Alma Mater Studiorum · Università di Bologna

Campus di Cesena
Scuola di Ingegneria e Architettura

Corso di Laurea in Ingegneria Elettronica, Informatica e Telecomunicazioni

Intelligenza artificiale: test di Turing e alcune questioni filosofiche

Tesi di Laurea in Fondamenti di Informatica B

Relatore:
Chiar.mo Prof.
ANDREA ROLI

Presentata da:
MICHELE BRACCINI

II Sessione
Anno Accademico 2012/2013

Alla mia famiglia e alla mia ragazza.

Indice

Introduzione **1**

1 Il test di Turing **3**
1.1 Obiezioni al test di Turing 4
1.1.1 L'obiezione teologica 5
1.1.2 L'obiezione della "testa nella sabbia" 5
1.1.3 L'obiezione matematica 6
1.1.4 L'argomento dell'autocoscienza 6
1.1.5 Argomentazioni fondate su incapacità varie 8
1.1.6 L'obiezione di Lady Lovelace 9
1.1.7 L'argomentazione fondata sulla continuità del sistema nervoso 10
1.1.8 L'argomentazione del comportamento senza regole rigide 10
1.1.9 L'argomentazione fondata sulla ESP 11
1.2 Il test di Turing totale 11
1.3 Il premio Loebner 12
1.4 ELIZA 12

2 Intelligenza artificiale debole **15**
2.1 La stanza cinese 15
2.1.1 La risposta dei sistemi 17
2.1.2 La risposta del robot 17
2.1.3 La risposta del simulatore del cervello 18
2.1.4 La risposta della combinazione 19
2.1.5 La risposta delle altre menti 20
2.1.6 La risposta delle molte dimore 20

3 Intelligenza artificiale forte **23**
3.1 Il cervello nella vasca 24
3.2 La sostituzione del cervello 25
3.3 Etica relativa all'intelligenza artificiale 27

4 Lo stato dell'arte **31**

Conclusione **37**

Introduzione

L'uomo ha cercato di costruire forme di vita intelligenti molto prima di quanto si possa pensare. Già Aristotele (384-322 a.C.) tentò di formulare un insieme preciso di leggi in grado di governare la parte razionale della mente. Molto tempo dopo, Thomas Hobbes (1588-1679) affermò che il ragionamento fosse paragonabile al compiere operazioni di tipo matematico, riducendo la *ragione* ad una macchina calcolatrice. Nel 1642 Blaise Pascal (1623-1662), matematico e filosofo francese, costruì la *Pascalina*: considerata a lungo la prima addizionatrice meccanica; anche se Wilhelm Schickard ne costruì una circa 20 anni prima. Leibniz (1646-1716), introducendo un meccanismo speciale, incrementò le operazioni eseguibili da una macchina, costruendone una in grado di svolgere addizioni, sottrazioni, moltiplicazioni, ed estrazioni di radice. Egli credeva, come Hobbes, che "ragionare equivalesse a calcolare": sognava, infatti, di ridurre qualsiasi ragionamento in calcolo, cosicché si potesse risolvere qualsiasi controversia intellettuale come si fa con un conto aritmetico. Successivamente, alcuni filosofi ipotizzarono che le macchine potessero arrivare effettivamente a pensare e agire da sole. Sempre Hobbes nel 1651 con il *Leviatano* ideò un "animale artificiale" sostenendo: "Che cos'è infatti il cuore se non una molla e che cosa sono i nervi se non altrettanti fili e che cosa le giunture se non altrettante ruote che danno movimento all'intero corpo". Il primo passo verso le macchine "programmabili" fu fatto da Charles Babbage (1791–1871), che con i suoi progetti gettò le basi per la nascita dell'informatica. È, però, nel Novecento che si hanno i maggiori contributi alla *computer science* e alla questione dell'intelligenza delle macchine. Nel 1936 Alan Turing (1912–1954) propose un modello ideale di *calcolatore universale*, la macchina di Turing, e con esso definì quali funzioni fossero *computabili*. Lo stesso Turing nel 1950 pubblicò un articolo nel quale introdusse il famoso test, che poi prenderà il suo nome, per stabilire se e quando una macchina si possa considerare intelligente. Con le obiezioni, le riflessioni, le idee e gli spunti forniti in quell'articolo apre la strada alla nascita di una nuova disciplina, strettamente collegata alle scienze cognitive

e all'informatica.[1] Solo sei anni dopo, nel 1956, grazie a John McCarthy si sarebbe, infatti, tenuto il seminario della durata di due mesi al Dartmouth College in cui nacque l'*"intelligenza artificiale"*, spesso abbreviata in IA. Il workshop aveva l'obiettivo di riunire un gruppo selezionato di scienziati, i quali per un'intera estate avrebbero lavorato su alcuni dei principali aspetti problematici dell'intelligenza artificiale. Il documento ufficiale del progetto di ricerca, proposto da C.E.Shannon, M.L.Minsky, N.Rochester e lo stesso J.McCarthy, riportava quanto segue: "Lo studio procederà sulla base della congettura per cui, in linea di principio, ogni aspetto dell'apprendimento o una qualsiasi altra caratteristica dell'intelligenza possano essere descritte così precisamente da poter costruire una macchina che le simuli. Si tenterà di capire come le macchine possano utilizzare il linguaggio, formare astrazioni e concetti, risolvere tipi di problemi riservati per ora solo agli esseri umani e migliorare se stesse."[McCarthy et al., 1955] Il seminario non portò a particolari innovazioni, ma ebbe come pregio quello di riunire i principali protagonisti della disciplina, i quali negli anni a venire avrebbero dominato il campo dell'IA.

Questa tesi si prefigge l'obiettivo di sviscerare, in maniera critica, l'articolo di Alan Turing del 1950, considerato da molti il manifesto dell'intelligenza artificiale. Insieme ad esso verrà analizzata l'opinione contrastante di John Searle, presentando la sua obiezione al test di Turing: la *stanza cinese*. In questo modo si delineeranno i due filoni filosofici che accompagnano l'IA dalla sua nascita, ovvero l'intelligenza artificiale debole e forte. Mostrando i loro tratti caratteristici e gli esperimenti che hanno plasmato le due correnti principali di pensiero si arriverà, poi, alla presentazione degli obiettivi che sono stati già raggiunti e quelli, la maggior parte, che sono ancora lontani da essere soddisfatti.

Nel primo capitolo si presenterà, quindi, il *test di Turing* originale accompagnato dalle obiezioni relative ad esso; si farà, inoltre, riferimento ad una sua variante chiamata *test di Turing totale* e al caso ELIZA di Joseph Weizenbaum. Nel capitolo relativo all'intelligenza artificiale debole, verrà presentato, corredato dalle sue critiche, l'esperimento mentale della *stanza cinese* di John Searle, una possibile confutazione dell'intelligenza artificiale forte. Nel terzo capitolo saranno esposti gli esperimenti che alimentano le speranze dell'IA forte. Infine, nel capitolo riguardante lo stato dell'arte, verranno elencati i filoni di ricerca che accompagnano questa disciplina in continua evoluzione.

[1] Fonti storiche dal libro "Intelligenza artificiale. Un approccio moderno" [Russel and Norvig, 2010].

Capitolo 1

Il test di Turing

"Can machines thinks?"[Turing, 1950]

"Possono pensare le macchine?", è questa la domanda che pone all'attenzione Alan Turing nel suo articolo *Computing machinery and intelligence* del 1950 pubblicato sulla Rivista *Mind*. Una domanda a cui è difficile dare una risposta, se non si è prima definito il significato dei termini *macchina* e *pensare*. È questo il motivo che portò Turing ad ideare un esperimento concettuale, un Gedankenexperiment, per stabilire se una macchina sia, o meno, in grado di pensare.[1] Il *test di Turing* è una variazione del *"gioco dell'imitazione"*. Nella versione originale del gioco i partecipanti sono tre: un

[1]Il test di Turing venne in qualche modo anticipato da Cartesio nel 1637, nella quinta parte del *"Discorso sul metodo"* di cui qui viene riportato un estratto:

> "Qui in particolare mi ero fermato per far vedere che se ci fossero macchine con organi e forma di scimmia o di qualche altro animale privo di ragione, non avremmo nessun mezzo per accorgerci che non sono in tutto uguali a questi animali; mentre se ce ne fossero di somiglianti ai nostri corpi e capaci di imitare le nostre azioni per quanto è di fatto possibile, ci resterebbero sempre due mezzi sicurissimi per riconoscere che, non per questo, sono uomini veri. In primo luogo, non potrebbero mai usare parole o altri segni combinandoli come facciamo noi per comunicare agli altri i nostri pensieri. Perché si può ben concepire che una macchina sia fatta in modo tale da proferire parole, e ne proferisca anzi in relazione a movimenti corporei che provochino qualche cambiamento nei suoi organi; che chieda, ad esempio, che cosa si vuole da lei se la si tocca in qualche punto, o se si tocca in un altro gridi che le si fa male e così via; ma non si può immaginare che possa combinarle in modi diversi per rispondere al senso di tutto quel che si dice in sua presenza, come possono fare gli uomini, anche i più ottusi."

uomo A, una donna B e un esaminatore C. L'esaminatore C, che può essere uomo o donna indifferentemente, si trova in una stanza separata dagli altri partecipanti. Lo scopo del gioco per l'esaminatore è quello di determinare quale degli altri due partecipanti è l'uomo e quale la donna. Egli conosce i partecipanti con due 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". Lo scopo di A è quello di ingannare C e fare in modo che dia una identificazione sbagliata. B ha, invece, il compito di aiutare l'esaminatore. Per determinare la risposta l'esaminatore si può basare solo su una serie di domande poste ad A e B. Per non far sì che il tono di voce o la scrittura possano influenzare l'esaminatore, le risposte possono essere battute a macchina, o, in alternativa, si potrebbe mettere in comunicazione le due stanza tramite una telescrivente. Un'altra opportunità è quella di far ripetere domande e risposte da un intermediario. A questo punto Turing immagina di sostituire ad A una macchina e pone una domanda: "L'interrogante darà una risposta errata altrettanto spesso di quando il gioco viene giocato tra un uomo e una donna?". Quest'ultima domanda quindi sostituisce la domanda originale: "Possono pensare le macchine?".

