Gli automi e l'intelligenza artificiale tra filosofia e scienza[1]

Marco Salucci

Società Filosofica Italiana

La questione degli automi affascina l'uomo da tempi immemorabili. Per cominciare, appunto, dall'inizio, ricordo che già Omero raccontava il mito del dio-fabbro Efesto il quale aveva costruito mantici in grado di funzionare da soli, tripodi in grado di muoversi da soli per portare cibi agli dei, cani da guardia di bronzo viventi ed eternamente giovani, ancelle di aureo metallo che aiutavano il dio nel suo passo claudicante: "due ancelle si affaticavano a sostenere il loro signore, auree, simili a fanciulle vive; avevano mente nel petto e avevano voce e forza" [1, XVIII vv. 417-421]. Gli automi sono artefatti che destano meraviglia per la loro capacità di agire da soli: l'ambiguità dell'etimologia della parola in questo caso è benvenuta: "automa" può derivare da *taumazein*, cioè "meravigliare", e da *automaton* cioè "che agisce da solo".

La storia della letteratura è costellata di personaggi fantastici che rappresentano esseri non umani, sia naturali che artificiali, in grado di comportarsi come esseri umani. Per citarne alcuni tra i più famosi e in ordine sparso: il *Golem* (il gigante di terra della tradizione ebraica), l'*homunculus* (l'uomo creato *in vitro* dagli alchimisti), la shelleyana creatura di Frankenstein, la fanciulla meccanica dei racconti Hoffmann musicata da Hoffenbach. Lo stesso termine "robot" ha un'origine letteraria (nella letteratura teatrale cecoslovacca degli anni '20, ad opera dei fratelli Karel e Josef Capek). Infine le contemporanee officine della mitologia, il cinema e la fantascienza, ci hanno reso familiari molte specie di robot, replicanti, zombie, scienziati pazzi, alieni, mostri. In realtà niente di radicalmente nuovo rispetto agli esseri non umani immaginati dalle mitologie antiche. Ma ciò che è più importante è che non è neppure sostanzialmente cambiato lo stesso problema fondamentale che sta dietro a tali creature durante l'intera storia del pensiero umano: già Omero ce lo suggerisce quando afferma che le auree ancelle avevano "mente nel petto": è realmente possibile per gli automi avere una mente? In altre parole: è possibile replicare artificialmente l'intelligenza? Va da sé che l'aspetto esteriore che un dispositivo dotato di intelligenza artificiale dovrebbe avere è trascurabile: il fatto che assomiglino a cani di bronzo, a fanciulle d'oro, a giganti di fango o ai robot come quelli di *Guerre stellari* è di secondaria importanza. Ciò che conta è se sia possibile che all'interno di tali "gusci" possa abitare una mente. Dunque la forma più genuina di porre il

1 Lezione tenuta il 27 novembre 2012 presso il Liceo Ginnasio "Galileo Galilei" di Firenze.

problema in questione è quella di domandarsi se possa esistere una mente artificiale ovvero se ciò che attualmente si avvicina più di ogni altro artefatto a una mente, i computer, possa avere una mente.

C'è un'intera corrente di pensiero (pluridisciplinare poiché comprende filosofi, scienziati cognitivi, psicologi, informatici ecc.) la quale ritiene che, sì, è possibile che i computer possano avere una mente per la buona ragione che la mente è una sorta di computer. Ovviamente qui uso il termine "computer" per semplicità intendendo d'ora in poi non l'hardware ma il software. Quando dunque si dice che la mente umana è un computer si intende dire che gli stati mentali dell'uomo sono come programmi per computer.

H. Putnam, uno dei padri fondatori della contemporanea filosofia della mente è stato colui che più di ogni altro ha utilizzato situazioni fantascientifiche per discutere serissime questioni filosofiche: non si può sperare niente di meglio delle sue parole per entrare in argomento.

