Università
Ca'Foscari
Venezia

Corso di Laurea Magistrale in
Scienze Filosofiche

ordinamento D. M. 270/04

Tesi di Laurea

Siamo macchine?

Differenze e similitudini tra la nostra mente e i computers

Relatore
Ch. Prof. Luigi Perissinotto

Correlatore
Ch. Prof. Enrico Jabara

Laureando
Pietro Gambirasi
Matricola 853746

Anno Accademico
2021 / 2022

INDICE

Introduzione

Ultimamente si sente sempre più parlare di intelligenza artificiale. Nuove tecnologie e applicazioni di programmi intelligenti che promettono di rivoluzionare interi settori economici, scientifici e anche la nostra vita di tutti i giorni riempiono le pagine di giornali e notiziari. Ed è effettivamente indubbio che la cosiddetta "intelligenza artificiale" si stia evolvendo e raggiungendo risultati impensabili fino a qualche anno fa. Veicoli che si guidano da soli, programmi che producono testi quasi indistinguibili da quelli umani e realizzano persino quadri e composizioni musicali, sono al giorno d'oggi una vera e propria realtà. In aggiunta a queste appariscenti ed eclatanti applicazioni dell'intelligenza artificiale, ne esistono di innumerevoli di cui non ci accorgiamo, pur facendone un uso quotidiano. I programmi che ci permettono di comunicare foneticamente con gli assistenti vocali, i suggerimenti e le correzioni automatiche dei messaggi che digitiamo sulle tastiere di computer e telefoni, l'ordine con cui ci vengono proposti contenuti nei social media sono tutti frutto di sistemi di cosiddetta intelligenza artificiale. Tuttavia, dietro questo termine, di cui tutti pensiamo intuitivamente di conoscere il significato, si nascondono in realtà un grandissimo numero di problematiche e di questioni di natura non solo scientifica e tecnologica, ma anche filosofica.

L'idea di costruire delle macchine che pensano, parlano, si comportano e svolgono dei compiti nello stesso modo degli esseri umani non è di certo recente. Il concetto di "automa", ovvero di una macchina dalle fattezze e movimenti umani risale all'epoca degli antichi greci; in epoca moderna pensatori come Cartesio, Pascal e Leibniz hanno ipotizzato e discusso sulla possibilità di costruire delle macchine che simulassero pensieri e comportamenti umani.

Tuttavia, è solo dalla seconda metà del Novecento, con l'avvento dei primi computers digitali, che la realizzazione di programmi che si propongono di simulare alcune capacità del pensiero umano è diventata una vera e propria realtà. A partire da questo momento, i progressi e gli sviluppi nel campo dell'intelligenza artificiale sono sempre stati accompagnati da un prolifico e nutrito dibattito in ambito filosofico in merito non solamente alle potenzialità e alla natura dell'intelligenza artificiale, ma anche alle questioni etiche e morali legate al suo impiego.

In questa tesi ci concentreremo soprattutto sul rapporto tra l'intelligenza artificiale e quella naturale, con l'obiettivo di trovare e sottolineare similitudini e differenze tra le due. A partire da questo confronto si cercherà inoltre di spiegare cosa si intende con il termine "intelligenza artificiale" e in senso più generale cosa significano termini come "intelligenza" o "pensiero".

Nel primo capitolo della tesi ci concentreremo su uno dei principali e più famosi argomenti rivolti contro la possibilità per una macchina di rappresentare una vera e propria riproduzione del pensiero umano, e di conseguenza di possedere una mente e una coscienza al pari della nostra, una posizione nota come "intelligenza artificiale forte". Questo argomento venne formulato dal filosofo inglese John Lucas a partire dai famosissimi Teoremi di Incompletezza di Gödel e si tratta, di conseguenza, di un ragionamento di carattere puramente logico. Verranno sottolineati i punti di forza e le debolezze di questo argomento, prendendo in considerazione le posizioni di alcuni dei suoi principali sostenitori e detrattori.

Nel secondo capitolo ci concentreremo invece sul versante più epistemologico e semantico del tema dell'intelligenza artificiale. Tramite l'interpretazione del pensiero di Ludwig Wittgenstein cercheremo di dare una definizione più precisa di cosa intendiamo con termini come "intelligenza", "pensiero" e "coscienza". Applicheremo poi questi ragionamenti all'intelligenza artificiale per vedere se e fino a che punto possiamo considerare come veramente intelligenti questi sistemi. Verranno inoltre presi in considerazione alcuni dei principali problemi legati alla formalizzazione del cosiddetto "senso comune" e della conoscenza tacita, da sempre uno dei principali ostacoli dell'intelligenza artificiale "classica". Considereremo se questi elementi, indissolubilmente legati alla dimensione esperienziale, sociale e relazionale dell'essere umano, possano essere ritenuti come il fattore di distinzione invalicabile tra la forma di vita umana e quella artificiale delle macchine computazionali.

Nel terzo e ultimo capitolo prenderemo invece in considerazione le più recenti e rivoluzionarie tecnologie di intelligenza artificiale, ovvero le reti neurali, basate su metodi di apprendimento meccanico. Dopo aver brevemente spiegato il loro funzionamento proveremo a vedere se questi ultimi ritrovati tecnologici possano risolvere i tradizionali problemi dell'intelligenza artificiale indicati nel secondo capitolo. Verranno messi in evidenza i punti di forza e le inevitabili limitazioni delle reti neurali, e considereremo se esse siano un altro passo in avanti verso una sempre minore differenza tra l'intelligenza artificiale e quella naturale, o se le due siano ancora, e forse per sempre, fondamentalmente differenti.

CAPITOLO I - GLI ARGOMENTI GÖDELIANI CONTRO L'IA "FORTE"

Tra tutti i teoremi di logica, i due Teoremi di Incompletezza di Gödel sono certamente tra i più celebri, se non i più celebri, anche al di fuori del loro stretto ambito disciplinare, tanto da aver generato un gran numero di dibattiti in merito alle loro possibili conseguenze e implicazioni in settori filosofici e scientifici non direttamente connessi alla logica matematica.

Ora, l'individuare tali possibili implicazioni e conseguenze di un teorema di logica formale in ambiti che con essa hanno poco o nulla a che fare, pone sempre delle importanti problematiche di tipo metodologico e filosofico. Un teorema di logica formale infatti è dimostrato come vero a partire da una serie di ben definiti assiomi tramite altrettanto precise e definite regole di inferenza, dunque la sua validità è indissolubilmente legata alla correttezza formale del sistema nel quale viene dimostrato. Trovare una applicazione "reale" di un teorema di logica formale significa sostenere che, almeno per determinati aspetti, il contesto a cui lo si vuole applicare sia quantomeno perfettamente formalizzabile in termini logici, se non addirittura che la sua stessa essenza sia logica. Si tratta di una questione analoga a quella che viene tradizionalmente posta in relazione al rapporto tra la matematica e la fisica, ovvero se il mondo sia effettivamente "scritto in caratteri matematici", e perciò la sua rappresentazione in termini matematici sia "perfetta", oppure se le espressioni matematiche dei fenomeni fisici siano solamente una molto accurata, ma non esatta (esatta da intendersi come matematicamente perfetta) descrizione di esso. Se dal punto di vista della fisica applicata questo tipo di questione filosofica non comporta rilevanti implicazioni, poiché anche qualora la matematica fornisse semplicemente una approssimazione dell'universo e del suo funzionamento, lo scarto tra realtà e rappresentazione sarebbe talmente piccolo da non causare alcun tipo di problema "pratico", dal punto di vista filosofico e teoretico la questione è invece centrale. Un teorema di logica formale come quello di Gödel si applica solo a determinati tipi di sistemi formali, perciò una sua applicazione rigorosa al di fuori della logica formale non può che avere come prerequisito che il contesto al quale viene applicato possa essere descritto in termini formali. Per questo motivo molte delle presupposte conseguenze filosofiche dei teoremi di Gödel in ambiti quali la politica o le scienze sociali pongono più di qualche interrogativo sulla loro legittimità.

Una delle più convincenti e appropriate applicazioni del teorema di Gödel in ambiti esterni alla logica formale è sicuramente il cosiddetto "argomento gödeliano contro l'"ia forte"

proposto per la prima volta negli anni ’60 del Novecento dal filosofo John Lucas, e che ha generato un lungo e animato dibattito tra sostenitori e detrattori di questa determinata concezione di intelligenza artificiale.

Prima di entrare nel dettaglio di questo dibattito, è opportuna una breve esposizione, anche se esclusivamente discorsiva e non formale, dei teoremi di incompletezza di Gödel e delle loro conseguenze, e di alcune definizioni e considerazioni sui concetti di “macchina” e “intelligenza artificiale”.

1.1 Breve esposizione informale dei teoremi di incompletezza

Come già detto, i teoremi di incompletezza di Gödel sono tra i più famosi e conosciuti teoremi di logica, tanto che sono stati spesso usati al di fuori dell’ambito logico per “dimostrare” le teorie più disparate, spesso con risultati paradossali. Ma da cosa deriva questa fama? La loro interpretazione più comune e superficiale li vede come una sorta di prova matematica della incoerenza della matematica stessa, un definitivo affossamento della convinzione positivistica che la matematica sia il regno della verità incontrastabile e assoluta, la caduta del più solido baluardo della razionalità umana. È senza dubbio evidente il fascino che un tale argomento può produrre: il venir meno dei fondamenti della matematica ad opera della matematica stessa. Ma è davvero questo che i teoremi di incompletezza di Gödel dimostrano?

Gödel giunse alla dimostrazione dei teoremi di incompletezza nel 1931 lavorando sul secondo dei ventitré problemi matematici insoluti stilati da Hilbert all’inizio del ventesimo secolo, ovvero dimostrare tramite metodi finitari la coerenza dell’insieme degli assiomi dell’aritmetica[1]. I risultati a cui Gödel pervenne furono non solo sorprendenti ma secondo alcuni[2] minarono alle fondamenta lo stesso programma di Hilbert.

Cosa affermano dunque questi teoremi? Il primo dei due teoremi di incompletezza può essere espresso in linguaggio non formale in questo modo:

> Qualunque sistema formale **S** che sia coerente e in cui una certa quantità di aritmetica possa essere svolta è incompleto; ovvero esiste almeno un enunciato nel linguaggio del sistema che non può essere né provata né confutata all’interno del sistema stesso.

[1] Uno degli obbiettivi di Hilbert era quello di provare la coerenza di teorie matematiche fondamentali sulla base di assunzioni deboli e senza supporre l’esistenza di insiemi infiniti.

[2] “Thus today I am of the opinion that 1. Gödel has shown the unrealizability of Hilbert’s program. 2. There is no more reason to reject intuitionism” (John Von Neumann)

Con “sistema formale” si intende un insieme di assiomi dotato di regole di inferenza attraverso le quali è possibile formulare teoremi, ma a differenza dei semplici sistemi assiomatici il suo linguaggio è formalizzato e privo di concetti come verità e significato. All’interno di un tale sistema la deduzione di enunciati diventa un procedimento puramente meccanico, ossia la trasformazione, tramite le regole di inferenza, delle stringhe di simboli che ne costituiscono le formule. Per “completezza” si intende che per ogni enunciato di un dato sistema formale o l’enunciato stesso o la sua negazione possono essere derivati nel sistema. Un sistema formale è definito coerente se non si dà il caso che al suo interno ci sia un enunciato tale che possano essere provati, all’interno di quel sistema formale, sia l’enunciato stesso che la sua negazione. La coerenza è requisito necessario per i teoremi di Gödel poiché in logica da un sistema incoerente può essere dedotto qualsiasi enunciato (*ex falso quodlibet*), di conseguenza tutti i sistemi formali incoerenti sono anche trivialmente completi.

Il secondo teorema di incompletezza di Gödel afferma che:

> Per ogni sistema formale **S** che sia coerente e in cui una certa quantità di aritmetica possa essere svolta, la coerenza di **S** non può essere provata in **S**.

La “certa quantità di aritmetica” che viene richiesta in entrambi i teoremi è l’aritmetica necessaria a rappresentare sotto forma di numeri gli assiomi, gli enunciati, i teoremi e le dimostrazioni di un sistema formale, e ad esprimere proposizioni e asserzioni riguardanti questi stessi elementi del sistema sotto forma di formule aritmetiche riguardanti i numeri a loro corrispondenti. Questa operazione di aritmetizzazione della sintassi del linguaggio del sistema formale prende il nome di gödelizzazione o numerazione di Gödel. In questo modo la parte aritmetica di un sistema formale **S** contiene enunciati aritmetici che possiamo interpretare come riferiti a dimostrazioni e teoremi propri dei sistemi formali, **S** incluso.

È opportuno aggiungere qualche precisazione riguardante questi teoremi in modo da confutare alcune delle loro interpretazioni più superficiali e inesatte. Il primo teorema di incompletezza non esclude in nessun modo che esistano delle verità assolutamente non dimostrabili; il teorema non riguarda né la nozione di verità semantica né afferma l’impossibilità assoluta di indimostrabilità di una proposizione. Quello che il teorema afferma è che in particolari sistemi formali esistono enunciati non derivabili sintatticamente dagli assiomi e dalle regole di inferenza propri del sistema stesso, tali enunciati possono tuttavia essere tranquillamente dimostrabili in altri sistemi formali. Una conseguenza invece molto

significativa del primo teorema di incompletezza è che prendendo in esame sistemi formali estremamente potenti, come ad esempio la teoria insiemistica ZFC da cui è possibile derivare gran parte della matematica, sarà sempre possibile trovare almeno un enunciato aritmetico non dimostrabile al suo interno. In questo senso si può quindi affermare che esistono verità aritmetiche non dimostrabili tramite gli attuali metodi e assiomi matematici. Questa conseguenza del primo teorema di incompletezza è di centrale importanza per le argomentazioni di ispirazione gödeliana sul tema dell'intelligenza artificiale e verrà di conseguenza ripresa e approfondita in seguito.