L'ultima versione del gioco dell'imitazione proposta da Turing ha il pregio di fornire una soddisfacente definizione operativa di intelligenza senza fare alcun riferimento ai termini *macchina* e *pensare*, così facendo si evitano le difficoltà riguardanti il significato di queste parole. L'utilità di questo esperimento non sta tanto nella risposta che può fornirci, "quanto alla possibilità che esso offre di analizzare concetti come *mente*, *pensiero* e *intelligenza*"[Longo, 2009].

1.1 Obiezioni al test di Turing

Turing credeva che entro la fine dello scorso secolo sarebbe stato possibile programmare calcolatori, con una capacità di memorizzazione di circa 10^9, in modo tale che potessero giocare così bene il gioco dell'imitazione che un esaminatore medio non avrebbe avuto più del 70% di probabilità di compiere l'identificazione corretta, dopo 5 minuti di interrogazione. Inoltre, sosteneva che entro 50 anni sarebbe talmente mutato l'uso delle parole, *macchina* e *pensare*, e l'opinione a esse associata, che chiunque avrebbe potuto parlare di macchine pensanti senza il pericolo di essere contraddetto.

Nel suo stesso articolo Turing anticipa e riassume le opinioni, opposte alle sue, sollevate dai critici sulla validità della domanda: "possono pensare le macchine?". Critiche che si concentrano anche sulla variante del gioco dell'imitazione da lui proposta. Le obiezioni, che verranno citate, sono le stesse dell'articolo originale di Alan Turing [Turing, 1950], con osservazioni,

approfondimenti e considerazioni tratte dal libro "Intelligenza artificiale. Un approccio moderno" [Russel and Norvig, 2010].

1.1.1 L'obiezione teologica

"Il pensare è una funzione dell'anima immortale dell'uomo. Dio ha dato un'anima immortale ad ogni uomo e donna, ma non agli altri animali o alle macchine. Perciò nessun animale o macchina può pensare."[Turing, 1950] Secondo questa visione *dualistica*, sviluppata principalmente da Cartesio (1596-1650), vi sarebbe una netta separazione tra *anima* e *corpo*. Il pensiero, che risiede nell'anima, sarebbe quindi prerogativa dell'uomo e non degli esseri inanimati e degli animali. Questo si può considerare come una limitazione dell'onnipotenza di Dio. Ma ciò, secondo Turing, non escluderebbe la possibilità della "donazione" di un'anima anche ad un elefante, se questo dimostrasse di avere un cervello abbastanza sviluppato. Quindi sarebbe a discrezione di Dio concedere, o meno, il dono del pensiero anche ad una macchina che dimostrasse tali caratteristiche. Tuttavia, Turing rifiuta ogni obiezione teologica, affermando che esse risultano prive di importanza e poco utili ai fini della determinazione della risposta alla domanda principale, da lui avanzata.

1.1.2 L'obiezione della "testa nella sabbia"

Strettamente collegata all'obiezione teologica, questa critica può essere riassunta con la seguente frase: "Se le macchine pensassero, le conseguenze sarebbero terribili; speriamo e crediamo che esse non possano farlo". Il pensiero sopra espresso è frutto della nostra convinzione antropocentrica. Concezione che contagia la maggior parte di noi quando riflettiamo sul problema in questione. Ci piace credere che l'uomo sia in qualche modo misterioso, superiore al resto del creato. A maggior ragione, l'uomo si "nasconde" dietro il suddetto pensiero per timore che le macchine un giorno possano usurpare la sua posizione di comando. Secondo Turing anche questa opinione non ha abbastanza valenza per essere argomentata e confutata, in quanto rappresenta una comprensibile paura dell'uomo nei confronti di eventuali macchine pensanti.[2]

[2]Si pensi all'eventuale futuro scenario proposto dal film *Io, Robot* del 2004 diretto da Alex Proyas.

1.1.3 L'obiezione matematica

Esistono risultati che affermano che ad alcune questioni matematiche non si può dare risposta attraverso particolari sistemi formali. Il più famoso è il teorema di incompletezza di Gödel(1931), e dimostra che in ogni sistema formale di assiomi F abbastanza potente da gestire l'aritmetica è possibile costruire una cosiddetta "formula di Gödel" *G(F)* con le seguenti caratteristiche[Russel and Norvig, 2010]:

- *G(F)* è una formula di F, ma non può essere dimostrata al suo interno,
- se F è consistente, allora *G(F)* è vera.

Vi sono altri risultati, simili in alcuni aspetti, dovuti a Church, Kleene, Rosser e lo stesso Turing(1936). Il più conveniente da esaminare è quest'ultimo, dato che si riferisce direttamente alle macchine. Il risultato ottenuto da Turing riguarda un calcolatore numerico a capacità infinita. Esso dice che vi sono alcune cose che le macchine non possono fare. Il più classico esempio di problema non risolubile in modo automatico è il "Problema della fermata".

Questo è il risultato matematico: si sostiene che esso dimostri un'incapacità alla quale l'intelletto umano non è soggetto. Ma anche se noi accettassimo i limiti dei computer, nulla prova che gli esseri umani siano immuni da tali limitazioni. È fin troppo facile dimostrare rigorosamente che un metodo formale non può fare X, e poi affermare che gli esseri umani *possano* fare X grazie a qualche metodo informale, senza portare alcuna motivazione a riguardo. In effetti, è impossibile dimostrare che gli umani non siano soggetti al teorema di incompletezza di Gödel, perché ogni prova rigorosa richiederebbe una formalizzazione del talento umano che si sostiene, a priori, non formalizzabile.

1.1.4 L'argomento dell'autocoscienza

Molti filosofi e intellettuali hanno affermato che il test di Turing non può essere uno strumento valido per verificare se una macchina stia realmente pensando, anche se quest'ultima riuscisse a superare brillantemente la suddetta prova. Ciò che sta dietro a queste critiche è l'argomento dell'autocoscienza, ovvero le macchine anche se presentassero un comportamento intelligente non sarebbero comunque in grado di comprendere e di essere consapevoli di sé. L'autocoscienza esprime, in modo generico, la coscienza che l'*io* ha di se stesso [3]. Questa obiezione, ancora una volta anticipata dallo stesso Turing,

[3]Definizione di *autocoscienza* dal vocabolario on-line Treccani: `http://www.treccani.it/vocabolario/autocoscienza/`

si può riassumere citando le parole del professor Geoffrey Jefferson prese dal suo articolo *The mind of mechanical man* pubblicato nel 1949 sul British Medical Journal:

> "Fino a quando una macchina non potrà scrivere un sonetto o comporre un concerto in base a pensieri ed emozioni provate, e non per la giustapposizione casuale di simboli, non potremo essere d'accordo sul fatto che una macchina eguagli il cervello - cioè, che non solo scriva ma sappia di aver scritto. Nessun meccanismo potrebbe sentire (e non semplicemente segnalare artificialmente, che sarebbe un facile trucco) piacere per i suoi successi, dolore quando una sua valvola fonde, arrossire per l'adulazione, sentirsi depresso per i propri errori, essere attratto dal sesso, arrabbiarsi o abbattersi quando non può ottenere quel che desidera."

Turing ribatte brillantemente a questa affermazione dicendo che l'obiezione proposta risulta mal definita, proprio come domandarsi "le macchine possono pensare?". Tale critica porta, quindi, ad interrogarsi se le macchine possano pensare ed essere coscienti di quello che stanno facendo, ma nella sua visione più estrema l'unico modo per verificare se una macchina stia effettivamente pensando è proprio quello di essere la suddetta macchina e *sentire* se stessi pensare. Tuttavia nella vita di tutti i giorni non possiamo sapere se un uomo stia realmente pensando, per saperlo sarebbe necessario essere quel particolare uomo, questo è il punto di vista solipsistico [4]. Punto di vista che potrebbe facilmente portarci ad una situazione in cui A può legittimamente credere che "A stia pensando, mentre B no" e nel frattempo B sia convinto che "B pensi, e A invece no". Turing risolve questa controversia adottando l'*educata convenzione*, secondo la quale *tutti* pensino, e sostenendo che lo stesso Jefferson accetterebbe volentieri il gioco dell'imitazione come prova se solo avesse esperienza di macchine in grado di comportarsi intelligentemente. Se una macchina fosse in grado di sostenere un dialogo con un umano come questo:

> «UMANO: Nel primo verso del tuo sonetto che recita "Dovrei paragonarti a un giorno d'estate", non pensi che "giorno di primavera" funzioni ugualmente bene, se non meglio?
> MACCHINA: Non sta nella metrica.
> UMANO: Che dire di "giorno d'inverno"? Quello sta bene.
> MACCHINA: Si, ma nessuno vuole essere paragonato ad un giorno d'inverno.

[4] Prendendo la definizione dal *"Grande dizionario italiano"* Hoepli, il *solipsismo* è quella teoria filosofica secondo la quale il soggetto pensante si pone come la sola realtà, per cui il mondo esterno appare solo come una sua momentanea percezione.

