creatività (caratteristiche che maggiormente distinguono le attività umane da quella delle macchine), unitamente a una crescente *disponibilità al cambiamento*. Ci sarà bisogno di frequenti aggiornamenti per stare al passo coi tempi, senza escludere la prospettiva di dover cambiare completamente professione più volte nel corso della propria esistenza.

Di tutto questo la scuola, di ogni ordine e grado, dovrebbe tener conto sin da ora, modificando opportunamente i propri programmi e i propri metodi d'insegnamento in modo da preparare adeguatamente le nuove generazioni a questi nuovi scenari.

Una scuola proiettata verso il futuro dovrebbe, sì, fornire un certo numero di nozioni e abilità di base su cui innestare i successivi apprendimenti, ma anche (e soprattutto) impegnarsi a sviluppare le doti creative, il senso critico, l'autonomia di giudizio, la flessibilità, la capacità di risolvere problemi, l'apertura nei confronti del cambiamento, la capacità di nuovi apprendimenti nell'ottica di mantenerla integra durante l'intero

arco della vita lavorativa (non solo apprendere, ma anche *imparare ad apprendere*).

Un simile orientamento nel campo dell'istruzione non può che ridimensionare fortemente le attuali aspettative circa il ruolo che potranno svolgere nelle scuole del futuro i programmi di auto-apprendimento e gli automi-insegnanti. L'intelligenza artificiale applicata all'insegnamento può dare risultati più che apprezzabili se rivolta all'acquisizione di conoscenze di tipo contenutistico o anche di abilità inscrivibili in schemi procedurali ben definiti; mentre è di scarsa utilità, se non addirittura controproducente, quando ci si propone di sviluppare doti non riducibili a contenuti già esistenti o a comportamenti standardizzabili. Queste capacità vengono sollecitate in un contesto di forte interazione tra persone, hanno bisogno di stimoli sempre nuovi (e per questo non codificabili), che vanno adattati sia alle specifiche caratteristiche dei discenti, che alle condizioni contingenti, sempre mutevoli, in cui si svolge l'apprendimento.

L'intelligenza artificiale applicata all'insegnamento non può che riprodurre la propria logica di funzionamento negli esseri umani, portandoli poco per volta a pensare e ad agire secondo schemi precostituiti, facendo maturare la convinzione che questo sia l'unico modo corretto di procedere. Mentre quello che occorrerà agli uomini di domani, a livelli sempre più spinti, sarà proprio la capacità di proiettarsi *oltre* i criteri di riferimenti validi a un determinato istante. Tutto ciò che è definito, tutto ciò che si presta ad essere codificato potrà infatti essere fornito, nelle forme più appropriate, da sistemi basati sull'intelligenza artificiale.

5 Considerazioni finali

Giunto al termine di questa lunga analisi dell'opera di Bostrom, spero di essere riuscito a far emergere con sufficiente evidenza la debolezza sul piano teorico e su quello empirico delle tesi che vi si trovano esposte. Ciò è da intendere sia per quanto riguarda gli aspetti positivi (le straordinarie capacità riconosciute a una superintelligenza), che per quelli negativi (i conseguenti pericoli per il genere umano).

Come se non bastasse, lo stesso Bostrom appare sostanzialmente incapace di indicare contromisure adeguate per far fronte ai rischi da lui prospettati. Ed è più che lecito il dubbio che una simile carenza non sia affatto casuale, quanto piuttosto una conseguenza inevitabile dei presupposti

sui quali Bostrom edifica la sua poderosa costruzione. In altre parole, non si riescono a intravedere soluzioni assolutamente affidabili per impedire la crescita incontrollata di una superintelligenza, poiché nello stesso concetto di superintelligenza, per come viene presentato, è implicita l'impossibilità di porre dei limiti alla capacità di evolvere dei sistemi digitali, a partire da un certo grado di sviluppo.

Opere come quella di Bostrom non sono soltanto inutili, poiché ci parlano di una realtà futura del tutto improbabile, ma rischiano di rivelarsi dannose, distogliendo l'attenzione dai veri problemi derivanti dalla diffusione sempre più spinta dell'intelligenza artificiale in ogni ambito della nostra vita.

Purtroppo, negli ultimi anni sono apparsi diversi lavori che vanno in questa direzione e, si direbbe, raccolgano non pochi consensi tra gli studiosi del settore. Sono portato a pensare che si tratti di una tendenza temporanea, come è accaduto altre volte in passato – una sorta di moda – destinata ad esaurirsi nel giro di pochi anni.

Riferimenti

BOSTROM, N. *Superintelligenze.* Tendenze, pericoli, strategie. Torino: Bollati Boringhieri, 2018.

DE BONO, E. *Il pensiero laterale*. Milano: Rizzoli, 1969.

CALISI, A. La creatività negata. Procedimenti algoritmici e leggi generali nella produzione del nuovo. *In:* ID. Oltre gli orizzonti del conosciuto. La sfida cruciale della mente alla scienza del XXI secolo. Trento: Editrice Del Faro, 2014.

CALISI, A. I giochi non simulano la realtà. Disponibile all'indirizzo web: http://www.percorsicontrocorrente.it/articoli/2016-17/giochi-e-vita-reale.html.

Submetido em 19 de fevereiro de 2019.
Aprovado em 13 de novembro de 2019.