Il fatto che il secondo teorema di incompletezza affermi che sistemi formali come l'aritmetica di Peano o la teoria degli insiemi ZFC non possano provare la loro stessa coerenza ha generato una serie di argomenti di carattere scettico nei confronti della matematica: come possiamo essere sicuri della coerenza di un sistema formale che non riesca a provare la sua stessa coerenza? A questo riguardo è necessario fare alcune precisazioni. Innanzitutto, il secondo teorema di incompletezza non afferma in nessun modo che la coerenza di tali sistemi formali non possa essere dimostrata in assoluto, essa può infatti essere tranquillamente dimostrata all'interno di altri sistemi formali. Ovviamente, se proviamo la coerenza di un determinato sistema formale **S** all'interno di un altro sistema formale **T** dobbiamo avere fiducia nella coerenza di quest'ultimo, coerenza che tuttavia non potrà essere dimostrata all'interno di **T**, generando una sorta regresso all'infinito da cui sembra difficile uscire.

A questo riguardo ci si potrebbe inoltre chiedere che valore possa avere una prova della coerenza di un sistema formale all'interno del sistema formale in questione, ovvero, se si nutrono dei dubbi nei confronti della coerenza di un sistema formale, che valore può avere una prova della sua coerenza ottenuta in esso? Da un sistema formale incoerente è possibile dimostrare qualsiasi proposizione, dunque, in un'ottica scettica, una prova della sua coerenza avrebbe valore solamente se si avesse la certezza della sua coerenza, ovvero di ciò di cui si dubita. In questo senso il secondo teorema di incompletezza non sembra avvalorare posizioni scettiche riguardanti la coerenza delle teorie matematiche.

In secondo luogo, come già affermato, i teoremi di Gödel riguardano le nozioni di derivabilità e di provabilità all'interno di un sistema formale e non dicono nulla rispetto al valore semantico di verità di una proposizione; ovvero, il fatto che la coerenza di un sistema formale non possa essere sintatticamente derivabile all'interno di esso non esclude che essa possa essere dimostrata come vera tramite metodi e ragionamenti non formali. Questa distinzione tra derivabilità formale e verità semantica, come vedremo in seguito, è di

fondamentale importanza per le interpretazioni che considerano i teoremi di incompletezza come la prova della impossibilità di una totale equivalenza tra la mente umana e una macchina.

1.2 Macchine computazionali e Intelligenze artificiali

Prima di entrare nel vivo della discussione intorno alle implicazioni dei teoremi di incompletezza nel campo dell'intelligenza artificiale è necessario dare una precisa definizione di cosa si intenda per concetti come "intelligenza artificiale" o "macchina computazionale" e indicare se e per quale motivo sia possibile applicare alle macchine dei teoremi di logica formale. Abbiamo infatti già indicato come sia sempre problematico applicare tali teoremi in contesti estranei alla logica, tuttavia, nel caso delle macchine, non solo tale applicazione sembra essere giustificata, ma, almeno sul piano teorico, sussiste una equivalenza tra un certo tipo di macchine e i sistemi formali.

1.2.1 Macchine computazionali ideali e macchine computazionali reali

Le macchine a cui si riferiscono gli argomenti di ispirazione gödeliana che verranno discussi in seguito sono una particolare tipologia di macchine dette macchine computazionali (computers). Sebbene questo tipo di macchine possa essere costruito nei modi più disparati, da meccanismi analogici ad ingranaggi fino ai moderni computers digitali, esse sono tutte riconducibili ad un modello ideale di macchina calcolatrice ideato da Alan Turing e che prende appunto il nome di "macchina di Turing".

Come nel caso di Gödel e dei suoi teoremi di incompletezza, anche Turing giunse al concetto di macchina di Turing lavorando ad un problema matematico posto da Hilbert[3]. Il problema in questione riguardava il cosiddetto Entscheidungsproblem (problema della computabilità), ovvero se esista la possibilità di provare per qualsiasi enunciato di logica del primo ordine la sua derivabilità all'interno di tale logica tramite una procedura puramente meccanica. Per provare che non esiste alcuna procedura computazionale che riesca a stabilire ciò per ogni enunciato[4], Turing introdusse il concetto di una macchina ideale[5] composta da un nastro di lunghezza infinita diviso in caselle di grandezza uguale, ognuna delle quali contenente

[3] Il problema in questione, denominato *Entscheidungsproblem*, venne posto nel 1928 nel corso di una conferenza internazionale da David Hilbert e Wilhelm Ackermann.
[4] Nel 1936 sia Alan Turing che Alonzo Church dimostrarono indipendentemente e con metodi diversi (poi dimostrati equivalenti tra loro) che una soluzione generale a tale problema sia impossibile.
[5] Turing pubblicò i suoi risultati in un articolo del 1936 intitolato: *On computable numbers, with an application to the Entscheidungsproblem*.

un simbolo appartenente ad un insieme finito di caratteri. Su questo nastro opera una testina capace di leggere il simbolo presente in una casella, cancellarlo, scriverne un altro al suo posto, spostarsi di una casella verso destra o verso sinistra oppure rimanere nella casella attuale. La testina si muove sulla base di una serie di istruzioni chiamate "programma" che indicano: lo stato iniziale della macchina, il contenuto della casella che sta leggendo, il nuovo contenuto della casella una volta che sia stato riscritto, se spostarsi di una casella verso sinistra, verso destra o se rimanere nella stessa casella e infine il prossimo stato della macchina. A seconda del diverso tipo di programma che opera la macchina, essa può svolgere un gran numero di funzioni matematiche, tra cui moltiplicazione, addizione, sottrazione elevamenti a potenza.

Una macchina ideale di questo tipo è superiore a qualsiasi macchina reale costruibile per diversi motivi: il nastro su cui scrive (ovvero la sua memoria) è infinita, per cui la sua capacità di calcolo non è limitata dalla sua capacità di memoria; non è soggetta a limiti temporali, dunque può eseguire operazioni che ad una macchina reale richiederebbero tempi talmente lunghi da renderne impossibile il completamento; non soffre di alcuna problematica strutturale o difetto di programmazione, per cui i suoi risultati sono sempre credibili e mai affetti da errori. Queste caratteristiche permettono alla macchina di Turing di portare a termine il calcolo di qualsiasi funzione computabile, senza incorrere in limitazioni di spazio di archiviazione, di tempo o di fallire per colpa di malfunzionamenti del meccanismo o di errori di programmazione. Nella realtà, invece, possono esistere delle funzioni computabili che nessuna macchina reale potrà mai calcolare in quanto richiederebbero per esempio un supporto fisico di memoria composto da più materia di quella presente nell'universo stesso.

Si definisce "macchina di Turing universale" una macchina di Turing capace di calcolare tutto ciò che una qualsiasi macchina di Turing particolare riesce a calcolare; di conseguenza, da ciò segue che: ogni problema non calcolabile da una macchina di Turing universale è incalcolabile in senso assoluto, e che ogni problema potenzialmente calcolabile può essere calcolato dalla macchina di Turing universale[6].

[6] In modo più formale si può affermare che una funzione è calcolabile tramite un procedimento meccanico se e solo se è calcolabile da una macchina di Turing. Questa identificazione tra calcolabilità effettiva e calcolabilità da parte di una macchina di Turing prende il nome di "Tesi di Church-Turing"; Alonzo Church nel 1936 propose indipendentemente da Turing una definizione formale di algoritmo chiamata "λ-calcolo" che venne in seguito dimostrata equivalente alla nozione di computabilità fornita da Turing.

1.2.2 Macchine di Turing e sistemi formali

Una volta specificato cosa si intende per "macchina di Turing" è ora necessario illustrare in che modo una tale macchina sia correlata ai sistemi formali, e di conseguenza ai teoremi di incompletezza di Gödel. Per far ciò è necessario introdurre qualche altra definizione, ossia quella di algoritmo, di enumerabilità e di decidibilità.

Un algoritmo è una procedura meccanica[7] che applicata ad una stringa di simboli (non necessariamente ai soli numeri) termina dopo una serie finita di passaggi dando un'altra stringa di simboli come risultato. Un insieme di stringhe di simboli si dice "computabilmente enumerabile" se esiste un algoritmo capace di elencare tutti i suoi membri, senza tenere in considerazioni limitazioni di spazio e tempo di calcolo. In questo contesto il concetto di macchina di Turing risulta estremamente utile poiché, dato che, come abbiamo visto, essa è in grado di calcolare qualsiasi cosa sia potenzialmente computabile, è possibile programmarla in modo da svolgere un tale algoritmo senza dover sottostare alle limitazioni fisiche di esseri umani e calcolatori reali.

Un insieme *S* si dice "computabilmente decidibile" se esiste un algoritmo che sia in grado di dire per un qualunque elemento *x* se esso appartenga o meno a *S*. Abbiamo in precedenza definito che un sistema formale è formato da un linguaggio formale composto da simboli ben definiti e da un insieme ben specificato di assiomi e di regole di inferenza. In aggiunta a ciò, specifichiamo ora che deve essere possibile stabilire tramite una procedura meccanica se una determinata sequenza di simboli sia un enunciato del linguaggio, se un determinato enunciato sia un assioma del linguaggio e se un certo enunciato segua direttamente da altri enunciati del sistema tramite le regole di inferenza. Abbiamo definito la gödelizzazione o numerazione di Gödel come un procedimento di aritmetizzazione della sintassi di un sistema formale *S* che permette di assegnare in modo univoco ad ogni simbolo, formula e sequenza di formule di *S* un numero intero naturale. In questo modo è possibile non solo assegnare ad ogni dimostrazione[8] un numero naturale che possiamo determinare essere lo specifico numero di Gödel di quella dimostrazione in *S*, ma possiamo anche estrarre il numero che corrisponde all'ultimo enunciato di tale dimostrazione, ossia il teorema che viene dimostrato. Da ciò segue che è possibile scorrere meccanicamente ogni numero intero e vedere se esso corrisponde al numero di una

[7] Per "meccanica" si intende una procedura che non richieda alcun tipo di interpretazione o di decisione, ma che consista nella sola applicazione di una serie di ben definite regole, o, in altre parole, che possa essere eseguita da un computer.

[8] Sequenza di formule in un sistema formale *S* tale che ogni formula o è un assioma di *S* o è ottenuta mediante l'applicazione delle regole di inferenza a formule precedenti.

dimostrazione in *S* e, in caso corrisponda, è possibile ricavare da esso il numero del teorema provato da quella dimostrazione. Detto in termini più formali, l'insieme dei teoremi di un sistema formale è computabilmente enumerabile, o ancora "with each effectively given formal system is associated a Turing machine *M* which enumerates the set of theorems of *S*, or -more picturesquely- prints out the theorems of *S* one ofter another"[9]. Per converso si può immaginare che, dato un linguaggio formale *L*, si possa programmare una macchina di Turing *M* che produca tutti i numeri di Gödel delle formule costruibili in *L* e considerare la loro chiusura deduttiva[10] come l'insieme dei teoremi di un sistema formale *S* in *L*. In questo modo si riesce a far corrispondere una macchina di Turing *M* ad un sistema formale *S*. O ancora, più informalmente, "per ogni sistema formale è possibile programmare un computer in modo che produca meccanicamente tutti e soli i teoremi del sistema; viceversa, per un qualsiasi modo di programmare un computer in modo da operare come un «produttore meccanico» di teoremi, esiste un sistema formale che ha per teoremi tutte e sole le formule prodotte dal computer"[11]. In questo modo "talk of well-defined or effectively given formal systems can be converted into talk of Turing machines and vice versa"[12].

Questa equivalenza rende dunque possibile l'applicazione dei teoremi di incompletezza alle macchine di Turing; tuttavia, come abbiamo visto, le macchine di Turing sono concetti puramente ideali, completamente diverse sia per funzionamento che per capacità di calcolo rispetto alle macchine reali, è dunque necessario verificare se e in che modo sia possibile applicare i teoremi di incompletezza anche ai computers reali. Ora, come abbiamo visto, l'equivalenza tra un sistema formale *S* e una macchina di Turing *M* è data dal fatto che è possibile programmare quest'ultima in modo che produca tutti i teoremi di *S*, e che per ogni programma che generi enunciati in un determinato linguaggio formale, esiste un sistema formale che ha questi enunciati tra i suoi teoremi. È possibile programmare un computer reale in modo da svolgere questo tipo di programma? È possibile, ma con alcune limitazioni. In primo luogo, data la intrinseca finitezza della capacità di memoria e del tempo di elaborazione a cui è soggetto qualsiasi computer reale costruito o realizzabile, non tutti i sistemi formali possono essere equiparati ad una macchina reale, ce ne saranno infatti alcuni talmente grandi e complessi da risultare assolutamente incalcolabili per un qualsiasi calcolatore fisico. In secondo luogo, si

[9] Solomon Feferman, *Are there absolutely unsolvable problems? Gödel's Dichotomy*, *Philosophia Mathematica,* Volume 14, Issue 2, giugno 2006, p. 138
[10] Insieme delle formule che possono essere ricavate a partire dagli assiomi di un sistema formale.
[11] Francesco Berto, *Tutti pazzi per Gödel!*, Bari, Laterza, 2008, p.207
[12] Solomon Feferman, *Are there absolutely unsolvable problems? Gödel's Dichotomy*, *Philosophia Mathematica*, Volume 14, Issue 2, giugno 2006, p. 138

dovrà supporre che il computer "associato" ad un determinato sistema formale non sia soggetto a nessun difetto di realizzazione nella sua parte materiale e a nessun errore di programmazione nel suo algoritmo tali da produrre errori nei risultati da esso prodotti. A queste condizioni è lecito estendere l'equivalenza tra sistemi formali e macchine di Turing anche alle macchine calcolatrici reali.