> Chiunque abbia passato molte ore (ben spese o mal spese?) della propria fanciullezza a leggere storie di missili e di robot, di androidi e di telepati, di civiltà galattiche e di macchine del tempo, sa benissimo che i robot […] possono essere buoni o tremendi, amici fedeli dell'uomo o suoi acerrimi nemici. I robot possono essere esemplari o pateticamente buffi: incuterci sbigottita ammirazione coi loro poteri sovrumani […] oppure divertirci con il loro comportamento ingenuo e sempliciotto […]. Almeno nella letteratura fantascientifica, dunque, un robot può essere "cosciente", il che significa […] avere sensazioni, pensieri, atteggiamenti, tratti caratteriali. Ma è veramente possibile? E se lo è, quali sono le condizioni necessarie e sufficienti? E perché noi filosofi dovremmo comunque occuparcene? […] La mia speranza [è] che il problema della mente delle macchine si dimostrerà in grado di offrire, almeno per qualche tempo, un approccio nuovo e stimolante a questioni decisamente tradizionali di filosofia della mente. [2, pp. 416 e 425]

La forma che il problema in questione ha assunto nel pensiero moderno e contemporaneo è in sostanza dovuta al filosofo e matematico francese René Descartes (1596-1650) il quale si domandava che cosa gli avrebbe potuto garantire che - parafrasando le sue parole - i cappelli e i mantelli che vedeva passare giù per strada dalla sua finestra coprivano realmente degli uomini veri e non degli automi? Domanda retorica perché Descartes in realtà riteneva che esistessero due modi sicuri per distinguere gli automi da un uomo: "Il primo è che non potrebbero mai valersi di parole o di altri segni, componendoli come noi facciamo per esprimere agli altri i nostri pensieri", "Il secondo mezzo è che, anche se facessero alcune cose ugualmente bene e anzi meglio di noi, essi inevitabilmente sbaglierebbero in alcune altre" [3, pp. 40-41]. In altre parole Descartes confidava nella capacità caratteristica della mente umana di essere creativa e in quella opposta della ripetitività del comportamento degli automi. Tale modo di distinguere intelligenza artificiale da intelligenza naturale è però entrato in crisi nel Novecento quando ha cominciato ad essere possibile rendere flessibile il comportamento di mac-

chine governate da regole. In realtà, a pensarci bene, la creatività non è incompatibile con il seguire regole: un giocatore di scacchi "crea" la sua partita obbedendo a regole che non ammettono deroghe. Uno degli studiosi che maggiormente ha contribuito allo sviluppo della nozione di creatività governata da regole è stato il linguista americano N. Chomsky che, nel 1966, così ha scritto:

> Benché [già ai tempi di Descartes] si fosse compreso che i processi linguistici sono in un certo senso "creativi", gli strumenti tecnici per esprimere un sistema di [regole in grado di descrivere adeguatamente la creatività] non erano disponibili fino a tempi molto recenti: Infatti una comprensione piena di come una lingua può [...] "fare un uso infinito di mezzi finiti" si è sviluppata soltanto negli ultimi trent'anni, nel corso di studi sui fondamenti della matematica" [4, p. 48]

Oppure, secondo le parole di un altro studioso: "la peculiarità inconfondibile del lavoro in Intelligenza Artificiale dipende dal fatto che si cerca di stabilire [...] insiemi di regole che dicano a macchine non flessibili come essere flessibili" [5, p. 28].

Fra coloro che hanno reso disponibili gli strumenti matematici a cui allude Chomsky per "fare un uso infinito di mezzi finiti" si deve annoverare Alan Turing (1912-1954). Turing, del quale ricorre quest'anno il centenario della nascita, è stato uno dei geni del Novecento. Matematico inglese, aveva collaborato con i servizi segreti britannici durante la seconda guerra mondiale alla decifrazione dei codici di comunicazione tedeschi. A lui dobbiamo la nascita della tecnologia e della teoria dei computer e dell'intelligenza artificiale. Uno dei primi computer mai costruiti, il *Colossus*, si basava sulle idee di Turing e fu decisivo nella decifrazione dei codici segreti nazisti, tanto che è stato definito il "computer che vinse la guerra".

A partire da una riflessione sui concetti di ricorsività e di commutabilità, Turing giunse a risultati che costituiscono il vero e proprio atto di nascita dell'informatica contemporanea e che sono sintetizzabili nella cosiddetta *macchina di Turing*. Una macchina di Turing è una macchina astratta, cioè fisicamente non realizzata: in realtà è un concetto, un concetto attraverso il quale è definibile ciò che è calcolabile. Una macchina di Turing è costituita da un nastro diviso in caselle che scorre attraverso un organo (una testina) di lettura/scrittura. Su ogni casella la testina può scrivere un simbolo di un alfabeto dato. La testina legge una casella alla volta e può leggere, cancellare o scrivere un simbolo sulla casella in esame e muoversi a destra o a sinistra di una casella.