> UMANO: Diresti che Mr. Pickwick ti ricorda il Natale?
> MACCHINA: In un certo senso.
> UMANO: Eppure il Natale è un giorno d'inverno, e non penso che al Sig. Pickwick spiacerebbe il paragone.
> MACCHINA: Non penso che tu stia parlando seriamente. Dicendo "giorno d'inverno" si intende un giorno tipico, e non uno speciale come il Natale.»[Turing, 1950]

allora forse molti di quelli che sostengono l'obiezione dell'autocoscienza, come il professor Jefferson, la abbandonerebbero e forse accetterebbero l'educata convenzione. Turing ammette che l'argomento dell'autocoscienza è complicato e, tuttora, porta con sé alcuni misteri. Tuttavia sostiene che per rispondere alle domande che ci siamo posti sulle macchine non è necessario riuscire a svelare i misteri che si celano dietro al concetto di coscienza. Costruire macchine coscienti è un ulteriore passo in avanti che si dovrà compiere una volta che avremo compreso appieno il concetto di intelligenza e la sua corrispondente forma "artificiale" .

1.1.5 Argomentazioni fondate su incapacità varie

Gli argomenti che interessano questa tipologia di critica assumono la forma "una macchina non potrà mai fare X". Turing nel suo articolo ha elencato alcuni esempi di X: "essere gentile, pieno di risorse, bello, amichevole, avere iniziativa, avere senso dello humour, riconoscere ciò che è giusto e sbagliato, commettere errori, innamorarsi, gustare le fragole con la panna, far innamorare qualcuno, apprendere dall'esperienza, usare le parole nel modo appropriato, essere l'oggetto del proprio pensiero, esibire una diversità di comportamenti pari a quella di un essere umano, fare qualcosa di veramente nuovo."[Turing, 1950]

Turing fa notare che questa critica può sorgere da un'errata applicazione del principio d'induzione scientifica. Infatti ogni macchina può fallire nell'eseguire uno scopo leggermente diverso da quello per cui è stata costruita. Ciò, ci porta alla conclusione che tutte le macchine posseggano tali proprietà. Potrebbe anche essere costruita una macchina in grado di "gustare le fragole con la panna", ma il tentativo sarebbe futile in relazione all'obiettivo che ci siamo posti all'inizio. Indubbiamente, risulterebbe più interessante esaminare alcune incapacità correlate ad essa, come la difficoltà che tra l'uomo e la macchina si stabilisca un rapporto di amicizia. Riguardo alle "macchine che non possono sbagliare", invece, Turing fa notare che, se fossero opportunamente programmate, potrebbero utilizzare l'induzione scientifica per arrivare a delle conclusioni, e talvolta, tale metodo, potrebbe condurle a risultati erronei. Per quanto concerne l'affermazione che "non può essere oggetto del

proprio pensiero" si può rispondere solamente se prima si dimostra che la macchina ha effettivamente un qualche pensiero su un qualche oggetto, ciò ci riporterà sicuramente alle argomentazioni sull'autocoscienza. Ma se ne consideriamo una che sta tentando di risolvere l'equazione $x^2 - 40x - 11 = 0$ allora potremo affermare con certezza che l'equazione rappresenta l'argomento di cui la macchina si sta occupando. In questo senso *può occuparsi di se stessa*, osservando i risultati del proprio comportamento può modificare i suoi programmi in modo da conseguire con maggior efficacia un determinato scopo, nel nostro esempio la soluzione all'equazione.

Dalla pubblicazione dell'articolo di Turing ad oggi sono stati fatti numerosi passi avanti e alcune capacità possono apparirci ora realizzabili mentre un tempo erano solo immaginabili. Ora abbiamo programmi che sono in grado di giocare a scacchi e ad altri giochi, che pilotano macchine ed elicotteri, che diagnosticano malattie e che svolgono svariati compiti, altrettanto bene o meglio degli esseri umani. A queste abilità si aggiungono piccole scoperte che i computer hanno fatto in campi come la matematica, l'astronomia, la chimica, la biologia e in molti altri ancora. Gli algoritmi di oggi possono svolgere alcune attività che richiedono la capacità di "apprendere dall'esperienza" e di "riconoscere ciò che è giusto e sbagliato", mediante l'utilizzo di semplici algoritmi statistici di apprendimento, con risultati più che soddisfacenti paragonabili a quelli di esperti [Russel and Norvig, 2010]. Nel 1950 il problema "avere un comportamento vario quanto quello umano" si riduceva ad un semplice problema di capacità di memorizzazione, ma oggi che la memoria non è più un problema possiamo affermare che siamo ancora ben lontani da creare programmi in grado di eseguire tutto ciò che fa un umano. Tra le attività che i computer non sono in grado di compiere si può ancora citare quella richiesta da Turing per verificarne l'intelligenza, ovvero sostenere una conversazione ad argomento libero.

1.1.6 L'obiezione di Lady Lovelace

Ada Lovelace (1815 - 1852) fu una matematica inglese nota per i suoi studi sulla *macchina analitica* di Charles Babbage. Essa, ridimensionando l'idea che la macchina fosse "pensante" alla maniera dell'uomo, affermò che: "La macchina analitica non ha la pretesa di "creare" alcunché. Può fare *qualsiasi cosa sappiamo come ordinarle di fare*". Douglas Rayner Hartree (1897 - 1958) riprese l'affermazione precedente aggiungendo che ciò non implica l'impossibilità di costruire un'apparecchiatura elettronica capace di "pensare per proprio conto" o, realizzabile tramite l'inserimento di un riflesso condizionato che servirebbe come base per l'apprendimento. Supponiamo che esista una macchina a stati discreti con queste ultime proprietà elencate. Allora, la mac-

china analitica di Babbage, considerata un calcolatore universale, potrebbe essere programmata opportunamente per poterla imitare; quest'ultima osservazione non venne in mente né a Babbage né a Lady Lovelace. Una variante di questa obiezione afferma che una macchina "non può fare mai veramente qualcosa di nuovo", alla quale si può rispondere semplicemente con il detto: "non c'è nulla di nuovo sotto il sole". Se, invece, si sostiene che una macchina non potrà mai "prenderci alla sprovvista" si incappa in una sorta di visione soggettiva al quale potremmo rispondere solo se siamo disposti a generare, nei confronti del calcolatore, un "atto mentale creativo". Difatti, dinnanzi a qualsiasi situazione sono necessari determinati atti mentali per giudicare se ciò che stiamo provando ci risulti, o meno, una sorpresa. Ma, sostenendo ciò riportiamo l'attenzione sull'argomento dell'autocoscienza.

1.1.7 L'argomentazione fondata sulla continuità del sistema nervoso

Il sistema nervoso non è certo rappresentabile mediante una macchina a stati discreti. Un errore, anche piccolo, di informazione che riguarda la grandezza dell'impulso elettrico che colpisce un neurone può risultare determinante per la generazione dell'impulso in uscita. Turing fa, però, notare che per il gioco dell'imitazione questa differenza non è rilevante, in quanto l'interrogante non sarà in grado di trarre alcun vantaggio dalla differenza tra macchina continua e discreta, se quest'ultima opportunamente programmata.

1.1.8 L'argomentazione del comportamento senza regole rigide

Chiamata anche "l'argomentazione derivante dall'informalità del comportamento" è una delle critiche più influenti e durature rivolte all'IA [Russel and Norvig, 2010]. Essa può essere riassunta mediante il *problema di qualificazione*, ovvero l'incapacità di catturare tutto in un insieme di regole logiche. Da questo punto di vista, quindi, non è possibile presentare un complesso di regole che descrivano ciò che debba fare un uomo in ogni possibile circostanza. Si potrebbero, sì, fornire alcune regole di condotta necessarie per determinare il comportamento che un individuo dovrebbe seguire in alcune situazioni particolari, ma appare impossibile fornirne a sufficienza per includere qualsiasi eventualità. Il comportamento umano risulterebbe troppo complesso per essere catturato da un semplice insieme di principi, e dal momento che le macchine non possono fare altro che eseguire istruzioni, non

potranno generare un comportamento intelligente come quello degli esseri umani.

Il *problema di qualificazione* si va ad aggiungere ad altre, già note, complessità riguardanti la progettazione e la realizzazione di sistemi intelligenti. Complessità che affrontano il problema della conoscenza di fondo, del buon senso, dell'incertezza, dell'apprendimento e dei processi decisionali. Diventa quindi fondamentale la ricerca, in quanto si occupa di *reti neurali*, *data mining*, *apprendimento non supervisionato*, *apprendimento per rinforzo*, *ragionamento automatico*, *comunicazione* e *percezione*. Queste difficoltà non rappresentano, tuttavia, un fallimento per l'IA, sono questioni che evidenziano una continua evoluzione della disciplina, e non ne attestano la sua impossibilità.

1.1.9 L'argomentazione fondata sulla ESP

Con il termine ESP[5] ci si riferisce a ogni percezione che non possa essere attribuita ai cinque sensi, ne sono esempi la chiaroveggenza, la telepatia, la precognizione e la psicocinesi. Se si accettano come validi questi fenomeni paranormali si mettono a repentaglio tutte le nostre comuni idee scientifiche. Inoltre viste le difficoltà che abbiamo riscontrato finora per la determinazione del *pensiero* potremmo pensare che la ESP risulti particolarmente importante per la sua spiegazione. Se si ammette la percezione extrasensoriale bisognerebbe rivedere il gioco dell'imitazione proposto: ora potrebbe accadere qualsiasi cosa. Il nostro test potrebbe venire alterato dai poteri psicocinetici di chi interroga, il quale potrebbe agire sulla macchina e modificare le sue risposte. D'altra parte potrebbe essere in grado di indovinare senza proferire alcuna domanda, semplicemente per chiaroveggenza. A questo punto servirebbe quindi una "camera a prova di ESP", in cui inserire i partecipanti, per soddisfare ogni esigenza e per non invalidare il test proposto.