1.3 Intelligenza Artificiale

Come ultimo passo introduttivo alla discussione degli argomenti gödeliani riguardanti l'intelligenza artificiale è necessario indicare cosa si intenda per "intelligenza artificiale". Dare una definizione precisa e univoca del concetto di intelligenza artificiale è un'impresa estremamente difficile se non impossibile, data la vastità di approcci possibili al suo studio e la varietà del suo impiego in molteplici contesti sia pratici che teoretici. Proprio per questo motivo non esiste una sola disciplina che si occupi di studiare l'intelligenza artificiale, ma moltissime e diverse branche della scienza e della filosofia svolgono ricerche su di essa. Ora sarebbe impossibile elencare con precisione tutti i campi in cui l'intelligenza artificiale trova impiego pratico e teorico, perciò ci si limiterà ad indicare alcuni dei più significativi: in logica e matematica l'intelligenza artificiale trova impiego nello studio della computabilità e del ragionamento formale; in psicologia nello studio del modo di pensare e comportarsi degli esseri umani; nelle neuroscienze nello studio del modo in cui il cervello processa le informazioni; in economia nello studio delle decisioni e delle previsioni dei guadagni; infine in filosofia essa presenta importanti implicazioni in diversi settori quali l'epistemologia, il linguaggio, la morale e la logica. A livello pratico, oggigiorno l'intelligenza artificiale trova applicazioni in un sempre crescente numero di contesti, tra i quali: l'analisi di dati e informazioni, il riconoscimento di immagini, la guida di veicoli autonomi, il riconoscimento vocale, la traduzione di testi e addirittura le diagnosi mediche. Una tale eterogeneità di applicazioni e approcci al suo studio rende senza dubbio ardua l'impresa di dare una definizione di intelligenza artificiale che riesca a comprendere tutte le sue sfaccettature e caratteristiche.

Da un punto di vista storico si può datare il primo utilizzo del termine "Artificial Intelligence" (spesso semplicemente abbreviato in "AI") al 1956, anno in cui si tenne al Dartmouth College di Hanover, nel New Hampshire, una piccola conferenza a cui parteciparono una decina di logici e matematici e nel corso della quale vennero gettate alcune delle basi per i successivi sviluppi dello studio sull'intelligenza artificiale. Sebbene il termine venne coniato nel 1956, ricerche e pubblicazioni sul campo dell'intelligenza artificiale erano presenti già da

alcuni anni. Basti pensare al famosissimo articolo *Computing machinery and intelligence* pubblicato da Alan Turing sulla rivista *Mind* nel 1950[13], nel quale, per provare a rispondere all'interrogativo se sia possibile per una macchina pensare, Turing introdusse l'"imitation game", che sarà poi generalmente conosciuto come il "test di Turing". Conscio della difficoltà insita nel dare una definizione formale di "pensiero", Turing propose di stabilire la capacità di una macchina di pensare non sul piano teoretico, ma sulla base di un esperimento pratico, nel corso del quale un intervistatore umano, ponendo delle domande, deve riuscire a distinguere le risposte date da una macchina da quelle date da un altro essere umano. Se per almeno la metà delle risposte la macchina riesce a ingannare l'intervistatore il test viene considerato superato. Secondo Turing il superamento da parte di una macchina di questo test è condizione sufficiente per poter considerarla intelligente, mentre, come vedremo, secondo molti altri autori, tra i quali John Searle[14], non è possibile attribuire l'intelligenza ad una macchina sulla base di ciò.

Ora, l'idea che possano essere costruite macchine che imitino i comportamenti e le risposte umane è presente in ambito filosofico e letterario da ben prima del ventesimo secolo. Il concetto stesso di automa, ossia di un costrutto meccanico dalle sembianze umane dotato di autonomia di movimento e azione, ha le sue radici nella tradizione mitologica greca, mentre riferimenti all'automa Talos (creatura bronzea dotata di vita) si possono trovare in uno scolio alla *Repubblica* di Platone attribuito a Simonide[15]. Il più celebre brano filosofico, antecedente il ventesimo secolo, riguardante le macchine (intese come repliche artificiali degli esseri umani) si trova nel *Discorso sul Metodo* di Cartesio, il quale fu il primo a porre una fondamentale e assoluta impossibilità per una qualsiasi macchina di imitare la ragione umana. Tramite una sorta di Test di Turing ante litteram, egli individuò due caratteristiche che rendono sempre possibile la distinzione tra gli esseri umani e le macchine: "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: poiché si può ben immaginare una macchina che profferisca delle parole, e anzi ne profferisca alcune riguardanti azioni corporali che producano qualche alterazione nei suoi organi, come domandare qualcosa, se toccata in una parte, o gridare che le si fa male se toccata in altra parte, e simili cose; ma non già che essa disponga le parole diversamente per rispondere a tono a tutto quello che uno può dirle"[16] e "il secondo mezzo è che, anche se facessero alcune cose ugualmente bene e anzi meglio di noi, esse inevitabilmente sbaglierebbero in alcune altre, e si

[13] Alan Turing, *Computing machinery and intelligence*, in *Mind* LIX, no. 2236, ottobre 1950, pp. 433-60
[14] Cfr. John Searle, *Minds, Brains, Programs*, "Behavioral and Brain Sciences", 3 (1980), pp. 417-457
[15] Cfr. Gianni Micheli, *Il concetto di automa nella cultura greca dalle origini al sec. IV A.C.*, "Rivista della Storia della Filosofia", vol. 53, no. 3, 1998, editore Franco Angeli, pp. 421-462
[16] Cartesio, *Discorso sul metodo*, trad. di Armando Carlini, Bari, Laterza, 1965, p. 165

scoprirebbe così che non agiscono per conoscenza, ma solo per una disposizione dei loro organi"[17]. Interessante notare che questi due argomenti siano tuttora validi. Il primo, infatti, può essere informalmente considerato equivalente ad un Test di Turing, e al giorno d'oggi nessuna macchina è ancora riuscita a superarlo, nonostante la previsione dello stesso Turing che riteneva che ciò sarebbe stato possibile già a partire dai primi anni del 2000[18], mentre il secondo descrive alla perfezione l'attuale stato dell'avanzamento tecnologico nel campo dell'intelligenza artificiale. Sebbene infatti ormai esistano macchine che svolgono i più disparati compiti in modo molto più efficiente e preciso di un qualunque essere umano (dall'analisi di enormi quantità di dati al giocare a scacchi), nessuna di esse non solo non riesce a svolgere tutte le semplici attività che un normale essere umano svolge ogni giorno, ma non è nemmeno in grado di portare a termine compiti computazionali che vadano oltre la sua specializzazione (ad esempio, Deep Blue, il famoso computer che per primo sconfisse a scacchi Gary Kasparov, uno dei più grandi giocatori di tutti i tempi, non avrebbe potuto tradurre un testo scritto, come d'altra parte un qualunque programma di traduzione non possiede alcuna capacità di giocare a scacchi[19]). Da questi rapidi cenni storici risulta evidente che concetti assimilabili a quello che intuitivamente si intende con "intelligenza artificiale" sono presenti nella cultura occidentale da ben prima che venissero concepiti e costruiti i moderni computer digitali, ovvero il supporto fisico che ha permesso l'effettiva realizzazione di macchine che secondo alcuni aspetti possono essere definite intelligenti.

A questo punto è dunque necessario provare a dare una definizione più generale e precisa di "intelligenza artificiale" che riesca a includere quanto più possibile le sue fondazioni teoriche e le sue applicazioni pratiche. Ora a questo proposito, in aggiunta alla eterogeneità dalle discipline che si occupano di intelligenza artificiale e alla trasversalità e ramificazione delle sue implicazioni in innumerevoli settori, c'è un ulteriore elemento da tenere in considerazione: ossia che esistono due diverse e incompatibili concezioni riguardanti le possibilità e gli obbiettivi dell'intelligenza artificiale sotto le quali si possono raggruppare la maggior parte delle posizioni filosofiche e scientifiche.

Secondo i sostenitori della cosiddetta "IA forte" è possibile la realizzazione di intelligenze artificiali che riproducano in tutto e per tutto il funzionamento di una mente umana e che quindi

[17] *Ibidem*

[18] Va fatto notare che il superamento del Test di Turing non è assolutamente uno degli obbiettivi principali delle ricerche odierne sull'IA, anzi è da molti considerato come un obiettivo irrilevante sul piano pratico.

[19] Anche in questo caso è doverosa la precisazione che non c'è alcun vantaggio pratico nel progettare un computer che sappia svolgere anche compiti che vanno al di là della sua specializzazione. In ogni caso, allo stato attuale della tecnologia, un'intelligenza artificiale che abbia anche solo le capacità intellettive generali di un bambino di pochi anni è ancora irrealizzabile.

siano dotate di coscienza e che riescano ad operare non solo sul piano della semplice sintassi del loro linguaggio di programmazione ma possiedano anche il concetto di significato. Questa tesi è fortemente contestata, specialmente in ambito filosofico, da coloro che invece considerano impossibile, non solo sul piano della reale ed effettiva costruzione di un tale tipo di macchine, ma anche e soprattutto sul piano teoretico, riprodurre artificialmente la coscienza. Questo secondo tipo di posizione viene generalmente definito come "IA debole". I sostenitori dell'IA debole affermano che al massimo una macchina potrà sembrare di possedere le stesse facoltà intellettive di una mente umana, ovvero sarà in grado di superare un test di Turing[20], ma non potrà mai avere una forma di coscienza paragonabile a quella degli esseri umani.

Vista questa divisione, al momento ancora aperta (non sono ancora stati formulati argomenti ritenuti completamente convincenti né per l'una né per l'altra ipotesi), si daranno almeno due diverse formulazioni della definizione di "intelligenza artificiale". In una prospettiva di IA forte si potrebbe definire l'intelligenza artificiale come il "campo che mira a costruire e a studiare sistemi artificiali che sappiano pensare come un essere umano", mentre per un sostenitore dell'IA debole, viste le molteplici difficoltà che si incontrano anche solamente nel definire cosa siano il pensiero e la coscienza in ambito umano, l'intelligenza artificiale si dovrebbe limitare allo studio e alla realizzazione di sistemi che agiscano come un essere umano.

Ora, se da un lato una concezione di intelligenza artificiale modellata sulla mente e sul comportamento umano può senza dubbio essere di estremo interesse per discipline come la filosofia, la psicologia e le scienze cognitive, dall'altro, dal punto di vista delle applicazioni pratiche in cui maggiormente trovano impiego sistemi di intelligenza artificiale, un'intelligenza artificiale che pensi o che si comporti come un essere umano non è solamente inefficiente ma anche potenzialmente dannosa (basti pensare ai problemi che causerebbe un veicolo a guida autonoma che, proprio come un qualsiasi essere umano, di tanto in tanto sbagliasse strada o si distraesse). Da questo punto di vista quindi l'obiettivo delle ricerche non è tanto quello di creare un'intelligenza artificiale "umana", quanto piuttosto di realizzare sistemi che pensino in maniera razionale (prospettiva forte) o si comportino come degli agenti razionali (prospettiva debole).

[20] La teoria dell'IA debole ammette che sia teoricamente possibile per una macchina superare anche il cosiddetto "test di Turing totale" (un test di Turing non solo linguistico ma che preveda anche una componente fisica), ossia si presuppone che una macchina possa essere indistinguibile da un essere umano anche nelle fattezze, nelle azioni e nei comportamenti. Ovviamente, parimenti al test di Turing linguistico, anche un eventuale superamento di questo tipo di test non è condizione sufficiente per poter attribuire ad una macchina una qualche forma di coscienza.

Ma cosa si intende per "pensare in modo razionale" e per "agente razionale"? In questo contesto per pensiero razionale si intende un ragionamento che porti al conseguimento di un obbiettivo nel modo matematicamente più corretto ed efficiente (ad esempio impiegando il numero minore di risorse possibili, nel minor tempo possibile etc.), e per agente razionale ciò che concretamente mette in atto il pensiero razionale, e che dunque agisce in vista del raggiungimento di uno scopo secondo regole logicamente e matematicamente esatte.

Ora, anche questo approccio "razionale" all'intelligenza artificiale comporta una serie di problematiche, soprattutto di carattere morale. I mezzi che un agente razionale decide di impiegare per raggiungere nel modo più efficiente il suo obbiettivo infatti essere moralmente inaccettabili o comportare conseguenze e danni collaterali imprevisti dagli stessi esseri umani che lo hanno programmato e realizzato, rendendo quindi necessari sia un'estrema cautela nell'impiego di tali intelligenze artificiali, che delicati quesiti di carattere morale che riguardino quali e quanti "danni collaterali" siano accettabili in nome del raggiungimento dello scopo che abbiamo affidato ad esse. Inoltre, va ulteriormente specificato che una perfetta razionalità è nella maggior parte dei casi impossibile da raggiungere, sia perché la realtà offre scenari molto complessi a cui è molto difficile applicare delle regole logiche certe, sia perché, come abbiamo già visto, le macchine reali presentano una serie di limitazioni intrinseche e insuperabili di capacità e velocità di calcolo che impediscono nel concreto una perfetta analisi dei sistemi. Di conseguenza è necessario usare strumenti di calcolo probabilistici, che nonostante possano risultare molto efficaci e precisi, sono comunque lontani dalla teoretica e ideale perfetta razionalità.

Queste quattro diverse concezioni dell'intelligenza artificiale e dei suoi scopi, sotto cui si possono raggruppare la maggior parte delle posizioni e degli studi in questo campo, non devono essere concepite come rigidamente separate e incompatibili tra loro, anzi molto spesso è vitale nella ricerca lavorare sia sulla componente "pensate" che su quella agente. Il problema della differenza tra la concezione "forte" dell'intelligenza artificiale e quella "debole" ha infatti una rilevanza principalmente filosofica e nella maggior parte dei casi ha scarsa importanza dal versante pratico e nelle applicazioni reali dei sistemi di intelligenza artificiale.

1.4 Gli argomenti gödeliani contro l'intelligenza artificiale forte

È ora giunto il momento di esporre nello specifico i cosiddetti "argomenti gödeliani" contro l'IA forte. Sotto questa definizione viene solitamente raggruppata una serie di argomentazioni di carattere filosofico e logico-matematico che, utilizzando alcune implicazioni

dei teoremi di incompletezza, mirano a stabilire, nelle loro versioni più deboli, una differenza e incompatibilità di fondo tra una mente umana e una macchina computazionale, e in quelle più forti una vera e propria superiorità della prima sulla seconda. Se il più celebre, discusso e contestato di questi argomenti è senza dubbio quello proposto da John Lucas nel suo articolo del 1961 *Minds, Machines and Gödel*[21], quello che può essere considerato il primo in termini cronologici fu posto dallo stesso Kurt Gödel nel corso di una conferenza dal titolo "Some Basic Theorems on the Foundations of Mathematics and Their Implications"[22]. Tuttavia, come vedremo, a differenza di Lucas, l'obiettivo principale di Gödel non era tanto quello di attaccare una certa concezione di intelligenza artificiale, quanto piuttosto di sfruttare la possibile superiorità della mente umana sulle macchine computazionali come sostegno alla sua concezione platonistica della matematica e degli enti matematici[23]. Oltre ad una seconda formulazione del suo stesso argomento gödeliano da parte di Lucas[24], scritta per rispondere alle molte critiche ricevute al suo articolo del 1961, un altro celebre argomento di questo tipo è quello proposto dal fisico e matematico inglese Roger Penrose nel suo libro *The Emperor's New Mind*[25] e riproposto e perfezionato nel seguente *Shadows of the Mind*[26]. Questa seconda formulazione dell'argomento gödeliano di Penrose è forse la più forte e completa delle due, anche se rimane comunque aspramente criticata da studiosi come Per Lindstrom[27] e Stewart Shapiro[28].