Una macchina di Turing è interamente descritta dal suo registro o tavola di macchina. Il registro o tavola di macchina è una matrice (cioè una tabella con righe e colonne) nella quale vengono specificate le coppie di output e stato interno per ogni coppia di input e stato interno. Detto ciò, già dovrebbe essere abbastanza chiaro che un programma per computer è una macchina di Turing. Ma quello che adesso dobbiamo notare è che una macchina di Turing è interamente descritta dal suo registro e che in tale descrizione non compaiono riferimenti agli elementi fisici con cui, eventualmente, potrebbe essere realizzata:

> la "descrizione logica" di una macchina di Turing non include alcuna specificazione della natura fisica degli "stati"; e anzi nemmeno della natura fisica della macchina (consisterà di relé elettronici, di cartone, di impiegati umani seduti a tavolino, o di che altro?). In altre parole, una data macchina di Turing è una macchina astratta, fisicamente realizzabile in un numero pressoché infinito di modi diversi [6, p. 401].

Dunque: "la materia di cui siamo fatti non impone restrizioni alla nostra forma intellettuale" [5, p. 330]. Ciò significa che esseri completamente diversi dal punto di vista fisico possono condividere gli stessi stati mentali e pertanto è del tutto sensato dire che, ad esempio, un animale, un marziano, un robot o un computer "credono che stia per piovere", purché tale credenza possa essere descritta per tutti gli esseri in questione dalle stesse triplette di input, stati interni e output: cioè possa essere descritta dalla stessa macchina di Turing. Definiti gli stati mentali come funzioni di input e output indipendenti dal materiale in cui sono realizzati, non solo diventa possibile che anche i computer abbiano stati mentali, ma i computer hanno costituito il modello al quale i filosofi della mente e gli scienziati cognitivi si sono ispirati per descrivere l'attività mentale degli esseri umani. Il motto di tale punto di vista è: "la mente sta al cervello come il software sta all'hardware".

Se gli stati mentali stanno al cervello nella stessa relazione in cui gli stati di una macchina di Turing stanno alla sua realizzazione fisica, allora

> potrebbe trattarsi di un sistema di leve e ingranaggi come un vecchio calcolatore meccanico; di un sistema idraulico attraverso cui scorre acqua; di una rete di transistor stampati in un chip di silicio attraverso cui passa corrente elettrica; o addirittura di un cervello. Ognuna di queste macchine utilizza un proprio mezzo peculiare per rappresentare i simboli [su cui opera]: le posizioni degli ingranaggi, la presenza o assenza di acqua, il livello di tensione elettrica, e, forse, gli impulsi nervosi [7, p. 39].

Ora, se si concorda sull'impossibilità di distinguere la mente naturale dalla mente artificiale a partire dalla realizzazione fisica e dalla capacità o meno di agire in modo creativo allora non resta che trarre una conclusione: "il solo modo per cui si potrebbe esser sicuri che una macchina pensa è quello di *essere* la macchina e di sentire se stessi pensare" [8, p. 169], il che è ovviamente impossibile. Non solo ma "secondo questa opinione la sola via per sapere che *un uomo* pensa è quella di essere quell'uomo in particolare […] Invece di discutere in continuazione su questo punto, è normale attenersi alla convenzione – suggerita dalla buona creanza – che ognuno pensi" [*ibidem*]. Ma perché non dovremmo attenerci a tale convenzione anche nei confronti degli automi? E infatti Turing percorre l'unica via *praticabile* per affrontare la questione della mente delle macchine attraverso il "gioco dell'imitazione", successivamente noto anche come "test di Turing".

> Mi propongo di considerare la domanda: "possono pensare le macchine?" [...] La nuova forma del problema può essere descritta nei termini di un gioco, che

> chiameremo il "gioco dell'imitazione". Questo viene giocato da tre persone, un uomo (A), una donna (B) e l'interrogante (C), che può essere dell'uno o dell'altro sesso. L'interrogante viene chiuso in una stanza, separato dagli altri due. Scopo del gioco per l'interrogante è quello di determinare quale delle altre due persone sia l'uomo e quale la donna. Egli le conosce con le etichette X e Y, e alla fine del gioco darà la soluzione "X è A e Y è B" o la soluzione " X è B e Y è A". L'interrogante può far domande di questo tipo ad A e B: "vuol dirmi X, per favore, la lunghezza dei [suoi] capelli?" Ora supponiamo che X sia in effetti A, quindi A deve rispondere. Scopo di A nel gioco è quello di ingannare C e far sì che fornisca una identificazione errata [...] Le risposte, in modo che il tono della voce non possa aiutare l'interrogante, dovrebbero essere scritte, o, meglio ancora, battute a macchina. [...]
>
> Poniamo ora la domanda: "che cosa accadrà se una macchina prenderà il posto di A nel gioco?" L'interrogante darà una risposta errata altrettanto spesso di quando il gioco viene giocato tra un uomo e una donna? Queste domande sostituiscono quella originale: "possono pensare le macchine?". Il metodo delle domande e risposte sembra essere adatto per introdurre nell'esame quasi ogni campo della conoscenza umana che desideriamo. Non desideriamo penalizzare la macchina per la sua incapacità di brillare in un concorso di bellezza, né penalizzare un uomo perché perde una corsa contro un aeroplano. Le condizioni del nostro gioco rendono irrilevanti queste incapacità [8, p. 116]