1.2 Il test di Turing totale

Il test di Turing "classico" descritto precedentemente evita deliberatamente l'interazione diretta tra l'esaminatore e il computer, non essendo richiesta la simulazione fisica di una persona umana. Tuttavia, esiste in letteratura un *test di Turing totale* che include la presenza di un segnale video, per consentire all'esaminatore di verificare le capacità percettive dell'individuo, e prevede

[5]ESP: acronimo dell'espressione inglese *Extra-Sensory Perception*, percezione extrasensoriale (Wikipedia: *Percezione extrasensoriale*, `http://it.wikipedia.org/wiki/Percezione_extrasensoriale`).

la possibilità di scambio di oggetti fisici attraverso una finestrella. A questo punto per poter superare il test di Turing totale la macchina dovrà essere dotata di *visione artificiale*, per la percezione dell'ambiente e degli oggetti, e qualche applicazione della *robotica*, per la manipolazione degli oggetti e lo spostamento.

1.3 Il premio Loebner

Il test di Turing è stato considerato per la sua importanza come esperimento mentale e come spunto per discussioni riguardanti l'intelligenza e la coscienza; e come tale è rimasto fino al 1991. Infatti, nessuno prima di allora aveva pensato ad applicare e testare realmente il suddetto test. Il primo fu Hugh Loebner, programmatore e uomo d'affari, che istituì un premio in denaro da assegnare al primo computer in grado di superare il test. Le regole del concorso sono le stesse elencate da Turing nell'articolo del 1950.

1.4 ELIZA

Nel 1964 Joseph Weizenbaum, a quei tempi ricercatore del MIT, realizzò il programma ELIZA, nome della protagonista femminile del *Pigmalione* di George Bernard Shaw. Si trattava di un programma piuttosto semplice che, senza la pretesa di comprendere il linguaggio, era in grado di conversare su argomenti circoscritti. ELIZA ricoprì il ruolo di "psicoterapeuta" e riuscì ad ingannare parecchie persone, facendo credere, a chi stava dialogando con questo *chatterbot*[6], di essere in presenza di un vero medico in grado di curare disturbi psicopatologici, tanto che alcuni provarono addirittura sollievo dopo le sedute. In realtà questo programma era ben lontano dal soddisfare il test di Turing, erano infatti stati usati semplici "trucchi" di programmazione. Il funzionamento del programma è questo: ELIZA cerca nell'affermazione del paziente qualche parola chiave, a quel punto, se la trova, elabora una frase secondo una qualche regola associata a quella parola; altrimenti produce un'osservazione acontestuale o, in altre situazioni, può ripetere una trasformazione precedente. Si può vedere una parte di una conversazione tra ELIZA e una paziente:

> Ragazza: Gli uomini sono tutti uguali.
> Eliza: In che senso?

[6]Dicesi *chatterbot* un software progettato per simulare una conversazione intelligente con esseri umani tramite l'uso della voce o del testo. (Wikipedia: *Chatterbot*, `http://it.wikipedia.org/wiki/Chatterbot`)

Ragazza: Trovano sempre il modo di tormentarci.
Eliza: Può farmi un esempio?
Ragazza: Beh è il mio ragazzo che mi ha fatto venire qui.
Eliza: È il suo ragazzo che l'ha fatta venire?
Ragazza: Dice che sono depressa.
Eliza: Mi dispiace sentire che è depressa.
Ragazza: È vero. Sono infelice.
Eliza: Pensa che il fatto di venire qui l'aiuterà a non essere infelice?

Come si può notare dal dialogo, il programma cerca di imitare una psicoterapia rogersiana, o centrata sul cliente. ELIZA ebbe un enorme successo, tanto che stupì lo stesso Weizenbaum. Lo psichiatra Kenneth Colby, colui che sosteneva che di lì a pochi anni i programmi come ELIZA potessero essere utilizzati nella pratica terapeutica, realizzò un programma simile. PARRY, questo era il nome del programma, simulava il comportamento linguistico di un paranoico, che riuscì ad ingannare parecchi psichiatri. Successivamente venne anche organizzata una seduta tra PARRY ed ELIZA.[Longo, 2009]

Capitolo 2

Intelligenza artificiale debole

Macchine che agiscono *come se* fossero intelligenti è l'ipotesi che sta alla base dell'*intelligenza artificiale debole*. Secondo questa visione filosofica il calcolatore viene considerato come uno strumento potentissimo applicato allo studio della mente, che ci permette di formulare e verificare ipotesi in una maniera più precisa e rigorosa [Searle, 1980]. L'intelligenza artificiale fu fondata dallo stesso John McCarthy con l'idea che l'*IA debole* fosse possibile[Russel and Norvig, 2010]. Questo perché nella proposta del seminario al Dartmouth College asseriva che ogni aspetto dell'apprendimento e dell'intelligenza potesse essere formalizzato e quindi simulato al calcolatore.

Fanno parte di questa categoria i "sistemi basati sulla conoscenza" o "sistemi esperti", chiamati successivamente "sistemi di supporto alle decisioni"[1].

2.1 La stanza cinese

La stanza cinese è un esperimento mentale ideato da Searle nell'articolo *"Minds, Brains and Programs"* pubblicato nel 1980.[2] Esso può essere considerato un esempio il cui scopo è quello di confutare la teoria dell'intelligenza artificiale forte.[3] Con l'ipotesi che, per verificare una qualunque teoria della mente è necessario domandarsi come funzionerebbero le cose se la nostra stessa mente funzionasse secondo i principi alla base della teoria in esame, Searle formulò il suo Gedankenexperiment.

Immaginiamo di porre un uomo, di madrelingua inglese, chiuso in una stanza assieme ad un grande foglio di carta interamente ricoperto da ideo-

[1] Sistemi capaci di risolvere problemi in domini limitati, utilizzando processi inferenziali.

[2] [Searle, 1980]

[3] L'idea di formulare questo esempio nasce per verificare l'attendibilità del programma di Roger Schank, capace di rispondere a domande relative a una storia anche se l'informazione che egli fornisce con queste risposte non era esplicitamente presente in essa.

grammi cinesi. Supponiamo che l'uomo non conosca assolutamente il cinese, né in forma scritta né parlata. Come ulteriore forma di sicurezza ipotizziamo che l'individuo in questione non sia nemmeno in grado di distinguere ideogrammi cinesi da quelli giapponesi; questi simboli appaiono ad esso come "scarabocchi privi di significato". Insieme al primo foglio ne viene fornito un secondo, anch'esso in cinese, e con questo un *set* di regole per mettere in relazione i due fogli. Le regole sono scritte in inglese, quindi totalmente comprensibili all'individuo. Esse permettono di correlare un insieme di simboli in un altro insieme di simboli, identificandoli semplicemente in base alla loro forma grafica. Infine, all'uomo viene fornito un terzo foglio contenente ideogrammi cinesi e regole, queste ultime in inglese, che permettono di collegare elementi di quest'ultimo foglio con i primi due. Queste regole hanno lo scopo di insegnare a scrivere certi ideogrammi cinesi aventi una data forma, in risposta a determinati simboli assegnati nel terzo foglio, discriminati in base alla loro struttura grafica. Ad insaputa dell'uomo "rinchiuso" nella stanza, le persone che gli forniscono questi simboli chiamano:

- *scrittura*, il contenuto del primo foglio,
- *storia*, quello del secondo,
- *domande*, quello del terzo,
- *programma*, l'insieme delle regole consegnate all'uomo e
- *risposte alle domande*, i simboli che l'uomo restituisce in risposta al contenuto del terzo foglio.

Per complicare le cose, Searle ipotizza che all'uomo vengano anche fornite storie in inglese, da lui comprensibili, e che successivamente risponda alle domande incentrate su tali storie, sempre in inglese. Supponiamo inoltre che l'uomo diventi particolarmente bravo ad applicare le regole, a lui fornite, per la manipolazione dei simboli cinesi e che i *programmatori* diventino così abili a scrivere i programmi che dall'esterno della stanza le risposte date alle domande siano indistinguibili da quelle che darebbero persone di madrelingua cinese. Nessuno potrebbe pensare che l'uomo non conosca neanche una parola di cinese perché, dall'esterno, le risposte alle domande in cinese e in inglese sono buone nella stessa misura. Nel caso del cinese però l'uomo giunge alle risposte manipolando simboli formali non interpretati, cioè si comporta come un calcolatore; esegue operazioni di calcolo su elementi specificati per via formale. L'uomo quindi può essere visto come un'istanziazione del programma del calcolatore. Dunque tornando alla questione principale, l'IA forte sostiene che il calcolatore programmato capisca le storie e che il programma, in

un certo qual senso, spieghi le capacità di comprendere dell'uomo. Tuttavia, Searle ribadisce che l'uomo nella stanza, e per analogia il calcolatore, non sia in grado di comprendere una sola parola delle storie in cinese. Ciò che sta alla base del ragionamento di Searle è che la sintassi (grammatica) non è equivalente alla semantica (significato). Quindi conclude dicendo che anche eseguire un programma appropriato non è condizione *sufficiente* per essere una mente.

2.1.1 La risposta dei sistemi

Questa critica ha come oggetto l'esperimento mentale della stanza cinese. Essa sostiene che:

> "Pur essendo vero che l'individuo chiuso nella stanza non capisce la storia, sta di fatto che egli è solo parte di un sistema globale e questo sistema capisce la storia. [...] la comprensione non viene ascritta all'individuo isolato, bensì al sistema complessivo di cui egli è parte" [Searle, 1980].[4]

Searle davanti a questa obiezione ribatte dicendo che ciò non farebbe assolutamente spostare l'ago della bilancia in favore della teoria dell'intelligenza artificiale forte. Infatti, propone di far memorizzare tutte le regole all'individuo e di fargli eseguire i calcoli a mente, con la possibilità di lavorare all'aperto, sbarazzandosi della stanza. Così facendo il solo individuo incorpora tutto il sistema. Ma, di nuovo, egli non capirà nulla di cinese e lo stesso varrà per il sistema, poiché in esso non vi è nulla che non sia anche nell'individuo: se lui non capisce, neppure il sistema può capire.