1.4.1 Kurt Gödel: Gibbs Lecture

Come già accennato precedentemente, Gödel fu il primo a intuire come possibile conseguenza dei suoi teoremi di incompletezza una differenza tra le potenzialità di una mente umana e di una macchina computazionale. Questa ipotesi venne da lui proposta nel corso di una

[21] J. R. Lucas, *Minds, Machines and Gödel*, "Philosophy", XXXVI (1961), pp. 112-127
[22] Gödel tenne questa conferenza il 26 dicembre del 1951 nell'ambito di un ciclo di conferenze dedicate al matematico J. W. Gibbs e ospitate dalla Brown University di Providence, ed è perciò generalmente conosciuta come *Gibbs Lecture*.
[23] Usando le parole dello stesso Gödel, gli enti e i concetti matematici "form an objective reality of their own, which we cannot create or change, but only perceive and describe". Kurt Gödel, *Collected Works, Vol III: Unpublished Essays and Lectures*, a cura di S. Feferman et al., Oxford University Press, New York, Oxford, 1995, pp. 320
[24] J.R. Lucas, *Minds, Machines and Gödel: a Retrospect*, in *Machines and Thought*, a cura di Peter Millican e Andy Clark, Oxford University Press, 1996, pp. 103-124
[25] Roger Penrose, *The Emperor's New Mind*, Oxford, Oxford University Press, 1989
[26] Roger Penrose, *Shadows of the Mind*, Oxford, Oxford University Press, 1994
[27] Cfr. Per Lindstrom, *Penrose's New Argument,* "Journal of Philosophical Logic", XXX (2001), pp. 241-250, e *Remarks on Penrose's New Argument,* "Journal of Philosophical Logic", XXXV (2006), pp. 231-237
[28] Cfr. Stewart Shapiro, *Mechanism, Truth and Penrose's New Argument*, "Journal of Philosophical Logic", XXXII (2003), pp. 19-42

conferenza, pubblicata solamente postuma[29], dal titolo "Some Basic Theorems on the Foundations of Mathematics and Their Implications", che aveva appunto come oggetto alcune conseguenze dei suoi due teoremi di incompletezza nell'ambito della matematica, specialmente in riferimento a ciò che può essere matematicamente provato e a possibili limitazioni della mente umana nello scoprire le verità matematiche. Lo scopo primario di Gödel non era quindi tanto quello di trarre come conseguenza dei suoi teoremi di incompletezza un argomento che confutasse la possibilità di equiparazione tra la mente umana e una macchina computazionale, quanto quello di usare questa possibile differenza tra le due come sostegno per una certa concezione filosofica della matematica stessa.

In particolare, Gödel sottolinea come, specialmente dal suo secondo teorema di incompletezza, si deduce che è impossibile che:

> Someone should set up a certain well-defined system of axioms and rules and consistently make the following assertion about it: All of these axioms and rules I perceive (with mathematical certitude) to be correct, and moreover I believe that they contain all of mathematics. If someone makes such a statement he contradicts himself.[30]

Questo perché, se si considerassero questi assiomi come corretti, si dovrebbero anche considerare coerenti, ma, come abbiamo visto, la coerenza di un sistema *S* non è dimostrabile all'interno di esso, e di conseguenza la coerenza di un sistema formale che si suppone contenga tutta la matematica non può essere dimostrato in esso, dunque non può essere dimostrata matematicamente.

A questo proposito Gödel introduce una distinzione tra ciò che egli chiama "objective mathematics" e "subjective mathematics". Con la prima egli intende l'insieme di tutte le proposizioni matematiche vere, e in questo caso è valida l'affermazione che non esista alcun sistema formale ben definito che le contenga tutte. La seconda invece indica l'insieme delle proposizioni matematiche che possono essere dimostrate, e in questo caso potrebbe essere possibile che esista una regola che produca tutte le sue proposizioni. Tuttavia, anche qualora una tale regola esistesse, noi come umani non potremmo mai sapere con certezza matematica che tutte le proposizioni che produce siano coerenti, o meglio potremmo prendere una alla volta

[29] In Kurt Gödel, *Collected Works, Vol III: Unpublished Essays and Lectures,* a cura di S. Feferman et al., New York-Oxford, Oxford University Press, 1995, pp. 304-323
[30] *Ivi*, p. 309

ogni singola proposizione e stabilire che è vera, ma non potremmo dire che *tutte* queste proposizioni lo siano.

Ora, certamente si potrà empiricamente ritenere un tale insieme coerente sulla base di un sufficiente numero di istanze, o tramite altri metodi induttivi, ma come umani non potremmo mai avere la certezza matematica della sua coerenza. Per questo motivo Gödel ritiene che, se così fosse, almeno nel campo della pura matematica, la mente umana sarebbe equivalente ad una macchina di Turing che non riesca a comprendere completamente il proprio funzionamento. In questo caso, l'inabilità di comprendere completamente il proprio funzionamento non è da intendersi come l'impossibilità di produrre una teoria scientifica che descriva il comportamento della componente fisica della mente (ad esempio i meccanismi chimici e neuronali del cervello), ma alla assoluta impossibilità per la mente umana di dimostrare matematicamente la propria coerenza[31]. Infatti, qualora fosse possibile per la neurofisiologia dimostrare che "the brain suffices for the explanation of all mental phenomena and is a machine in the sense of Turing"[32] e che "such and such is the precise anatomical structure and physiological functioning of the part of the brain which performs mathematical thinking"[33], la mente umana (almeno nella sua capacità matematica) sarebbe equiparabile ad una macchina di Turing finita e, applicando ad essa i teoremi di incompletezza, si potrebbe dimostrare non solo la sua incompletezza, ma anche la sua incapacità di dimostrare la propria coerenza. Questa l'incompletezza della mente potrebbe erroneamente apparire ad essa stessa come la propria illimitatezza e inesaustività.

Ma da cosa deriva questa associazione tra la mente e una macchina di Turing? Abbiamo già illustrato l'equivalenza che sussiste tra sistemi formali e macchine di Turing, ovvero dato un qualsiasi sistema formale *S* si potrà "costruire" una macchina di Turing *T* che enumeri uno dopo l'altro tutti i teoremi derivabili dagli assiomi di *S*. Ora, se prendiamo in considerazione un sistema formale che contenga tutto ciò che la mente umana può dimostrare matematicamente, la macchina di Turing ad esso associata altro non sarebbe che la stessa capacità matematica della mente umana. Ma, essendo questa macchina di Turing equivalente ad un sistema formale che si presuppone coerente[34], si potranno applicare ad essa i due teoremi di incompletezza, di

[31] Gödel propone anche una versione più forte di questo argomento che al posto della semplice coerenza pone la correttezza della mente. Per correttezza si intende la capacità di dimostrare solo proposizioni vere, e non semplicemente coerenti (la coerenza non esclude che si possano dimostrare cose false, ma solo che si possa dimostrare il contraddittorio). Un sistema formale corretto è anche coerente, quindi è soggetto ai teoremi di incompletezza.

[32] Kurt Gödel, *Collected Works, Vol III: Unpublished Essays and Lectures,* a cura di S. Feferman et al., Oxford University Press, New York, Oxford, 1995, p. 309

[33] *Ibidem*

[34] Ovviamente si deve presupporre che la capacità matematica delle mente umana sia coerente, altrimenti potrebbe trivialmente dimostrare qualunque enunciato.

conseguenza esisteranno delle proposizioni che non potranno essere dimostrate da essa a partire dai propri assiomi, e nello specifico proprio la sua stessa coerenza. Questo significherebbe che, qualora la capacità di dimostrazione matematica della mente umana potesse essere formalizzata in un sistema formale (o, in altre parole, se la mente umana fosse, almeno nella sua capacità matematica, assimilabile ad una macchina di Turing) ci sarebbe almeno un enunciato matematico indimostrabile da essa (la sua stessa coerenza), e, rappresentando questo sistema la totalità delle dimostrazioni umanamente possibili, esisterebbero di conseguenza enunciati matematici *assolutamente* indimostrabili tramite metodi matematici per una mente umana. Da ciò, secondo Gödel, conseguirebbe che, non solo la matematica sarebbe incompletabile nel senso che non potrebbe essere interamente contenuta in alcun sistema formale ben definito, ma esisterebbero "*absolutely* unsolvable diophantine[35] problems (…), where the epithet "absolutely" means that they would be undecidable, not just within some particular axiomatic system, but by *any* mathematical proof the human mind can conceive"[36].

A questo punto ci si troverebbe davanti ad una disgiunzione: o la mente umana non può essere assimilata in tutto e per tutto ad una macchina di Turing e, almeno nell'ambito della matematica, sorpassa infinitamente il potere di ogni macchina di Turing, oppure esistono problemi diofantei assolutamente irrisolvibili, dove per "assolutamente irrisolvibili" non si intende semplicemente che siano irrisolvibili in un qualche determinato sistema assiomatico, ma che sia assolutamente impossibile che la mente umana possa mai produrre una loro dimostrazione. Questa disgiunzione, come afferma lo stesso Gödel, non deve intendersi come strettamente esclusiva, e si potrebbe dare il caso in cui entrambi i suoi termini siano veri. Ovvero la mente, nella sua componente matematica, potrebbe essere effettivamente equivalente ad una macchina di Turing ma rimanere comunque superiore ad essa nella risoluzione di *alcuni* problemi diofantei[37].

[35] Si definisce equazione diofantea un'equazione polinomiale con una o più incognite e coefficienti interi di cui si cercano le soluzioni intere. Prendono il loro nome dal matematico greco del III secolo Diofanto di Alessandria. Gödel, tramite l'aritmetizzazione della sintassi, ha dimostrato che il problema della coerenza di un sistema formale è equivalente ad un problema diofanteo, in quanto per coerenza si può intendere il fatto che nessun numero intero sia il numero di Gödel della dimostrazione di una contraddizione in quel sistema formale.

[36] Kurt Gödel, *Collected Works, Vol III: Unpublished Essays and Lectures,* a cura di S. Feferman et al., Oxford University Press, New York, Oxford, 1995, p. 310

[37] Potrebbe anche darsi il caso in cui la mente umana sia superiore alle macchine di Turing nello svolgimento di problemi di semplice aritmetica, ma non nella dimostrazione di teoremi di matematica superiore. Alcuni teoremi legati ai concetti di "infinito" e "transfinito" ancora indimostrati, tra cui la celebre ipotesi del continuo di Cantor, potrebbero essere assolutamente indimostrabili per una mente umana.

Ma cosa intende Gödel, nella prima parte di questa disgiunzione, quando afferma che la mente "infinitely surpasses the powers of any finite machine"[38]? Dato un sistema formale *S* a cui si possano applicare i teoremi di incompletezza (dunque che sia coerente e capace di aritmetica), la macchina di Turing *T* che enumera tutti e soli i teoremi di *S* non potrà dimostrare un enunciato vero di *S* (ovvero la sua coerenza) che è invece provabile dalla mente umana. Generalizzando si può affermare che la mente umana sarebbe in grado di dimostrare almeno un teorema in più di una qualsiasi macchina di Turing, indipendentemente dalla grandezza e complessità del sistema formale al quale essa corrisponde. Perciò ogni tentativo di formalizzare in un sistema ben definito la capacità matematica umana verrebbe continuamente superato dalla mente stessa, rendendo di conseguenza impossibile racchiudere la totalità degli enunciati matematici dimostrabili in un sistema formale e confermando la posizione filosofica di Gödel sull'inesaustività e incompletabilità della matematica.

Gödel riteneva che entrambi i termini della disgiunzione da lui proposta costituissero un argomento a sfavore di una concezione filosofica materialistica della mente e della matematica. Se fosse vera la prima parte della disgiunzione, infatti, egli afferma che sarebbe impossibile ridurre l'intera attività mentale umana ai semplici meccanismi fisici del cervello, che dal punto di vista fisico altro non è che "a finite machine with a finite number of parts, namely, the neurons and their connections"[39], spingendo verso un'interpretazione vitalistica del mondo; mentre, anche qualora fosse vero il secondo argomento della disgiunzione, ossia che la mente, in quanto assimilabile ad una macchina di Turing, non può dimostrare matematicamente ogni possibile enunciato matematico, se ne potrebbe comunque trarre un'argomentazione che confuti la teoria secondo la quale la matematica sarebbe un'invenzione umana. Questo perché, secondo l'autore, il creatore conosce necessariamente "all properties of his creatures, because they can't have any others except those he has given to them"[40]. Non ci addentreremo nella discussione di questo secondo argomento, concernente soprattutto temi di filosofia della matematica, per concentrarci invece sulla componente relativa al rapporto tra menti e macchine.

Ora, ci sono diverse testimonianze al di fuori della *Gibbs Lecture* che indicano chiaramente che Gödel avesse una posizione filosofica contraria ad una concezione meccanicistica della mente umana. Come riporta Hao Wang, nel corso di alcune discussioni private Gödel espresse l'opinione che l'ipotesi che la mente umana non sia separata dalla

[38] Kurt Gödel, *Collected Works, Vol III: Unpublished Essays and Lectures,* a cura di S. Feferman et al., Oxford University Press, New York, Oxford, 1995, p. 310
[39] *Ivi*, p. 311
[40] *Ibidem*

materia "sia un pregiudizio del nostro tempo, che sarà confutato scientificamente (forse in base al fatto che le cellule nervose non sono in numero sufficiente per eseguire tutte le operazioni osservabili della mente)"[41] e ancora "una confutazione consisterà, secondo Gödel, in un teorema matematico che dimostri che la formazione del corpo umano in tempi geologici secondo le leggi della fisica (…), a partire da una distribuzione casuale delle particelle elementari e del campo, è pressappoco altrettanto improbabile della separazione casuale dell'atmosfera nei suoi componenti"[42].