Un celebre tentativo di mettere in crisi l'approccio di Turing è stato fatto dal filosofo americano J. Searle. Egli ha rovesciato il test di Turing cercando di fare proprio quello che Turing considerava impossibile e cioè "*essere* la macchina e sentire se stessi pensare" (vedi *supra*). In altre parole per sapere se i computer pensano dobbiamo metterci noi stessi nei panni del computer. L'argomento di Searle è noto come "argomento della stanza cinese" poiché Searle immagina che un uomo sia chiuso in una stanza e riceva dall'esterno dei fogli sui quali sono stampati degli ideogrammi cinesi [9, pp. 48-49]. Consultando un manuale nel quale sono date le regole con le quali, tenendo conto soltanto della forma degli ideogrammi, è possibile far corrispondere certi ideogrammi a certi altri ideogrammi, l'uomo nella stanza deve trascrivere gli ideogrammi corrispondenti a quelli ricevuti e trasmetterli all'esterno. L'uomo nella stanza non conosce il cinese, ma, nonostante ciò, risponde sensatamente a domande che gli vengono poste su una storia cinese.

L'uomo nella stanza, argomenta Searle, non fa niente di più di ciò che farebbe un computer programmato per eseguire lo stesso compito, e cioè manipolare simboli in base alla loro forma. Ma l'uomo nella stanza *sa* di non capire il cinese, dunque se l'uomo non capisce la storia in cinese, allora non la capirà neppure il computer. Rispetto ai sostenitori della teoria computazionale della mente, Searle opera un rovesciamento di prospettiva: non è la macchina che simula gli stati mentali di un essere umano, ma un essere umano che simula gli stati della macchina. Quindi anche se non sappiamo ciò che accade nella macchina, sappiamo che, se la macchina manipola simboli, non comprende il significato dei simboli.

Il dibattito suscitato dall'argomento della stanza cinese è molto ampio ed ha trovato sostenitori e avversari e dunque, implicitamente, avversari e sostenitori della posizione di Turing. Nessuno dei due partiti ha avuto, al momento, partita vinta. Ma la discussione ha messo in evidenza che uno dei compiti della ricerca futura è quello di chiarire se le proprietà mentali derivano da proprietà logiche (come quelle dei programmi) o da proprietà fisiche caratteristiche del cervello. E non è da escludere neppure una terza possibilità, del tutto alternativa o che rappresenti una sintesi di quelle attualmente in competizione.

BIBLIOGRAFIA

[1] Omero, *Iliade*, Padova, Marsilio 2003, trad. M.G. Ciani.

[2] Putnam H., *I robot: macchine o vita creata artificialmente?*, in Mente, linguaggio e realtà, Adelphi, Milano 1987, pp. 416-438.

[3] Descartes R., *Discorso sul metodo*, Laterza, Bari 1978.

[4] Chomsky N., *Linguistica cartesiana. Un capitolo di storia del pensiero razionalistico*, *Saggi Linguistici*, vol. III, Boringhieri, Torino 1969.

[5] Hofstadter D.R., *Gödel, Escher, Bach: un'eterna ghirlanda brillante*, Milano, Adelphi

[6] Putnam H., *Mente, linguaggio e realtà*, Milano, Adelphi 1987.

[7] Johnson-Laird P.N., *La mente e il computer. Introduzione alla scienza cognitiva*, Bologna, Il Mulino 1990

[8] Turing A., *Macchine calcolatrici e intelligenza*, (1950), in Somenzi V. (a cura di), *La filosofia degli automi*, Torino, Boringhieri 1965.

[9] Searle J., *Menti, cervelli e programmi. Un dibattito sull'intelligenza artificiale*, Padova, CLUP-CLUED 1984.