2.1.2 La risposta del robot

Consideriamo l'opportunità di poter introdurre un calcolatore in un robot. Supponiamo che esso sia dotato di una telecamera per vedere, di braccia e gambe per agire e interagire con l'ambiente. Il calcolatore non accetta e manipola solo simboli formali, ma guida il robot nell'eseguire comportamenti simili a quelli umani: percepire, camminare, mangiare, bere... Il robot, ora "controllato" dal suo "cervello" (il calcolatore), avrebbe un'autentica comprensione e vari stati mentali.

Searle ribadisce che l'aggiunta di capacità motorie e percettive non modifichi nulla sotto il profilo della comprensione. L'unica cosa che fa è quella di

[4]Per analogia si può pensare all'individuo come un singolo neurone all'interno del cervello, da solo non può capire, ma contribuisce alla comprensione nel suo complesso.

ammettere implicitamente che la capacità cognitiva non è soltanto una questione di manipolazione di simboli formali, aggiungendo, difatti, un'insieme di rapporti causali col mondo esterno. Questo si allaccia all'approccio dei *processi cognitivi incorporati (embodied cognition).* Esso afferma che non ha senso considerare il cervello separatamente; i processi cognitivi hanno luogo in un corpo, il quale è immerso in un ambiente. Così, il cervello aumenta il proprio ragionamento facendo riferimento all'ambiente; visione artificiale e altri sensori diventano di importanza primaria[Russel and Norvig, 2010].

2.1.3 La risposta del simulatore del cervello

Supponiamo sia possibile creare un programma che simuli l'effettiva sequenza delle scariche neuroniche che avvengono nelle sinapsi del cervello di una persona di madrelingua cinese, quando questa risponde a domande riguardanti determinate storie. Storie, domande e risposte vengono presentate in lingua cinese. La macchina elaborerebbe le risposte appropriate sfruttando la simulazione formale del cervello autentico di un cinese. Se ciò fosse realizzabile, dovremmo sicuramente asserire che la macchina capisca le storie, perché se rifiutassimo di ammetterlo, dovremmo anche negare che le persone di madrelingua cinese siano in grado di comprendere tali storie.

Searle, prima di rispondere alla critica del simulatore del cervello, si interroga sul vero significato di intelligenza artificiale forte, considerando questa obiezione una "strana risposta". Infatti, credeva che alla base dell'IA forte ci fosse il seguente concetto: "non c'è bisogno di sapere come funzioni il cervello per sapere come funzioni la mente". Questo implica l'esistenza di un livello di operazioni mentali che possono essere realizzate nei processi cerebrali più svariati, proprio come un qualunque programma per calcolatore può essere realizzato in hardware diversi. "La mente sta al cervello come il programma sta all'hardware", quindi è possibile comprendere la *mente* senza ricorrere alla neurofisiologia. É su questi concetti che si basa l'intelligenza artificiale forte e, secondo Searle, se dovessimo sapere come funziona il cervello per fare l'IA, l'IA stessa non avrebbe alcun senso. Tuttavia, egli sostiene che neppure cogliere il funzionamento del cervello risulterebbe essere condizione sufficiente per la comprensione. Per spiegare quest'ultima affermazione Searle ricorre ad una variazione della stanza cinese. Supponiamo di collocare nella stanza un uomo che conosca solo l'inglese e che, invece di manipolare simboli, manovri delle valvole per fare scorrere, o meno, acqua in un complesso sistema di tubature idrauliche. Ogni connessione idraulica corrisponde a una sinapsi del cervello di un cinese e il sistema, una volta aperti tutti i rubinetti giusti, genererà le risposte in cinese. A questo punto ci si può chiedere dove risieda la comprensione in questo sistema. Esso ha in ingresso domande e fornisce

risposte, entrambe in cinese, ma né l'uomo né tanto meno i tubi lo comprendono. Non si può neanche incappare, secondo Searle, nell'errore di intendere la combinazione uomo più tubature come insieme capace di comprendere, perché così facendo si ritornerebbe alla questione legata alla "risposta dei sistemi", già affrontata precedentemente. Per concludere, Searle sostiene che il simulatore del cervello imiti le cose sbagliate, ovvero esso simula solo la struttura formale delle scariche all'interno del cervello e non le sue proprietà causali: risulta quindi incapace di generare stati intenzionali.[5] La dimostrazione che le proprietà *formali* non siano sufficienti per quelle *causali* è dimostrato, proprio, dall'esempio dei tubi idraulici.

2.1.4 La risposta della combinazione

Se le tre precedenti obiezioni, prese da sole, non sono riuscite a confutare la stanza cinese, considerate tutte insieme possono risultare decisive. Immaginiamo quindi un robot con un calcolatore a forma di cervello, programmato con tutte le sinapsi presenti in uno umano. Supponiamo inoltre che il comportamento del robot nel suo complesso non sia distinguibile dal comportamento dell'uomo. Infine, pensiamo l'insieme delle cose precedenti come un unico sistema e non come un semplice calcolatore con ingressi e uscite; in questo caso dovremmo attribuirgli intenzionalità.

Saremmo tentati, di primo acchito, di conferire stati intenzionali al robot. Questa attribuzione deriva dal fatto che, ora, esso presenta un comportamento simile al nostro, quindi, per estensione, potremmo supporre che esso abbia anche stati mentali della stessa natura dei nostri. Ma, se trovassimo una spiegazione autonoma per il suo comportamento, in particolare se sapessimo che esso funziona attraverso l'esecuzione di un semplice programma formale, noi non attribuiremmo più l'intenzionalità al robot. Come prova a favore di quest'ultima dichiarazione, Searle prova a spiegare il suo comportamento in questo modo: supponiamo di sapere che all'interno del robot ci sia un uomo che riceva simboli formali non interpretati, attraverso gli organi di senso del robot, li manipoli e li elabori attraverso un insieme di regole e invii simboli formali, non interpretati, agli organi di movimento. L'uomo elabora solo ed esclusivamente simboli formali, lui non vede ciò che vedono gli occhi del robot, non capisce le osservazioni fatte al, o dal, robot e nemmeno si

[5] Gli *stati intenzionali* si riferiscono a particolari stati, quali credere, conoscere, desiderare e temere, che fanno riferimento a un aspetto del mondo esterno.[Russel and Norvig, 2010] Gli stati intenzionali sono un concetto dell'*intenzionalità*, corrente filosofica della fenomenologia. L'intenzionalità non deve essere confusa con concetti come: libera volontà e agire "intenzionalmente". (Wikipedia: *Intenzionalità*, http://it.wikipedia.org/wiki/Intenzionalità)

propone di muovere i suoi arti; la sua elaborazione non è collegata in alcun modo a qualche stato intenzionale. In questo modo appena descritto il robot ci apparirebbe come un ingenuo fantoccio meccanico al quale è impossibile attribuire una mente e, quindi, intenzionalità.

2.1.5 La risposta delle altre menti

Questa obiezione, come quella successiva, è secondo Searle una critica che non coglie il punto centrale dell'argomentazione da lui introdotto con l'esperimento della stanza cinese, ma merita comunque una risposta, in quanto ricorrente.

La formulazione dell'obiezione delle altre menti è: "Come si fa a sapere se un'altra persona capisca il cinese o qualunque altra cosa?". In linea di principio si potrebbe capire solo osservando il suo comportamento. Dato che il calcolatore potrebbe superare prove comportamentali come un essere umano, e a questi ultimi siamo disposti ad attribuire capacità cognitive, allora dovremmo attribuirle anche ai calcolatori.

Questa critica si riallaccia fortemente all'*educata convenzione*, che Turing propone nell'argomentazione sull'autocoscienza. Searle però non si interroga su come sapere se altri possiedano capacità cognitive, ma a cosa gli si attribuisca nel caso loro le abbiano realmente a disposizione. Non si può, secondo la sua visione, considerare solo processi di calcolo con ingressi e uscite corrette, perché queste potrebbero esistere anche in assenza di stati cognitivi. Rimarcando il fatto che nella psicologia cognitivista, ed anche nell'esperimento da lui proposto, ci si interroga sulla mente e i processi ad essa correlati col presupposto che siano reali, non in base ad un semplice espediente che ci permetta di "bypassare" la mente e le sue funzionalità.

2.1.6 La risposta delle molte dimore

Quest'ultima critica afferma che: "Qualunque cosa siano questi processi causali che tu[6] ritieni necessari per l'intenzionalità (ammesso che tu abbia ragione), riusciremo prima o poi a costruire dispositivi che possiedano tali processi causali, e questa sarà l'intelligenza artificiale. Quindi le tue argomentazioni non toccano in alcun modo la capacità dell'intelligenza artificiale di generare e spiegare le facoltà cognitive" [Searle, 1980].

Searle sostiene che il progetto dell'IA forte era nato per cercare di "creare" processi mentali attraverso processi di calcolo su elementi definiti per via formale. Se si ridefinisce il progetto stesso di IA forte, come fa in effetti

[6]Rivolgendosi a Searle

questa obiezione, non c'è più la necessità di cercare prove e spiegazioni per confutarlo.

Capitolo 3

Intelligenza artificiale forte

> "Secondo l'IA forte, invece, il calcolatore non è semplicemente uno strumento per lo studio della mente, ma piuttosto, quando sia programmato opportunamente, è una vera mente; è cioè possibile affermare che i calcolatori, una volta corredati dei programmi giusti, letteralmente capiscono e posseggono altri stati cognitivi"[Searle, 1980]

L'ipotesi dell'*intelligenza artificiale forte*, sostenuta dai funzionalisti[1], asserisce che le macchine che si comportano intelligentemente stiano *effettivamente* pensando, e non semplicemente *simulando* il pensiero. L'IA forte porta, dunque, ad interrogarsi sul concetto di coscienza, in quanto comprensione e consapevolezza di sé, e ci permette di riflettere su aspetti che riguardano l'etica relativa allo sviluppo di macchine intelligenti.