Per quale motivo allora nel corso della *Gibbs Lecture* Gödel ha preferito mantenere il suo argomento in una forma disgiuntiva, non propendendo esplicitamente per la prima parte dell'argomentazione, ossia per la superiorità della mente rispetto a qualsiasi macchina finita, e di conseguenza la sua irriducibilità al cervello, inteso come organo materiale e finito? Molto probabilmente poiché era conscio di non avere una "unassailable proof of the falsity of the mechanist position"[43], ovvero del fatto che il cervello possa costituire una causa sufficiente per spiegare completamente il funzionamento della mente umana, e che dunque la sua capacità di dimostrazione matematica possa essere in qualche modo formalizzata in un qualche sistema formale ben definito (rendendo di conseguenza impossibile la dimostrazione della coerenza almeno della sua componente matematica). Questo motivo, in aggiunta al fatto che, come vedremo nelle critiche rivolte all'argomento di Lucas, l'affermare che la mente umana possa sempre riconoscere la coerenza di una macchina di Turing finita, e di conseguenza esserle superiore, è estremamente difficile da provare (se non impossibile), spinsero Gödel a non affermare apertamente la sua posizione anti-meccanicistica e a mantenere valida la possibilità di una spiegazione puramente meccanica del funzionamento della mente umana.

Questa reticenza da parte di Gödel di affermare apertamente la sua posizione anti-meccanicistica è ulteriormente evidenziata dalla contesa che avvenne qualche anno in seguito alla Gibbs Lecture tra il logico austriaco e gli autori di *Gödel's Proof*, una popolare esposizione dei suoi teoremi di incompletezza, Ernest Nagel e James Newman[44]. Gödel chiese esplicitamente a costoro di avere il controllo sull'utilizzo dei propri materiali impiegati per la stesura del libro, una richiesta che venne polemicamente rifiutata da Nagel e Newman. La causa di questa disputa è probabilmente "their provocative related but contrasting views on the

[41] Hao Wang, *Dalla matematica alla filosofia*, tr. it. di Alberto Giacomelli, Torino, Boringhieri, 1984, p. 342
[42] *Ibidem*
[43] Solomon Feferman, *Gödel, Nagel, Minds and Machines*, "The Journal of Philosophy", 106 (2009), p. 211
[44] Per una approfondita ricostruzione di questa contesa cfr. Solomon Feferman, *Gödel, Nagel, Minds and Machines*, "Journal of Philosophy", 106 (2009), p. 201-219; e Kurt Gödel, *Collected Works, Volume V: Correspondence H-Z*, a cura di S. Feferman et al., Oxford University Press, New York, Oxford, 1995, p. 135

possible significance of Gödel's theorems for minds versus machines in the development of mathematics"[45], in aggiunta ad alcune imprecisioni di natura tecnica nell'esposizione da loro proposta dei teoremi di incompletezza. Ma quali sono nello specifico le affermazioni contestate da Gödel? Nel capitolo finale di *Gödel's Proof* intitolato "Conclusive Reflections", Nagel e Newman scrivono:

> Gödel's conclusions bear on the question whether a calculating machine can be constructed that would match the human brain in mathematical intelligence. (…) But, as Gödel showed in his incompleteness theorem, there are innumerable problems in elementary number theory that fall outside the scope of a fixed axiomatic method, and that such engines are incapable of answering, however intricate and ingenious their bult-in mechanisms may be and however rapid their operations.[46]

Sebbene Nagel e Newman non abbiano ben specificato cosa intendano con "calculating machines", e il fatto che i teoremi di incompletezza di Gödel si possano applicare a dei computers reali se e solo se essi si intendano come realizzazioni fisiche di una determinata macchina di Turing (solo in questo caso, infatti, varrebbe la loro equivalenza ai sistemi formali, e di conseguenza il loro essere soggetti ai teoremi di incompletezza), queste affermazioni sono tutto sommato corrette. È vero, infatti, data l'equivalenza già indicata tra macchine di Turing e sistemi formali, che una macchina calcolatrice reale, costruita sul modello di una definita macchina di Turing, sia soggetta ai teoremi di incompletezza e che quindi ci siano degli enunciati aritmetici che non sarà in grado di dimostrare. Ciò che probabilmente causò la contesa con Gödel furono le considerazioni seguenti, ovvero: "The theorem does indicate that in structure and power the human brain is far more complex and subtle that any nonliving machine yet envisaged"[47].

Ora, Nagel e Newman non aggiungono nessuna argomentazione dettagliata per spiegare il motivo di questa affermazione, limitandosi a indicare genericamente che: "the human brain appears to embody a structure of rules of operation which is far more powerful than the structure of currently conceived artificial machines"[48], per cui è difficile capire con esattezza secondo

[45] Solomon Feferman, *Gödel, Nagel, Minds and Machines*, "Journal of Philosophy", 106 (2009), p. 201
[46] Ernest Nagel, James R. Newman, *Gödel's Proof,* NYU Press, 2001, p. 111
[47] *Ivi*, p. 113
[48] *Ivi*, p. 112

quale procedimento logico abbiano dedotto la superiorità della mente umana sulle macchine calcolatrici. È possibile che abbiano derivato una tale superiorità dal fatto che, come da loro indicato, sia teoricamente possibile, dato un qualsiasi problema matematico costruire una macchina capace di risolverlo, ma che non sia invece possibile costruire una macchina che possa risolvere *tutti* i problemi matematici. Tuttavia, ciò non implica in nessun modo né che la mente possa invece farlo, né che la mente sia un qualcosa di differente o di superiore rispetto ad una macchina calcolatrice. Conscio di ciò, Gödel preferì non dichiarare apertamente la sua posizione, che, come abbiamo visto, era tuttavia molto probabilmente in accordo con una tale concezione della mente umana. Mantenendo il suo argomento in forma dicotomica, Gödel ha infatti evitato molte delle più immediate e facili critiche alla sua posizione filosofica, che invece, come vedremo, sono state rivolte sia a queste considerazioni di Nagel e Newman che all'argomento di Lucas. La preoccupazione da parte di Gödel che una formulazione sommaria e imprecisa di questi argomenti potesse essere facilmente oggetto di critiche, e di conseguenza screditata, fu il motivo che con ogni probabilità spinse il logico austriaco a chiedere a Nagel e Newman di avere il controllo e la supervisione sull'uso dei propri materiali usati in *Gödel's Proof*.

1.4.2 Lucas: Minds, Machines and Gödel

Il primo argomento che ha apertamente ed estesamente affermato la possibilità di stabilire un'assoluta differenza tra la mente umana e una qualsiasi macchina computazionale è quello proposto dal filosofo John Lucas nel suo articolo del 1961 *Minds, Machines and Gödel* pubblicato nella rivista *Philosophy*. Come abbiamo visto, nonostante una tale possibilità fosse stata proposta, seppur in forma ipotetica, da Gödel già una decina di anni prima nella sua *Gibbs Lecture*, il testo di tale conferenza non venne mai pubblicato nel corso della vita del logico austriaco. Per questa ragione, in aggiunta probabilmente al fatto che l'argomento proposto da Gödel venne posto in forma di dicotomia, lasciando dunque spazio anche alla possibilità che la mente umana possa essere effettivamente rappresentata totalmente da una macchina di Turing, la conferenza tenuta da Gödel non generò il vasto dibattito, e la pioggia di critiche, che invece si scatenarono in seguito alla pubblicazione dell'articolo di Lucas. Una reazione tale che spinse lo stesso Lucas, una trentina di anni più tardi, ad affermare: "I must have touched a raw nerve. That, of course, does not prove that I was right. Indeed, I should at once concede that I am very likely not to be entirely right (…) But I am increasingly persuaded that I was not entirely

wrong"[49]. Se infatti l'argomento di Lucas, almeno nella sua iniziale formulazione del 1961, è quasi unanimemente considerato come fallace, ha senza dubbio avuto il merito di generare un nutrito dibattito sul tema del rapporto che intercorre tra mente e macchine e ha ispirato ulteriori formulazioni di argomentazioni "gödeliane" di stampo anti-meccanicista, la più notevole delle quali, oltre ad una seconda formulazione del suo argomento originario da parte dello stesso Lucas, è senza dubbio quella proposta dal fisico matematico, recentemente insignito del premio Nobel per la fisica, Roger Penrose.

Ora, indipendentemente dalla posizione che si possa avere su questo tema e dal giudizio sulla validità di tali argomentazioni, è indubbio che una tale quantità di critiche e argomentazioni successive indichino che si tratti di un argomento di notevole interesse filosofico, le cui implicazioni si estendono oltre il semplice rapporto tra macchine e menti, andando a toccare argomenti come la capacità di auto comprensione del pensiero umano, i suoi limiti, le sue potenzialità e la sua capacità o incapacità di pensare concetti filosoficamente e matematicamente fondamentali come quello dell'infinito o della differenza tra la verità semantica e la provabilità di una proposizione.

Entriamo ora nel vivo dell'argomento proposto da Lucas. Il suo articolo *Minds, Machines and Gödel* si apre subito con la dichiarazione che: "Gödel's theorem seems to me to prove that Mechanism is false, that is, that minds cannot be explained as machines"[50], poiché tale teorema:

> Must apply to cybernetical machines, because it is of the essence of being a machine, that it should be a concrete instantiation of a formal system. It follows that given any machine which is consistent and capable of doing simple arithmetic, there is a formula which it is incapable of producing as being true -i.e. the formula is unprovable-in-the system- but which we can see to be true. It follows that no machine can be a complete or adequate model of the mind, that minds are essentially different from machines.[51]

Lucas, in questo caso, con il termine "cybernetical machine" si riferisce a un tipo di macchina "which performs a set of operations according to a definite set of rules"[52], ovvero quella particolare categoria di macchine dette calcolatrici, o computers, che, come abbiamo

[49] J.R. Lucas, *Minds, Machines and Gödel: a Retrospect*, in *Machines and Thought: the Legacy of Alan Turing*, vol. 1, a cura di P. J. R. Millican e A. Clark, Oxford, Oxford University Press, 1996, p. 106
[50] J.R. Lucas, *Minds, Machines, and Gödel*, "Philosophy", XXXVI (1961), p. 112
[51] *Ivi*, p. 113
[52] *Ibidem*

indicato precedentemente, possono, sotto alcune condizioni, essere considerate il corrispettivo reale e materiale del modello ideale di macchina calcolatrice che prende il nome di "macchina di Turing".

Ora, non tutte le macchine sono macchine di Turing, anzi propriamente nessuna macchina reale lo è. Quello di "Turing Machine" è infatti un concetto puramente ideale di una macchina calcolatrice perfetta che può operare senza alcuna limitazione di tempo o spazio, di conseguenza al massimo si può affermare che una macchina calcolatrice reale, a determinate condizioni, può corrispondere sul piano delle operazioni e del suo output ad una macchina di Turing. L'equivalenza che Lucas afferma sussistere generalmente tra macchine calcolatrici e sistemi formali, in realtà è formalmente possibile solo tra un determinato sistema formale e una macchina di Turing che abbia come prodotto tutti e soli i teoremi di quel sistema formale. Per questo motivo, al fine di applicare i teoremi di incompletezza di Gödel, che fanno riferimento a determinati tipi di sistemi formali, a delle macchine calcolatrici reali, Lucas avrebbe dovuto specificare che con "cybernetical machine" si intende una macchina calcolatrice reale, che operi in modo simile ad una macchina di Turing, e che si supponga che non incorra in errori di calcolo nel suo lavoro dovuti a limitazioni fisiche di spazio di memoria, al tempo di elaborazione, a errori nel suo programma o a un errato funzionamento delle sue componenti materiali.

Fatta questa precisazione sull'utilizzo del termine "cybernetical machine", specifichiamo ora cosa si intenda quando si considera la possibilità che una macchina di questo tipo possa essere considerata come un modello adeguato e completo della mente umana. Il modello generale di una macchina calcolatrice presuppone che alla macchina vengano fornite delle informazioni (input) sulle quali essa svolge delle operazioni meccaniche specificate dal suo programma e alla fine delle quali viene prodotto un risultato (output). Applicando un tale modello al cervello umano, lo si considera a tutti gli effetti una macchina calcolatrice i cui circuiti neuronali svolgono delle operazioni sulle informazioni sensoriali che esso riceve e forniscono un output determinato meccanicisticamente dal modo in cui è strutturato. Di conseguenza "given the way in which it is programmed (…) and the information which has been fed into it, the response is determined, and could, granted sufficient time, be calculated"[53].

Una macchina di questo tipo, sebbene meccanicamente determinata, non è necessariamente anche deterministica. Infatti, sostiene Lucas, si potrebbe ritenere che la mente umana sia una macchina che opera secondo una serie ben precisa di regole, ma la scelta della regola da applicare ad ogni passaggio del calcolo possa essere decisa casualmente attraverso

[53] *Ibidem*

uno strumento che generi un numero casuale[54]. Tuttavia, anche questo tipo di modello presenta dei limiti, specificatamente per il fatto che la scelta, sebbene casuale, non può essere tra un insieme qualsiasi di alternative: "any randomizing device must allow choices between those operations which will not lead to inconsistency"[55]. È contro un tipo di macchina di questo tipo, ovvero che possieda un "randomizing device, that acted whenever there were two or more operations possible, none of which could lead to inconsistency"[56], che Lucas rivolge il suo argomento. Questo perché il modello di un cervello solamente meccanico è molto meno forte di quello di un cervello sia meccanico che deterministico e, di conseguenza, l'argomento risulta più generale e vengono evitate al contempo anche complicazioni di carattere filosofico riguardanti, per esempio, il problema del libero arbitrio e della libertà di scelta, che avrebbero complicato l'argomentazione.

Consideriamo dunque la possibilità che una macchina di questo tipo venga costruita. Essa, come tutte le macchine reali, avrà un numero finito di componenti, un numero finito di tipologie di operazioni che è in grado di svolgere e un numero finito di assunzioni iniziali sulle quali può lavorare. Ma se sia il numero di assunzioni che il numero di operazioni possibili sono finiti, allora sarà teoricamente possibile per un essere umano trascrivere tramite un'appropriata simbologia, ignorando limitazioni di carattere fisico e temporale, l'intero funzionamento di una tale macchina usando carta e penna. Le assunzioni iniziali altro non sarebbero che gli assiomi di un sistema formale e le operazioni che la macchina svolge su di esse potrebbero essere tradotte in regole di inferenza. In questo modo si traduce l'intero funzionamento della macchina in un sistema formale, ovvero: "the conclusions it is possible for the machine to produce as being true will therefore correspond to the theorems that can be proved in the corresponding formal system"[57]. Da ciò, assieme all'assunzione iniziale dell'argomento, ossia che la macchina che stiamo considerando sia coerente e capace di svolgere una certa quantità di aritmetica, consegue che tra i teoremi che una tale macchina può dimostrare non ci potrà essere il cosiddetto "enunciato gödeliano" del sistema formale su cui essa stessa è basata.

Un tale enunciato, nella sua formulazione non formale, afferma, riferendosi a se stesso, di non essere dimostrabile in quel sistema formale. Si tratta di una sorta di paradosso del

[54] La generazione di sequenze di numeri casuali (nota anche come *RNG*, ossia *random number generation*) trova applicazione in molti campi della matematica e dell'informatica. I computers, tuttavia, operando esclusivamente tramite procedure meccaniche, non sono in grado di generare sequenze numeriche completamente casuali, ossia che non siano determinabili o riproducibili in nessun modo, ma solo sequenze apparentemente causali, chiamate *pseudo-random*. La generazione di sequenze numeriche veramente casuali è basata su eventi fisici non prevedibili quali ad esempio il decadimento atomico o la radiazione cosmica di fondo.
[55] J.R. Lucas, *Minds, Machines, and Gödel*, "Philosophy", XXXVI (1961), p. 114
[56] *Ibidem*
[57] *Ivi*, p. 115

mentitore[58], tuttavia, a differenza della classica formulazione di questo famoso paradosso logico, Gödel, tramite l'aritmetizzazione della sintassi ha reso inequivocabile la sua auto-referenzialità. Ora, se un tale enunciato che afferma di se stesso: "Questo enunciato non è dimostrabile nel sistema", fosse effettivamente dimostrabile in quel sistema formale, si verrebbe a creare una contraddizione, di conseguenza, se sappiamo che il sistema che stiamo considerando è coerente, un tale enunciato non vi potrà essere dimostrato. Ma se non può essere dimostrato allora quello che dice di sé è vero e quindi l'enunciato è vero: "we can see that the Gödelian formula is true: any rational being could follow Gödel's argument, and convince himself that the Gödelian formula, altough unprovable-in-the-system, was nonetheless -in fact for that very reason- true"[59]. Può la macchina che stiamo considerando fare altrettanto e stabilire la verità del suo enunciato gödeliano? No, perché un tale argomento non può essere formalizzato in essa, in quanto "richiederebbe di definirvi la nozione di verità, il che non possiamo fare per via del teorema di Tarski[60]"[61]. La verità (semantica) dell'enunciato di Gödel di un determinato sistema formale può essere stabilita solamente nel metalinguaggio di quel sistema formale, ossia su un piano logico ad esso superiore. L'argomento di Lucas, dunque, afferma che una mente umana è sempre capace di superare il piano logico di una qualsiasi macchina computazionale (che sia soggetta ai teoremi di Gödel), ossia di "uscire dal sistema" e stabilire nella metateoria una verità che essa non riesce a produrre.

Una qualsiasi macchina che venga realizzata con l'obiettivo di essere una rappresentazione perfetta di una mente umana, dovrà per forza possedere delle capacità aritmetiche, poiché gli esseri umani ne sono dotati. Ma se possiede questo tipo di capacità allora esisterà per forza un enunciato di cui gli esseri umani possono provare la verità, mentre essa non sarà in grado di fare altrettanto. Con ciò non si vuole implicare che gli esseri umani posseggano una migliore capacità aritmetica delle macchine computazionali, anzi sono loro nettamente inferiori sia per quanto riguarda la velocità di calcolo che la sua precisione e

[58] Famoso paradosso logico la cui prima formulazione è tradizionalmente fatta risalire al filosofo cretese Epimenide (VI secolo A.C). Di questo paradosso sono state proposte innumerevoli versioni nel corso della storia, ma fondamentalmente possono essere ricondotte tutte al modello di un enunciato (A) che riferendosi a se stesso afferma di essere falso: (A): l'enunciato (A) è falso. All'interno di una logica che rispetti il principio di bivalenza, qualunque valore di verità si attribuisca ad (A) risulta in una contraddizione; se infatti fosse vero sarebbe falso, e se fosse falso sarebbe vero, da qui la sua natura paradossale. Gödel tramite l'aritmetizzazione della sintassi ha fornito una versione di questo paradosso che rende inequivocabile la sua natura autoreferenziale.

[59] J.R. Lucas, *Minds, Machines, and Gödel*, "Philosophy", XXXVI (1961), p. 115

[60] Il Teorema di Indefinibilità di Alfred Tarski può essere informalmente espresso come l'impossibilità di definire formalmente concetti semantici come quello di verità o falsità all'interno di un certo linguaggio formale *L*. Tali concetti possono essere definiti per *L* solamente all'interno di un metalinguaggio *L'* che abbia una capacità di rappresentazione maggiore del linguaggio-oggetto originario *L*.

[61] Francesco Berto, *Tutti pazzi per Gödel!*, Bari, Laterza, 2008, p. 209

correttezza, né che sia impossibile costruire delle macchine che imitino delle caratteristiche e funzioni particolari della mente umana, lo stesso Lucas ammette che è perfettamente possibile, e altamente probabile, che si possano realizzare delle macchine che imitino ciascuna funzione cerebrale dell'essere umano. Quello che risulta impossibile, secondo il suo argomento, è che riescano a imitarle *tutte*:

> A machine cannot be a complete and adequate model of the mind. It cannot do *everything* that a mind can do, since however much it can do, there is always something which it cannot do, and a mind can. This is not to say that we cannot build a machine to simulate any desired piece of mind-like behaviour: it is only that we cannot build a machine to simulate *every* piece of mind-like behaviour. (…) We can never not even in principle, have a mechanical model of the mind.[62]

Ora, verrebbe forse naturale pensare di costruire una seconda macchina, uguale in tutto e per tutto alla prima, ma che riesca a produrre come vero l'enunciato gödeliano della precedente. Tuttavia, anche questa seconda macchina, sarà necessariamente basata su un sistema formale, sebbene più "grande" di quello della macchina originaria, e di conseguenza possiederà un suo proprio enunciato gödeliano che non potrà produrre come vero, e di cui invece una mente umana sarebbe perfettamente in grado di riconoscerne la verità. Si potrebbe procedere all'infinito nella costruzione di macchine sempre più grandi e complesse che cerchino di superare l'incompletezza della macchina loro precedente, senza tuttavia riuscirci: "However complicated a machine we can construct, it will, if it is a machine, correspond to a formal system, which in turn will be liable to the Gödel procedure for finding a formula unprovable-in-that-system. This formula the machine will be unable to produce as being true, although a mind can see that it is true. And so the machine will still not be an adequate model of the mind"[63].

Alternativamente, invece di costruire macchine via via sempre più grandi e complesse, si potrebbe semplicemente pensare di dotare una macchina di un "Gödelizing operator"[64], ovvero di una particolare regola di inferenza che, tramite una procedura ricorsiva, di volta in volta aggiunga l'enunciato gödeliano indimostrabile in essa ai suoi assiomi. Verrebbe, perciò, in questo modo generato un nuovo sistema formale, che naturalmente a sua volta possiederà un

[62] J.R. Lucas, *Minds, Machines, and Gödel*, "Philosophy", XXXVI (1961), p. 116
[63] *Ibidem*
[64] *Ivi*, p. 117

suo proprio enunciato gödeliano , il quale verrà nuovamente aggiunto ai suoi assiomi tramite la stessa regola di inferenza, e così via all'infinito. Tuttavia, un sistema formale di questo tipo altro non sarebbe che il sistema formale originario a cui è stato aggiunto l'insieme infinito di tutti questi enunciati gödeliani tra gli assiomi, tramite una finita e ben definita regola di inferenza, per cui una mente umana sarebbe comunque in grado, tenendo ciò in considerazione, di trovare una "falla" in esso, e dunque di superarlo: "in a sense, just because the mind has the last word, it can always pick a hole in any formal system presented to it as a model of its own workings"[65].

Un'altra possibile obiezione che si potrebbe rivolgere all'argomento di Lucas, consiste nell'affermare che, sebbene sia vero che una mente umana possa sempre dimostrare una verità che una certa macchina non riesce a produrre come vera, è tuttavia impossibile che ciò avvenga se invece di considerare una macchina particolare si considera l'insieme di *tutte* le macchine, sia quelle effettivamente costruite sia quelle costruibili. Un argomento simile era già stato formulato da Turing, il quale, considerando il caso in cui un essere umano riesca a dare una risposta che una macchina computazionale non è in grado di produrre, o riconosca un errore commesso da essa, affermò: "our superiority can only be felt on such an occasion in relation to the one machine over which we have scored our petty triumph. There would be no question of triumphing simultaneously over *all* machines. In short, then, there might be men cleverer than any given machine, but then again there might be other machines cleverer again, and so on"[66].

Tuttavia, l'obiettivo dell'argomento di Lucas non è tanto quello di stabilire quale tra mente umana e macchina sia superiore, o più intelligente, quanto di dimostrare l'impossibilità di realizzare una macchina che riesca in tutto e per tutto a imitare una mente umana, o in altre parole di fornire un modello meccanico che sia una perfetta rappresentazione di essa. In questo senso non ha importanza quante cose una macchina o l'intera totalità di esse sia in grado di fare in modo migliore rispetto ad un essere umano, ma è semplicemente sufficiente che una mente sia in grado di fare una qualsiasi cosa, anche insignificante, in più di un qualsiasi modello meccanico proposto, per rendere impossibile una sua completa rappresentazione da parte di una macchina computazionale:

> It is like a game. The mechanist has first turn. He produces a -any, but only a *definite* one- mechanical model of the mind. I point to something that it cannot do, but the

[65] *Ibidem*
[66] Alan Turing, *Computing machinery and Intelligence*, in *The Philosophy of Artificial Intelligence*, a cura di Margaret A. Boden, New York, Oxford University Press, 1990, p. 52

> mind can. The mechanist is free to modify his example, but each time he does so, I am entitled to look for defects in the revised model. If the mechanist can devise a model that I cannot find fault with, his thesis is established; (…) but since he cannot, in principle cannot, produce any mechanical model that is adequate, even though the point of failure is a minor one, he is bound to fail, and mechanism must be false.[67]

Questo è il cuore dell'argomento di Lucas, ovvero cercare di dimostrare, tramite i teoremi di incompletezza, la non riducibilità della mente umana ad un modello puramente meccanico e di conseguenza anche la falsità delle posizioni di tipo meccanicistico nei confronti della natura umana: "Since the time of Newton, the bogey of mechanist determinism has obsessed philosophers. (…) It seemed that we must look on human beings as determined automata"[68]. Da ciò, tuttavia, non deriva necessariamente un atteggiamento antiscientifico o "mistico":

> Our argument has set no limits to scientific enquiry: it will still be possible to investigate the working of the brain. It will still be possible to produce mechanical models of the mind. Only, now we can see that no mechanical model will be completely adequate, nor any explanation in purely mechanist terms.[69]

In questo senso Lucas, tramite il suo argomento, vuole creare una sorta di conciliazione tra la rinuncia alla libertà individuale che consegue ad un atteggiamento scientifico puramente meccanicistico nei confronti della nostra mente e l'opposta completa rinuncia ad ogni spiegazione scientifica della stessa, qualora si volesse mantenere la possibilità del libero arbitrio: "no longer on this count will it be incumbent on the natural philosopher to deny freedom in the name of science: no longer will the moralist feel the urge to abolish knowledge to make room for faith"[70]. Sarà infatti sempre possibile creare dei modelli scientifici che descrivano alla perfezione alcuni meccanismi del nostro cervello, ma non si sarà mai in grado di produrre un modello meccanico completo e perfettamente rappresentativo della mente: "we can produce models and explanations, and they will be illuminating: but, however far they go,

[67] J.R. Lucas, *Minds, Machines, and Gödel*, "Philosophy", XXXVI (1961), p. 118
[68] *Ivi*, p. 126
[69] *Ivi*, p. 127
[70] *Ibidem*

there will always remain more to be said. There is no arbitrary bound to scientific enquiry: but no scientific enquiry can ever exhaust the infinite variety of the human mind"[71].

1.5 Critiche agli argomenti gödeliani

Prenderemo ora in considerazione alcune delle principali e più importanti critiche che sono state rivolte alle diverse formulazioni degli argomenti gödeliani. Come abbiamo già indicato, questi argomenti hanno nel corso degli anni suscitato un acceso dibattito e una grande quantità di critiche sono state loro rivolte da parte di logici e filosofi.

È ormai quasi unanimemente riconosciuto che, almeno le prime due formulazioni degli argomenti gödeliani, ovvero quelle di Lucas e di Nagel e Newman, siano state confutate, mentre la versione proposta da Gödel, sia a causa del fatto che venne pubblicata solo postuma, sia perché venne proposta in forma ipotetica, non ha ricevuto lo stesso numero di attacchi.

Il principale argomento impiegato dai critici degli argomenti gödeliani è stato formulato per la prima volta dal filosofo Hilary Putnam e come vedremo si basa sul concetto di "coerenza" e sulla effettiva applicabilità dei teoremi di incompletezza, mentre altre critiche, tra cui quella di Solomon Feferman[72] e Stewart Shapiro[73], prendono in considerazione anche aspetti di tipo filosofico, come ad esempio l'eccessiva idealizzazione di molti dei concetti presi in considerazione in essi, tra cui proprio quello di "mente".