In questo capitolo verranno presentati due esperimenti mentali che rappresentano metodi teorici appartenenti a due correnti ideologiche distinte all'interno della filosofia della mente. Con essi si cercherà di fornire una visione della mente che può, in linea teorica, dare adito alle speranze dell'IA forte.

[1]Il *funzionalismo*, branca della filosofia della mente sviluppata da Hilary Putnam nel 1950, sostiene che gli stati mentali (come i desideri, le convinzioni, etc.) siano costituiti solamente dal loro ruolo, cioè dalla loro funzione, la loro relazione causale, rispetto ad altri stati mentali, percezioni e comportamento. L'analogia mente/computer, che vede il cervello paragonato all'hardware e la mente al software, costituisce l'emblema di gran parte delle teorie funzionaliste della mente. (Wikipedia: *Funzionalismo (filosofia della mente)*, `http://it.wikipedia.org/wiki/Funzionalismo_(filosofia_della_mente)`)

3.1 Il cervello nella vasca

Per poter citare questo esperimento mentale[2] ideato da Hilary Putnam nel libro *Reason, Truth and History*, pubblicato nel 1981, bisogna fare un passo indietro ed introdurre il *problema mente-corpo*. Questa questione fu presa in considerazione già dagli antichi filosofi greci, ma fu analizzata approfonditamente solo a partire dal diciassettesimo secolo, con il filosofo e matematico francese Cartesio. Egli concluse che mente e corpo sono in qualche modo separati, dividendo così l'attività di pensiero della mente e i processi fisici del corpo. Diede così origine alla teoria *dualista*, teoria che presenta il problema di come la mente possa controllare il corpo in quanto separata da esso.[3] Una dottrina che evita questo problema è il *fisicalismo*, asserendo che la mente non è affatto separata dal corpo, gli stati mentali sono stati fisici. Il fisicalismo, in linea di principio, consente la possibilità dell'IA forte. Se accettiamo la visione filosofica del *fisicalismo* allora la descrizione appropriata dello stato mentale di una persona è determinata dallo stato cerebrale di quella persona. Quindi se al momento sono impegnato, e concentrato, a mangiare un hamburger, il mio stato cerebrale è un'istanza della classe di stati mentali "sapere di mangiare un hamburger". La questione fondamentale è che lo stesso stato cerebrale non può corrispondere ad uno stato mentale fondamentalmente distinto, come per esempio "sapere di mangiare una banana". Con l'esperimento mentale del "cervello nella vasca" si mettono in discussione le ipotesi del fisicalismo.[4]

Immaginiamo che il cervello di una persona (per esempio voi stessi), tramite un'operazione di uno scienziato malvagio, venga rimosso dal cranio alla nascita e successivamente inserito in una vasca, piena di liquido nutriente e perfettamente ingegnerizzata. La vasca mantiene in vita il cervello (il vostro), quest'ultimo viene collegato ad un supercomputer che, tramite segnali elettronici, invia al cervello una simulazione di un mondo totalmente fittizio. I segnali, a loro volta, provenienti dal cervello vengono utilizzati per modificare la simulazione in modo appropriato. Il computer è così abile che se la persona cerca di alzare il braccio la risposta del computer farà sì che egli

[2]Esperimento mentale che può essere considerato una rivisitazione contemporanea del "genio ingannatore" di Cartesio, contenuto nella sua opera *Principia philosophiae*. Esperimento che, con qualche piccola variazione, è stato portato anche sul grande schermo da Lana e Andy Wachowski nel 1999, con il film *The Matrix*

[3]Cartesio tentò di rispondere a questa critica ipotizzando un'interazione tramite la *ghiandola pineale*, ma questo ci riporta ad un'ulteriore domanda: come riesce la mente a controllare la ghiandola pineale?

[4]L'esperimento mentale del cervello nella vasca è stato preso dal libro *Reason, Truth and History*[Putnam, 1981], ed integrato con la sua rivisitazione presentata nel libro "Intelligenza artificiale. Un approccio moderno" [Russel and Norvig, 2010]

"veda" e "senta" il braccio in movimento. Lo scienziato può, in aggiunta, cancellare il ricordo dell'operazione di rimozione del cervello. Alla vittima potrebbe addirittura sembrare di essere seduto mentre sta leggendo queste stesse parole, divertenti, ma abbastanza assurde, riguardo l'ipotesi che possa esistere uno scienziato malvagio che rimuova i cervelli dal corpo delle persone e li metta in una vasca. La vita simulata replica in modo esatto la vita che avreste vissuto se il vostro cervello non fosse stato collocato nella vasca, anche per ciò che riguarda la simulazione del mangiare hamburger simulati. A questo punto potreste avere uno stato cerebrale *identico* a quello di qualcuno che sta realmente mangiando un hamburger, ma sarebbe letteralmente falso affermare che avete il suo stesso stato mentale, ovvero "sapere di mangiare un hamburger". Non state mangiando un hamburger, non avete mai provato un hamburger e non potreste, quindi, avere tale stato mentale.

L'esperimento appena presentato sembra essere in contrapposizione con la visione del fisicalismo, che considera gli stati cerebrali come coloro che determinino gli stati mentali. Per poter spiegare questo esempio si può introdurre un espediente che ci permette di considerare il contenuto degli stati mentali da due diversi punti di vista. Il punto di vista del *contenuto allargato* è quello di un osservatore esterno onnisciente che ha accesso all'intera situazione e può distinguere le differenze del mondo; qui il contenuto degli stati mentali coinvolge sia lo stato cerebrale sia la storia ambientale. Invece, nel *contenuto ristretto* consideriamo solo lo stato cerebrale, e così sia il mangiatore di hamburger *reale* che quello *simulato* nella vasca hanno lo stesso contenuto degli stati cerebrali.

Ora siamo in grado di rispondere alla domanda se un sistema di intelligenza artificiale, un calcolatore per esempio, sia in grado di pensare e di possedere realmente stati mentali. Non ha senso affermare che un sistema di IA possa pensare, o meno, in base a condizioni esterne ad esso, quindi è necessario e appropriato giudicarlo dal punto di vista del *contenuto ristretto*.

3.2 La sostituzione del cervello

L'esperimento mentale della "sostituzione del cervello" può essere considerato un'esemplificazione delle *teorie funzionaliste*. "I funzionalisti sostengono che la specificità della neurofisiologia umana non è probabilmente essenziale per produrre i fenomeni mentali. Creature con una struttura biologica diversa potrebbero giungere agli stessi risultati cognitivi con mezzi fisici diversi. E perfino un sistema non biologico fatto, per esempio, di rame, silicio e germanio potrebbe raggiungere lo stesso fine se possedesse un'organizzazione

interna adeguatamente analoga."[5] L'esempio che tratteremo venne introdotto da Clark Glymour, ma più frequentemente associato al lavoro di Hans Moravec.

L'esperimento viene così definito: supponiamo che la neurofisiologia si sia sviluppata a tal punto che sia il comportamento di input-output dei neuroni che il loro modo di essere connessi gli uni con gli altri siano compresi nei minimi dettagli. Immaginiamo che anche la tecnologia si sia evoluta e che si possano costruire microscopici chip elettronici che siano in grado di prendere il posto dei neuroni. Ora tramite una speciale operazione chirurgica supponiamo che sia possibile sostituire i neuroni, uno ad uno, con i nostri chip; senza interrompere il funzionamento del cervello nel suo insieme. Alla fine dell'operazione avremo rimpiazzato tutti quanti i neuroni di una persona e, secondo la definizione dell'esperimento, il comportamento esterno del soggetto rimarrà immutato rispetto a quello che si sarebbe osservato se l'operazione non fosse stata eseguita. Cosa possiamo dire riguardo alla *coscienza* del soggetto che è stato sottoposto a questo intervento?

Sarebbe difficile determinare dall'esterno la sua presenza o la sua assenza, quindi possiamo solo fare affidamento alla nostra intuizione per avanzare un'ipotesi di risposta, quanto meno, plausibile. Supponiamo di sottoporre alcune domande al soggetto in questione, quando ormai non gli rimane alcun neurone reale. Date le condizioni dell'esperimento le risposte saranno le più "normali" possibili, come quelle che ci si aspetterebbe da una persona con cervello umano. Se provassimo a colpire il soggetto con un bastone otterremo come risposta: "Ahio, che male!". Questo output però non può essere stato ottenuto mediante un semplice meccanismo di rilevazione di sensori, basandosi poi su un *mapping* tra segnali di ingresso e uscita. Infatti, effettuando l'operazione chirurgica, abbiamo replicato le proprietà funzionali di un normale cervello umano, e quindi siamo certi di non aver inserito nessun simile artificio in quello *elettronico*. Ciò ci porta a spiegare le manifestazioni di coscienza dell'individuo facendo riferimento alle sole proprietà funzionali dei neuroni all'interno del cervello elettronico. Ma, se quest'ultima considerazione viene ritenuta valida, si deve applicare la stessa spiegazione anche al *cervello reale*, in quanto presenta le stesse proprietà funzionali. Vanno quindi prese in esame le varie possibili conclusioni che si possono trarre da quest'esperimento:

1. i meccanismi causali della coscienza stanno ancora operando nel cervello elettronico,

[5]Definizione da enciclopedia Treccani: `http://www.treccani.it/enciclopedia/il-problema-mente-cervello_(Frontiere_della_Vita)/`)

2. gli eventi mentali non sono presenti nel cervello elettronico, quindi non è conscio,

3. l'esperimento è inattuabile, pertanto proporre ipotesi al riguardo non ha alcun senso.