1.5.1 Hilary Putnam: la coerenza delle macchine

Hilary Putnam, nel suo articolo del 1960 *Minds and Machines*[74], che aveva come oggetto il rapporto mente-corpo e i possibili parallelismi che si possono costituire con una macchina che sappia rispondere a delle domande sul proprio funzionamento, inserì una breve, ma efficace, critica all'argomento proposto da Nagel e Newman in *Gödel's Proof*. Ora, anche lo stesso Putnam si trovò in difficoltà nel ricostruire con esattezza quale potesse essere il ragionamento di questi due autori ("Nagel and Newman give no argument, but I assume they must have this one in mind"[75]) e ne diede questa interpretazione:

[71] *Ibidem*
[72] Cfr. Solomon Feferman, *Gödel, Nagel, Minds and Machines*, "The Journal of Philosophy", 106 (2009), p. 201-219
[73] Cfr. Stewart Shapiro, *Incompleteness, Mechanism and Optimism*, "The Bulletin of Symbolic Logic", Vol. 4, No. 3 (1998), p. 273-302
[74] Hilary Putnam, *Minds and Machines*, in *Dimensions of Mind*, a cura di Sydney Hook, New York, New York University Press, 1960, pp. 138-164
[75] *Ivi*, p. 144

> Let *T* be a Turing machine which "represents" me in the sense that *T* can prove just the mathematical statements I can prove. Then the argument (…) is that by using Gödel's technique I can discover a proposition that *T* cannot prove, and moreover *I* can prove this proposition. This refutes the assumption that *T* "represents" me, hence I am not a Turing machine.[76]

Ovvero, si prenda in considerazione una macchina di Turing il cui output siano esattamente tutti gli enunciati matematici che una mente umana può provare; questa macchina rappresenterà di conseguenza un ben definito sistema formale *S* e, essendoci tra gli enunciati matematici che un essere umano può provare anche i teoremi aritmetici sufficienti per la Gödelizzazione, vi si potranno applicare i teoremi di incompletezza e quindi vi sarà un enunciato che *T* non riuscirà a dimostrare. Ma possiamo noi esseri umani "vedere" come vero questo enunciato e di conseguenza provare qualcosa in più rispetto ad una macchina che all'apparenza rappresenta la totalità delle proposizioni che una mente umana può dimostrare? Sì, ma se e solo se riusciamo a dimostrare la coerenza di *T*:

> Given an arbitrary machine *T*, all I can do is find a proposition *U* such that *I* can prove: (3) If *T* is consistent, *U* is true, where *U* is undecidable by *T* if *T* is in fact consistent. However, *T* can perfectly well prove (3) too! And the statement *U*, which *T* cannot prove (assuming consistency), *I* cannot prove either (unless I can prove that *T* is consistent, which is unlikely if *T* is very complicated)![77]

Infatti, quello che dimostrano i teoremi di incompletezza è che è dimostrabile in un determinato sistema formale *S*, che sia capace di una certa quantità di aritmetica, l'implicazione: "se *S* è coerente allora *G* (il suo enunciato gödeliano) è vero", ovvero: "whenever we know a theory *S* to be consistent, we also know the truth of a statement not provable in *S*. But in those cases when we have no idea whether or not *S* is consistent, we also have no idea whether or not a Gödel sentence *G* for *S* is true, and if we merely believe or guess *S* to be consistent, we merely believe or guess *G* to be true"[78].

---

[76] *Ibidem*
[77] *Ibidem*
[78] Torkel Franzén, *Gödel's Thorem: an Incomplete Guide to Its Use and Abuse*, Wellesley, A K Peters, 2005, p. 117

Nonostante la critica di Putnam precedette di un anno la pubblicazione dell'articolo di Lucas, essa può essere applicata anche a quest'ultimo. Infatti, sebbene l'argomento di Lucas sia molto più sviluppato e complesso di quello solo leggermente accennato da Nagel e Newman, il nucleo centrale dei due argomenti è sostanzialmente lo stesso, ovvero data una qualsiasi macchina computazionale ci sarà sempre almeno un enunciato (aritmetico) che essa non sarà in grado di produrre come vero, mentre una mente umana sarà in grado di vederne la verità. Di conseguenza, nessuna macchina sarà mai in grado di essere una perfetta rappresentazione delle capacità intellettive umane.

Ora, questo ragionamento si basa sull'applicazione dei teoremi di incompletezza ad una macchina computazionale, e di conseguenza è necessario preventivamente dimostrare la coerenza di una tale macchina. Ma possiamo noi esseri umani dimostrare sempre la coerenza di una qualsiasi macchina computazionale? Consideriamo il caso in cui ci venga presentata una macchina *M* che sia apparentemente una perfetta rappresentazione delle capacità matematiche umana, e supponiamo anche che siamo perfettamente in grado di dimostrarne la coerenza e di conseguenza dimostrare che esiste una certa proposizione *G* di cui noi possiamo riconoscere la verità e che la macchina in questione non riesce a produrre come vera. È possibile a questo punto, come anche Lucas ammette, costruire una seconda macchina *M'* che sia la macchina iniziale *M* a cui è stata aggiunta la proposizione *G* come un assioma. Ora, per dimostrare che nemmeno questa nuova macchina *M'* è una perfetta rappresentazione delle capacità intellettive di un essere umano, dovremmo innanzitutto dimostrarne la coerenza e una volta fatto ciò applicare nuovamente i teoremi di incompletezza, trovando una nuova proposizione *G'* indimostrabile da *M'*. A questo punto però si potrebbe costruire una macchina *M''* e così via all'infinito.

Come abbiamo visto, Lucas, per cercare di uscire da questa spirale infinita di costanti implementazioni di nuovi enunciati, aveva immaginato di dotare la macchina iniziale *M* di un "Gödelizing operator" che tramite una procedura ricorsiva ben definita fornisse a *M* l'insieme infinito di enunciati gödeliani che si potrebbero creare, e dimostrando l'incompletezza di questa nuova macchina ne aveva dichiarato la sua non adeguatezza come modello della mente umana. Purtroppo, però, questo ragionamento non solo non risolve il problema, ma lo rende ancora più complesso; vediamo il perché.

Immaginiamo di costruire una nuova macchina M_G dotata dell'"operatore gödeliano" proposto da Lucas, che renda di conseguenza possibile esprimere l'altrimenti infinito insieme di enunciati gödeliani. A questo punto, qualora riuscissimo a dimostrare la coerenza di questa macchina M_G, proveremmo il suo enunciato gödeliano G_{MG} e di conseguenza potremmo

affermare la sua inadeguatezza nei confronti della nostra mente. Ora, il problema è che anche questo enunciato G_{MG} può essere formalmente aggiunto a M_G come un suo assioma, ricadendo così nella situazione iniziale con tuttavia in aggiunta la complicazione derivante dal fatto che M_G possiede rispetto alla macchina *M* iniziale un insieme infinito di enunciati gödeliani. In altre parole, ci troveremmo di fronte ad una serie infinita di macchine dotate di un numero infinito di insiemi infiniti di enunciati gödeliani, di cui, per poterne provare l'inadeguatezza nel confronto con la nostra mente, dovremmo essere in grado di dimostrare la coerenza. Aggiungendo questo operatore gödeliano "abbiamo fatto un "salto" importante (in termini cantoriani: siamo passati dagli ordinali finiti al primo ordinale transfinito)"[79].

Ci troviamo dunque nel regno del transfinito, abbiamo ora di fronte sistemi formali talmente grandi e complessi di cui non è semplicemente messa in discussione la possibilità per una mente umana di dimostrarne la coerenza, ma la stessa possibilità di comprensione e raffigurazione. Indichiamo con "ω"[80] l'insieme di tutti gli infiniti enunciati gödeliani che abbiamo aggiunto alla macchina iniziale *M* per ottenere M_G. Questo insieme è composto da tutti gli infiniti enunciati gödeliani che si possono creare per *M* e per tutte le sue successive iterazioni, ovvero $\omega=\{G,G_1,G_2,\dots\}$. Abbiamo tuttavia visto che la medesima procedura si può applicare anche a M_G, per cui essa possiederà non solo tutti gli enunciati gödeliani ottenuti a partire da *M* ma anche quelli ottenuti applicando un ulteriore operatore gödeliano a se stessa, dunque avremo ω+1, ω+2, ω+3, … (dove per ω+1, ω+2, etc. si intendono gli ordinali successori di ω). Ma anche questa ulteriore espansione può essere iterata all'infinito per cui arriveremo a ω+ω, ovvero 2ω, e poi 3ω, 4ω, fino ad arrivare a ωxω, ossia ω^2 e ancora ω^3, …, ω^ω: "in these terms, the Lucas-Penrose contest to write and assert Gödel sentences becomes a contest to enumerate recursive ordinals. One might think that all Lucas has to do is iterate the procedure of adding Gödel sentences (…) far enough. The problem, however, is with the crucial notion of "far enough""[81]. Oltre un certo punto, infatti, sembra venire meno la nostra stessa capacità di "Gödelizzare". Possiamo pensare questo problema come un problema di denominazione. Lucas cerca di riassumere un insieme infinto di enunciati gödeliani in un unico schema che consiste in una procedura algoritmica; abbiamo chiamato questo schema ω, questo perché tutti i nomi dei numeri naturali sono già stati "utilizzati" per nominare l'infinito numero di enunciati gödeliani che costituiscono ω. Tuttavia, come abbiamo visto, anche questo schema ω può essere

[79] Francesco Berto, *Tutti pazzi per Gödel!*, Bari, Laterza, 2008, p. 216
[80] Georg Cantor utilizzò la lettera greca "ω" per indicare il primo ordinale transfinito.
[81] Stewart Shapiro, *Incompleteness, Mechanism and Optimism*, "The Bulletin of Symbolic Logic", Vol. 4, No. 3 (1998), p. 288

esteso, come quello a lui successivo, e così via all'infinito, generando una sempre più esponenziale degenerazione in infinità di infinità. A questo proposito Hofstadter scrive:

> This is a move of a different character than any that has gone before, and is given the new name "ω". The newness of the name is quite important. It is the first example where the old naming scheme -which only included names for all the natural numbers- had to be transcended. Then come some more extensions, some of whose names seem quite obvious, others of which are rather tricky. But eventually, we run out of names once again- at the point where the answer-schemas
>
> $$\omega, \omega^{\omega}, \omega^{\omega^{\omega}}, \ldots$$
>
> are all subsumed into one outrageously complex answer schema. The altogether new name "ε_0"[82] is supplied for this one.[83]

È chiaro che con questo metodo qualunque schema di risposte, indipendentemente dalla sua grandezza e complessità, possa essere trasceso. Ed è altrettanto evidente la possibilità che un essere umano ben presto non riesca più a stare dietro a questa corsa nel transfinito. Una delle ragioni di ciò è che questa procedura non può essere schematizzata e riprodotta algoritmicamente, o in altre parole, non esiste alcuna macchina che possa farlo:

> Now offhand you may think that these irregularities in the progression from *ordinal* to *ordinal* (as these names of infinity are called) could be handled by a computer program. (…) It turns out that the irregularities themselves happen in irregular ways, and one would need also a second-order program -that is, a program which makes new programs which make new names. And even this is not enough. Eventually a third-order program becomes necessary. And so on, and so on.[84]

Questa impossibilità di generare meccanicamente nomi per gli ordinali è stata espressa in un complesso teorema da Alonzo Church e Stephen C. Kleene, che informalmente si può

[82] Generalmente si indica con ε_0 il primo ordinale che non è possibile ottenere tramite un numero finito di operazioni di addizione, moltiplicazione o elevamento a potenza a partire da ω.
[83] Douglas Hofstadter, *Gödel, Escher, Bach: An Eternal Golden Braid*, New York, Basic Books, 1979, p. 475
[84] *Ivi*, p. 476

riassumere in questo modo: "There is no *recursive* enumeration of every *recursive* ordinal"[85]. Ma, nello specifico, cosa comporta ciò per l'argomento proposto da Lucas?

> No algorithmic method can tell how to apply the method of Gödel to all possible kinds of formal systems. (…) Therefore one must conclude that any human being simply will reach the limits of his own ability to Gödelize at some point. From there on out, formal systems of that complexity, though admittedly incomplete for the Gödel reason, will have as much power as that human being.[86]

In questo senso si può pensare l'argomento di Lucas come una sfida tra un uomo e una macchina nell'enumerare numeri ordinali ricorsivi. Poiché per il teorema di Church-Kleene nessuna macchina è in grado di nominare tutti gli ordinali ricorsivi, se un essere umano ne fosse capace, questo comporterebbe una superiorità della mente umana rispetto ai calcolatori, al pari di quella ricercata da Lucas tramite il suo argomento gödeliano originario. Ora, anche ammettendo che possa essere *idealmente* possibile per un essere umano fare ciò, sarebbe *praticamente* impossibile per un qualsiasi essere umano reale e vivente enumerare *tutti* gli ordinali, poiché, essendo essi infiniti, sarebbe necessario avere a disposizione un tempo infinito per elencarli tutti. Dunque, si tratterebbe di verificare se, data una qualsiasi macchina computazionale, un essere umano sarebbe sempre in grado di indicare almeno un ordinale in più di essa. Ma abbiamo poco sopra indicato che non esiste una procedura algoritmica per denominare tutti gli algoritmi ricorsivi, come invece può accadere nel caso dei numeri naturali, che possono essere tutti ottenuti tramite un algoritmo ricorsivo che dato un qualsiasi numero naturale, ne generi il successivo semplicemente aggiungendo 1 ad esso, e così via. Non esistendo una procedura algoritmica, è molto difficile provare la possibilità per un essere umano di essere sempre in grado di enumerare almeno un ordinale in più di una macchina, a meno di non prendere in considerazione supposte capacità umane di intuizione matematica.