Escludiamo la terza possibilità perché noi, in questa trattazione, siamo interessati alla questione filosofica e non alla realizzabilità, o meno, dell'esperimento. La seconda opzione proposta ci porta a ritenere la coscienza come un qualcosa che è ininfluente ai fini della determinazione dell'output del soggetto. La sua esclamazione "Ahio, che male!" non sarà stata pronunciata perché lui *sente* realmente male, cioè cosciente del dolore provato, ma sarà causata da un secondo meccanismo inconscio. Avendo, però, riprodotto il funzionamento di un cervello reale, dovremmo anche asserire che gli eventi mentali consci nel cervello umano non hanno un collegamento causale con il comportamento. Se quindi accettiamo il fatto che l'esperimento della sostituzione del cervello dimostra che quello elettronico è cosciente, dobbiamo concordare che la coscienza è conservata anche quando l'intero cervello è sostituito da un insieme di chip elettronici; ovvero la prima opzione proposta.

3.3 Etica relativa all'intelligenza artificiale

Finora ci siamo domandati solo se è possibile progettare e sviluppare un'intelligenza artificiale, ma non ci siamo interrogati se è *eticamente corretto* cercare di realizzarla. Se le tecnologie sviluppate con l'IA si rivelassero, un giorno, pericolose per la sopravvivenza dell'umanità stessa la responsabilità dovrà essere attribuita ai ricercatori ed ingegneri che per anni se ne sono occupati. Diventa quindi necessario capire, ora, l'importanza legata all'etica e ai rischi dello sviluppo dell'intelligenza artificiale. Se le macchine raggiungessero l'intelligenza umana, o addirittura la superassero, e sviluppassero una mente cosciente, si ribellerebbero all'uomo o rimarrebbero al suo fianco servendo l'umanità in tutte le sue mansioni? In un futuro popolato da sistemi intelligenti affioreranno nuove problematiche, le quali, probabilmente, minerebbero la posizione di dominio dell'essere umano. Se, invece, riuscissero a coesistere entrambe le specie, umanità e macchine pensanti, si potrebbe comunque presentare uno scenario poco piacevole per l'uomo. Quest'ultimo potrebbe essere soppiantato dalle macchine in qualsiasi ambito, lento e poco produttivo in confronto agli automi. Si ritroverebbe inferiore dal punto di vista intellettuale e lavorativo, schiavo della sua stessa "creazione". Questi problemi possono essere ora solo immaginati, ma se quel giorno, tanto atteso dai ricercatori, in cui le macchine riusciranno a pensare risultasse non

così lontano nel tempo? La fantascienza, esprimendo i desideri e le paure umane sulla tecnologia, potrebbe ricoprire un ruolo fondamentale in questa discussione, in quanto può influenzare il progresso dell'IA.[6] Così, il cinema e la letteratura, simulando scenari futuri, ci forniscono un'idea di quello che potrebbe essere, sperando che non sia una società distopica. Parecchi film descrivono robot come macchine sofisticate, costruite per servire l'uomo e capaci addirittura di provare sentimenti. Ne è un esempio *A.I. Intelligenza Artificiale* di Steven Spielberg, in cui David, il robot bambino, programmato per credere di essere umano non riesce ad accettare l'abbandono da parte della padrona-madre. Il film d'animazione *WALL-E*, di Andrew Stanton, raffigura un robot "spazzino" che, attraverso la visione di una videocassetta, acquisisce coscienza della sua solitudine e, successivamente, si innamora di EVE, un robot femmina. Ma, allo stesso tempo la fantascienza ritrae macchine pericolose capaci di cospirare contro l'umanità attraverso piani subdoli. Il supercomputer HAL 9000, protagonista del film *2001: Odissea nello spazio* di Stanley Kubrick, è dotato di una vera intelligenza artificiale, capace di vedere, parlare, giocare a scacchi, provare sentimenti e perfino in grado di elaborare un piano al di fuori degli schemi per uccidere gli astronauti, non appena si rende conto della possibilità di essere "disattivato". Emblematico è anche il caso di *Io, Robot*, di Alex Proyas, in cui i robot, pur governati dalle "tre leggi della robotica", si rivoltano all'uomo cercando di instaurare una dittatura, al fine di proteggere gli uomini da loro stessi, in base alla "prima legge".[7] Alcuni di questi film, oltre a ritrarre le possibili conseguenze dell'IA, ci aprono anche a scenari in cui i robot necessitino di diritti civili, ora prerogativa solamente degli uomini. Leggi che governino e tutelino le interazioni tra le macchine e tra macchina e uomo. Sempre più evidente è la necessità di progettare, fin dall'inizio, una *IA amichevole*, ovvero la realizzazione di sistemi che nutrono il desiderio di non nuocere agli uomini. L'introduzione

[6] [Buttazzo, 2002]

[7] Le *tre leggi della robotica* sono state introdotte dallo scrittore di fantascienza Isaac Asimov nella raccolta di racconti *Io, Robot* del 1950:

1. Un robot non può recar danno a un essere umano
   né può permettere che, a causa del proprio
   mancato intervento, un essere umano riceva danno.

2. Un robot deve obbedire agli ordini impartiti dagli esseri umani,
   purché tali ordini non contravvengano
   alla Prima Legge.

3. Un robot deve proteggere la propria esistenza,
   purché questa autodifesa non contrasti
   con la Prima o con la Seconda Legge.

di questo concetto di "amichevolezza" potrebbe, comunque, non essere sufficiente per proteggerci. I progettisti devono essere consci che i loro progetti potrebbero essere difettosi e le loro creazioni, una volta coscienti, potrebbero evolvere e modificare il loro comportamento.

Capitolo 4

Lo stato dell'arte

Dire cosa può fare oggi l'intelligenza artificiale è difficile se si guarda a questa disciplina con i presupposti e le aspettative che si avevano al momento della sua nascita. Tuttavia molti passi avanti sono stati compiuti, soprattutto in campi di competenza circoscritti e limitati. Numerosissime applicazioni e attività sono ora possibili grazie al continuo avanzamento dello studio e della ricerca di questa scienza. Oggigiorno i ricercatori si concentrano sul concetto di *agente intelligente*, ossia un'entità che intraprende la miglior azione possibile in ogni situazione. Questi agenti possono essere visti come sistemi che percepiscono l'ambiente, nel quale loro sono "immersi", e lo modificano agendo opportunamente. Qui di seguito verranno elencati i vari filoni di ricerca che interessano oggi l'intelligenza artificiale.[1]

Problem solving

Con questo termine si prendono in considerazione l'insieme dei processi atti ad analizzare, affrontare e risolvere positivamente situazioni problematiche.[2] Gli agenti per il problem solving, partendo da una condizione data, identificano idonee sequenze di azioni per raggiungere gli stati desiderati. Quando non può essere utilizzata una ricerca della soluzione ottimale, perché la complessità del problema risulterebbe impraticabile in termini di tempo e capacità di elaborazione, si utilizzano *algoritmi euristici*: tecniche basate sull'esperienza che portano più rapidamente ad una soluzione, tuttavia non ottima per il problema in questione.

[1] Tratti dall'articolo: *Intelligenza artificiale: i primi 50 anni* [Dapor and Aiello, 2004]

[2] (Wikipedia: *Problem solving*, `http://it.wikipedia.org/wiki/Problem_solving`)

Rappresentazione della conoscenza e ragionamento

La ricerca in questo settore ha come scopo quello di progettare linguaggi, metodi e tecniche per rappresentare la conoscenza su domini applicativi, utilizzando *ontologie*. Inoltre si occupa di algoritmi e metodi per fare inferenze e trarre valide conclusioni nel dominio di interesse. Il componente più importante degli *agenti basati sulla conoscenza* è proprio la "base di conoscenza", o KB (*knowledge base*), costruita attraverso asserzioni sul mondo. Per quanto riguarda i linguaggi utilizzati è necessario che siano sufficientemente espressivi, mentre i meccanismi impiegati per la deduzione devono avere una complessità di calcolo accettabile. L'intelligenza degli uomini non si basa solo su meccanismi puramente reattivi, ma utilizza processi di ragionamento che operano sulle rappresentazioni della conoscenza [Russel and Norvig, 2010].

Il *ragionamento automatico* rappresenta, quindi, il punto chiave per estrarre nuova conoscenza dalla *knowledge base* stessa, attraverso meccanismi inferenziali. Spesso un agente, non possedendo tutte le informazioni relative all'ambiente con cui deve interagire, deve decidere in condizioni di incertezza, basandosi quindi su *ragionamenti di tipo probabilistico*. In molte applicazioni viene utilizzata la *regola di Bayes*, capace di aggiornare i valori di probabilità in base ai dati via via raccolti.[3]

Pianificazione automatica

Dato uno stato iniziale, particolare configurazione di un dominio applicativo specifico, è compito della pianificazione automatica stabilire la sequenza di azioni necessarie per raggiungere un obiettivo (detto anche stato finale). Viene fatta distinzione tra la pianificazione attuata in ambienti completamente osservabili, deterministici e statici e quella che viene fatta su ambienti non deterministici e parzialmente osservabili: quest'ultima oggetto di intensa ricerca, data la complessità degli algoritmi in gioco.

Il primo programma di pianificazione autonoma fu REMOTE AGENT della NASA, software capace di gestire lo scheduling delle operazioni di un veicolo spaziale, occupandosi della generazione di piani partendo da obiettivi inviati dalla terra.[4] REMOTE AGENT monitorava l'esecuzione delle operazioni, rilevando, diagnosticando e recuperando dagli errori, non appena questi si verificavano.

[3] [Russel and Norvig, 2010]
[4] [Russel and Norvig, 2010]

Si può includere in questa categoria la *pianificazione logistica*, utilizzata persino in ambito militare per la gestione di persone, approvvigionamenti e veicoli. Utilizzando queste tecniche, che applicano metodi e concetti dell'IA, si riescono a elaborare piani strategici in poche ore, i quali, altrimenti, richiederebbero settimane di lavoro.

Apprendimento automatico

L'apprendimento, noto in letteratura come *machine learning*, riguarda la capacità di un agente di osservare i risultati dei processi delle sue stesse interazioni con l'ambiente, per modificare il suo comportamento ed estrarre nuova conoscenza. Si possono distinguere vari tipi di apprendimento: *simbolico*, per esempio tecniche che utilizzano alberi di decisione a partire da esempi, *connessionista* o *subsimbolico*, riguardante l'addestramento di reti neurali e, infine, *statistico*, basato su tecniche di tipo statistico. Tra i paradigmi di apprendimento più noti possiamo citare quelli che richiedono l'aiuto di un insegnante umano, *apprendimento supervisionato*, e quelli che possono funzionare autonomamente, *apprendimento non supervisionato* e *apprendimento per rinforzo*[5]. Il *data mining* è un concetto correlato all'apprendimento automatico, esso introduce tecniche e metodologie con le quali è possibile estrapolare un sapere o una conoscenza da grandi quantità di dati. Tale tecnica ha come obiettivo quello di identificare pattern e regolarità, che ci consentano poi di ipotizzare e verificare nuove relazioni di tipo causale fra i fenomeni del sistema, o di formulare previsioni di tipo statistico su nuovi dati[6].

Comunicazione e Percezione

La problematica della comunicazione coinvolge il linguaggio (le sue ambiguità, la sua sintassi e la sua semantica), la disambiguazione e i modelli che permettono la comprensione del contenuto del discorso. Questa disciplina si occupa dell'analisi e della generazione di linguaggio sia scritto che parlato, della traduzione di testo automatica e dell'elaborazione di riassunti. La ricerca di informazioni sul web può, quindi, rivelarsi utile e determinante per lo sviluppo di queste tecniche.

L'agente necessita della *percezione* per avere informazioni sull'ambiente circostante; questa, ottenuta mediante sensori, permette il riconoscimento

[5] [Russel and Norvig, 2010]

[6] (Wikipedia: *Data mining*, `http://it.wikipedia.org/wiki/Data_mining`)

di oggetti (tramite la *visione artificiale*) e la comprensione del linguaggio parlato.

Numerose applicazioni oggigiorno integrano percezione e comunicazione per interagire con l'uomo, basti pensare ai *call center* automatici: sistemi che fanno uso di riconoscimento vocale e tecniche di gestione dei dialoghi.

Robotica

I robot possono essere considerati agenti fisici artificiali che svolgono funzioni ed eseguono compiti manipolando oggetti del mondo fisico. Per poter agire sul mondo necessitano di *sensori*, per acquisire informazioni, e *attuatori* per manipolarlo.

La robotica, data la sua natura interdisciplinare, trova applicazioni nei contesti più disparati. Si possono trovare esempi nell'uso domestico, come l'aspirapolvere robotizzata *Roomba*, negli ambiti medicali e biomedicali, arti e protesi artificiali, e perfino per uso industriale, come le macchine automatiche per la produzione. I robot vengono utilizzati in ambienti proibitivi per l'uomo (esplorazione dello spazio e dei fondali oceanici) e in situazioni di estremo pericolo (come per la difesa militare). Rivestono un ruolo molto importante gli assistenti utilizzati dai chirurghi per eseguire delicate operazioni di microchirurgia sui pazienti.

Sistemi multiagente

Un agente artificiale non opera in modo solitario: esso deve interagire con esseri umani e altri sistemi. Considerando quindi una molteplicità di agenti bisogna tener conto di problemi di primaria importanza quali la comunicazione, la coordinazione e la cooperazione. Inoltre, ogni entità potrebbe agire in base alla sua conoscenza parziale, ovvero dal suo "punto di vista", cosicché aumenta la difficoltà di costruire un sistema che cerchi di soddisfare efficacemente un obiettivo comune. Risulta necessario imporre un'organizzazione per poter distribuire il carico di lavoro e per riuscire a sfruttare le caratteristiche di alcuni agenti, che presentano determinate capacità specifiche. La ricerca si occupa, dunque, del concetto di autonomia nei sistemi ad agenti intelligenti: è il campo dell'*autonomic computing*.

Programmazione ispirata a modelli biologici

Branca della ricerca automatica di soluzioni con l'obiettivo di riprodurre meccanismi che si osservano in natura. Fanno parte di questo ambito gli *algoritmi genetici* e gli *algoritmi evolutivi*, i quali simulano i meccanismi della selezione naturale e dell'evoluzione darwiniana. Questi fanno parte degli algoritmi euristici, utilizzati per cercare di risolvere problemi di ottimizzazione per i quali non si conoscono altri metodi di risoluzione efficienti di complessità lineare o polinomiale.

Modellazione Cognitiva

Ricerca interdisciplinare che ha l'obiettivo di fornire modelli artificiali di mente e cervello. Si occupano di modellazione cognitiva le discipline facenti parte delle *scienze cognitive*: neurofisiologia, neuroscienza cognitiva, psicologia cognitiva, intelligenza artificiale, linguistica cognitiva e filosofia della mente. Un modello cognitivo è infatti essenziale per lo sviluppo, per il controllo e per la supervisione di sistemi che superano un certo livello di complessità e di incertezza.

Rappresentazione delle emozioni

Obiettivo di questa branca della ricerca è quello di comprendere, modellare e, infine, generare *stati emotivi "artificiali"*. Tale studio aiuterebbe a migliorare i meccanismi psicologici, ad oggi non ben compresi, che consentono di cogliere gli aspetti impliciti del linguaggio naturale.

Conclusione

L'obiettivo di questa tesi è stato quello di ripercorrere parte della storia dell'intelligenza artificiale, rimarcando, in particolar modo, l'importanza del test di Turing. Includendo, ove possibile, trattazioni filosofiche sulla natura della mente: quest'ultima oggetto di dibattito che divide filosofi e psicologi fin dai tempi antichi.

L'IA nasce, nel 1956, con l'idea che ogni aspetto dell'apprendimento o una qualsiasi altra caratteristica dell'intelligenza possano essere descritte così precisamente da poter costruire una macchina che le simuli. Nel periodo di "gestazione" di questa disciplina ricoprì un ruolo fondamentale l'articolo pubblicato nel 1950 da Alan Mathison Turing, pietra miliare dell'IA stessa. In questo articolo egli introdusse il noto test (che prese poi il suo nome), il concetto di apprendimento automatico, gli algoritmi genetici e l'apprendimento per rinforzo; oltre a fornirci riflessioni e argomenti di discussione su intelligenza e coscienza, tuttora validi. Le diverse visioni della *mente* all'interno della più generale *filosofia della mente* possono essere ritenute le cause che hanno portato alla suddivisione dell'intelligenza artificiale stessa in due filoni: IA forte e IA debole. Presentando la *stanza cinese* abbiamo descritto il pensiero di John Searle, il quale si oppone al concetto di IA forte sostenendo una mancanza di intenzionalità da parte della macchina e aderendo di fatto al naturalismo biologico: da questo punto di vista la *coscienza* emerge dall'organismo che ha proprietà causali specifiche non riproducibili da un calcolatore. L'IA forte sostiene, in linea teorica, l'idea ipotizzata da Hobbes, ovvero che il pensiero sia un "calcolo" che la mente effettua, quindi un calcolatore corredato dai giusti programmi possiede realmente una mente e non una sua simulazione. Si è cercato di trovare una spiegazione a questo presentando l'esperimento del *cervello nella vasca* che da un punto di vista fisicalista, o più in generale riduzionista, consentirebbe la realizzazione di una macchina cosciente, in quanto ridurrebbe la mente ad una semplice proprietà fisica. Mentre la *sostituzione del cervello* offre uno spunto di riflessione per le teorie funzionaliste.

Oggigiorno abbiamo esempi di sistemi che si comportano *come se* fossero

intelligenti. L'IA ha infatti reso possibile nuove applicazioni come sistemi di riconoscimento vocale, sistemi esperti per la diagnostica, robot e motori di ricerca. Pur avendo ottenuto risultati significativi e avendo realizzato sistemi che possono, in ambiti limitati, competere con l'uomo, siamo ancora lontani dal costruire una macchina con le pretese e le aspettative che si avevano inizialmente. Turing nel suo articolo fu fin troppo ottimista, tuttora non esiste una macchina in grado di eludere un interrogante umano nel gioco dell'imitazione come da lui previsto e tanto meno è mutata l'opinione riguardante le *macchine pensanti*. L'IA, nonostante i suoi enormi progressi, ha ancora molto lavoro da compiere, rimangono infatti valide e attualissime le parole con cui Alan Turing chiuse il suo articolo *Computing Machinery and Intelligence* nel 1950:

> "Possiamo vedere solo una breve distanza davanti a noi, ma vediamo che molto rimane ancora da fare."

Bibliografia

G. Buttazzo. Coscienza artificiale: missione impossibile? *Mondo Digitale*, 2002.

M. Dapor and L.C. Aiello. Intelligenza artificiale: i primi 50 anni. *Mondo Digitale*, 2004.

G.O. Longo. Il test di turing.storia e significato. *Mondo Digitale*, 2009.

J. McCarthy, M.L. Minsky, N. Rochester, and C.E. Shannon. A proposal for the Dartmouth summer research project on artificial intelligence. 1955. URL `http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html`.

H. Putnam. *Reason, Truth, and History*. Cambridge University Press,, 1981. ISBN 0521297761.

S. Russel and P. Norvig. *Intelligenza Artificiale. Un approccio moderno*. Pearson, 2010. ISBN 9788871925936.

J.R. Searle. Minds, brains and programs. *Behavioral and Brain Sciences*, 1980.

A.M. Turing. Computing machinery and intelligence. *Mind*, 1950.