Per questo motivo, Hofstadter conclude che prima o poi la capacità di Gödelizzazione umana giungerà inevitabilmente al proprio limite e di conseguenza essa non può essere usata come argomento per stabilire una differenza e superiorità di principio tra la nostra mente e una qualsiasi macchina computazionale. Ovviamente questo limite non sarà mai esattamente calcolabile e definibile, e sarà anche diverso da individuo a individuo, ma sicuramente esiste,

[85] Stewart Shapiro, *Incompleteness, Mechanism and Optimism*, in "The Bulletin of Symbolic Logic", Vol. 4, No. 3 (1998), p. 289
[86] Douglas Hofstadter, *Gödel, Escher, Bach: An Eternal Golden Braid*, New York, Basic Books, 1979, p. 476

nello stesso modo in cui è impossibile indicare con precisione quale sia il limite di peso che un essere umano può sollevare, ma di certo nessuno sarà mai in grado di sollevare mille tonnellate.

Ad ogni "salto" da un'infinità alla successiva abbiamo bisogno di adottare degli assiomi dell'infinito sempre più forti, mettendo sempre più in dubbio la nostra capacità di comprendere questi sistemi assurdamente complessi, e di conseguenza facendo vacillare anche la nostra capacità di stabilirne la coerenza. Ma "nel momento in cui non sappiamo (più) con certezza se il sistema è coerente, non sappiamo (più) con certezza se il suo enunciato gödeliano è vero. Sembra dunque che non vi sia una base per affermare che noi possiamo oltrepassare *qualsiasi* sistema formale dato, e quindi qualsiasi macchina. Ma gli ≪argomenti gödeliani≫ richiedevano nientemeno che questo."[87]

Lucas rispose a queste ed altre critiche nel corso della Turing Conference di Brighton tenutasi il 6 aprile 1990. A proposito dell'impossibilità per un essere umano di essere capace di stabilire la coerenza di una qualsiasi macchina computazionale egli affermò:

> There is a claim being seriously maintained by the mechanist that the mind can be represented by some machine. (…) It is reasonable to ask him not only what the specification of the machine is, but whether it is consistent. (…) The consistency of the machine is established not by the mathematical ability of the mind but on the word of the mechanist. The mechanist has claimed that his machine is consistent. If so, it cannot prove its Gödelian sentence, which the mind can none the less see to be true: if not, it is out of court anyhow.[88]

Tuttavia, spostare l'onere della prova della coerenza della macchina sul meccanicista che dichiara di aver realizzato un modello meccanico della mente umana, non risolve affatto la questione. Il meccanicista, infatti, incontra esattamente le stesse difficoltà che può incontrare Lucas nello stabilire la coerenza di una macchina computazionale, per cui la sua assicurazione della coerenza di essa non costituisce affatto una base sufficiente per l'applicazione dei teoremi di incompletezza.

La macchina dichiarata come coerente potrebbe effettivamente esserlo, oppure ci potrebbe essere stato un errore da parte del meccanicista nello stabilire la sua coerenza. Se la macchina fosse coerente, allora tramite i teoremi di incompletezza potremo stabilire la sua non

[87] Francesco Berto, *Tutti pazzi per Gödel!*, Bari, Laterza, 2008, p. 218
[88] J.R. Lucas, *Minds, Machines and Gödel: a Retrospect*, in *Machines and Thought*, a cura di Peter Millican e Andy Clark, New York-Oxford, Oxford University Press, 1996, p. 117

perfetta assimilabilità ad una mente umana, ma se non fosse coerente i teoremi di incompletezza non ci direbbero nulla a questo riguardo. Entrerebbero a questo punto in gioco altre argomentazioni e considerazioni, basate sul fatto che la nostra mente possa essere provata coerente, e di conseguenza non sarebbe perfettamente rappresentabile da una macchina computazionale incoerente. Ma questo tipo di considerazioni non dipendono in nessun modo dai teoremi di incompletezza. Il punto centrale in questo caso non è se esista o meno la possibilità di stabilire la fondamentale irriducibilità della mente umana a sistemi di calcolo meccanici tramite argomenti filosofici informali, ma se sia possibile farlo mediante l'applicazione di specifici teoremi di logica formale. In questo senso le critiche rivolte all'argomento di Lucas sulla base della capacità di riconoscere la coerenza di una determinata macchina computazionale, prescindono dalla posizione filosofica che si possa avere in questo dibattito, e sono rivolte alla stessa validità formale dell'argomento.

Ora, come abbiamo visto, non si può provare in modo inequivocabile l'incapacità di principio per un essere umano di essere sempre in grado di stabilire la coerenza di un determinato sistema formale, ma sicuramente si può affermare che essa sia estremamente probabile. Di conseguenza, quello che si può ricavare da queste critiche è che appare estremamente improbabile, anche se non necessariamente impossibile, che si possano usare i Teoremi di Incompletezza di Gödel al fine di trarre le conclusioni anti-meccanicistiche in relazione alla natura della mente umana proposte da Lucas.

1.5.2 Altre Critiche

Dopo aver analizzato le principali critiche di carattere logico-formale rivolte agli argomenti gödeliani che si oppongono ad una concezione meccanicistica della mente umana, prendiamo ora in esame anche alcune delle più significative tra le numerose critiche di natura più prettamente filosofica che nel corso degli anni logici e filosofi hanno rivolto ad essi.

Alcuni autori, tra i quali Solomon Feferman e Stewart Shapiro, hanno obiettato che questi argomenti richiedano e impieghino delle definizioni e dei concetti eccessivamente idealizzati riguardanti sia la natura delle macchine sia della mente umana: "The raw thesis that the human mind is, or can be modeled as, a digital computer or Turing machine, is too vague to apply something as sharp and delicate as the Gödel theorem"[89]. Effettivamente Lucas non definisce

[89] Stewart Shapiro, *Incompleteness, Mechanism and Optimism*, in "The Bulletin of Symbolic Logic", Vol. 4, No. 3 (1998), p. 275

affatto cosa intenda quando utilizza termini come "mente umana" o "macchina calcolatrice", e, come abbiamo già visto, i teoremi di logica formali richiedono delle precisissime condizioni per poter essere correttamente applicati. Se generalmente si possono intendere le macchine che Lucas menziona come delle realizzazioni materiali del ben definito concetto di "Turing machine", ben più difficile risulta capire cosa si possa intendere con il termine "mente umana", sempre che se ne possa dare una definizione formale.

Ora, il problema è che questi argomenti richiedono necessariamente l'impiego di concetti estremamente idealizzati, vediamo il perché. È evidente che il numero di pensieri, scritti e asserzioni prodotti durante la vita di un qualsiasi essere umano sia un insieme finito, per quanto grande o difficile da formalizzare. Così come è evidente che sia finito anche l'insieme di tutti i pensieri prodotti dagli esseri umani passati e futuri, data l'elevata probabilità che prima o poi l'essere umano si estingua. Ma, data la sua finitezza, è di conseguenza assolutamente possibile, in linea di principio, programmare una macchina che lo produca perfettamente come output. A questo punto questa macchina sarebbe in grado di produrre come vere tutte le medesime proposizioni che anche un essere umano (o la totalità degli esseri umani) riesce a produrre come vere, rendendo invalidi gli argomenti gödeliani. Inoltre, basterebbe che in questo sistema costituito da tutte le asserzioni umane vi si trovasse anche una sola contraddizione, per rendere incoerente l'intero insieme, e di conseguenza inapplicabili ad esso i teoremi di incompletezza. Ma è evidente che la storia del pensiero umano sia piena di contraddizioni, di conseguenza un tale sistema non potrebbe che essere incoerente. Infine, come abbiamo precedentemente indicato, l'argomento di Lucas implica un numero transfinito di enunciati gödeliani, che quindi non può certamente essere contenuto in un insieme finito.

Per queste tre ragioni, gli argomenti gödeliani non possono prendere in considerazione ciò che un singolo essere umano, o l'intera comunità di tutti gli esseri umani presenti, passati e futuri, possono *realmente* dire o pensare, ma solo ciò che possono dire o pensare *idealmente* e *potenzialmente*. Ma a questo punto "we must idealize on the "machines" as well. (…) We ignore finite limits and assume that our machines never run out of memory, space, time, and attention span. (…) In short we deal with *Turing machines*, with their fixed programs and unlimited tapes"[90]. Ovviamente una macchina di Turing che non debba sottostare alle fisiche limitazioni di un essere umano sarà sicuramente "superiore" a lui, almeno dal punto di vista della capacità di provare teoremi matematici. A causa di ciò, è a questo punto necessario introdurre lo stesso tipo di idealizzazione anche per quanto riguarda l'essere umano:

[90] *Ivi*, p. 276

> The principals to the present debate (try to) make idealizing assumptions about humans analogous to those of Turing machines. They do not speak of the theorems a subject does produce, but the theorems that she *can* produce. (…) In short, the envisioned creatures have unlimited lifetimes, unlimited attention spans and energy, and unlimited materials at their disposal. Yet they are like humans in every other respect- whatever that means.[91]

È evidente che a questo punto non si sta più trattando di esseri umani in carne ed ossa, e di macchine reali con le loro limitazioni, problemi ed errori, ma di esseri puramente ideali. Se da un lato è tutto sommato semplice indicare cosa si intenda per una macchina computazionale ideale (il modello di macchina di Turing è ben definito e formalizzato), dall'altro è ben più problematico definire la caratteristiche di questo essere umano ideale e individuare con precisione ciò che esso *potenzialmente* sarebbe in grado di provare in termini logico-matematici, in altre parole: "we must agree on a way to resolve this matter, and come up with a clear and unambiguous conception of idealized human mathematical ability. Otherwise, there is no meaningful debate. We need the idealizations *before* we can assess the relevance of the various theorems"[92].

Questa difficoltà nel dare una precisa definizione delle capacità potenziali della mente umana richieste dagli argomenti gödeliani venne ammessa dallo stesso Lucas: "In finite life-span only a finite number of the propositions can be recognized, only a finite set of problems can be solved. And a machine can be programmed to do that. Of course, we reckon that a man *can* go on do to more, but it is difficult to capture that sense of infinity potentiality. This is true. It is difficult to capture the sense of infinite potentiality"[93].

Tuttavia, è proprio questa potenzialità, sebbene difficile da definire ciò che, nel pensiero di Lucas, si pone come una delle differenze fondamentali tra uomo e macchina. Le macchine, infatti, definite una volta per tutte dal modo in cui sono state costruite e programmate, sono necessariamente limitate, nel loro output, ad un numero qualitativamente finito di espressioni ed enunciati, anche se potenzialmente infinito in termini quantitativi, qualora prendessimo in considerazione modelli di macchine computazionali ideali come le macchine di Turing. In

[91] *Ibidem*
[92] *Ibidem*
[93] J.R. Lucas, *Minds, Machines and Gödel: a Retrospect*, in *Machines and Thought*, a cura di Peter Millican e Andy Clark, New York-Oxford, Oxford University Press, 1996, p. 109

questo senso, una macchina non potrà mai "uscire" dal proprio sistema, ovvero produrre un qualcosa per il quale non è stata progettata, e per questo motivo essa può essere paragonata ad un essere umano considerato in una "*post-mortem* view"[94], che dunque non abbia più la possibilità di generare attivamente nuovi pensieri.

Lucas concede che sia un compito perfettamente possibile, e tutto sommato non difficilmente realizzabile, quello di costruire una macchina che sia una rappresentazione fedele della vita di un essere umano una volta che si sia conclusa, ma lo scopo del suo argomento è quello di dimostrare l'impossibilità per un computer di imitare un essere umano *vivo*, dunque ancora nel pieno delle sue potenzialità: "What is in issue is whether a computer can copy a living me, when I have not as yet done all that i shall do, and can do many different things"[95]. Per questo motivo il suo argomento richiede necessariamente il concetto di *potenzialità* della mente umana, e le conseguenti e problematiche idealizzazioni che ne conseguono, poiché: "a modally "flat" account of the mind in terms only of what it has done is as unconvincing as an account of cause which considers only constant conjunction, and not what would have been the case had circumstances been different"[96].

Lucas ha posto il suo argomento nella forma di una sfida aperta con il meccanicista proprio per riflettere questo carattere potenziale; ovvero è sempre possibile per un essere umano provare la sua differenza nei confronti di una qualsiasi macchina che sia costruita con l'obbiettivo di essere una sua perfetta riproduzione, ma non potrà fare lo stesso nei confronti dell'ideale insieme infinito di *tutte* le macchine computazionali costruite o costruibili. Il carattere di sfida continua serve proprio a sottolineare questa continua tensione, l'impossibilità di riprodurre perfettamente il funzionamento di una mente umana in modo puramente meccanico non può essere provato una volta per tutte, ma deve essere costantemente ribadito.

Ora, se da un lato questo tipo di struttura dinamica che Lucas ha dato al suo argomento è necessaria, in quanto un qualsiasi modello statico della mente potrebbe essere facilmente emulato da una macchina, dall'altro presta anche il fianco ad altre critiche. Se infatti Lucas può affermare che: "the mind, being in fact "alive", can always go one better than any formal, ossified, dead, system can. (…) The mind always has the last word"[97], il meccanicista potrebbe affermare invece di essere sempre in grado, dato un qualsiasi modello non computazionale della mente proposto da Lucas, di realizzare una macchina che ne sia una adeguata riproduzione. Se

[94] *Ibidem*
[95] *Ibidem*
[96] *Ibidem*
[97] J.R. Lucas, *Minds, Machines, and Gödel*, in "Philosophy", XXXVI (1961), p. 116

la sfida si pone come potenzialmente infinita non può esserci una "last word", nessuna delle due parti può avere l'ultima parola, se così fosse, infatti, il numero di iterazioni del confronto tra le due parti sarebbe finito.

Questo confronto può essere pensato, per fare un'analogia, come una sorta di Serie di Grandi[98]: il meccanicista inizia col proporre un modello computazionale della mente umana (1) e Lucas ne dimostra l'inadeguatezza (-1), a quel punto il meccanicista ne propone un altro (+1) e Lucas confuta anche quest'ultimo (-1), e così via all'infinito. In questo senso non è possibile stabilire un vincitore, poiché si tratta di una potenzialmente infinita fluttuazione tra le due posizioni, così come non è possibile assegnare un valore alla serie di Grandi, almeno nel senso convenzionale del termine. Ma se decidessimo di provare comunque ad assegnare un valore a questa serie, o a proclamare un vincitore della sfida, usando dei metodi informali, allora sarebbe possibile, a seconda di come scegliamo di considerare la situazione, ottenere addirittura tre diversi risultati, tutti ugualmente validi. Se infatti, applicando "impropriamente" dei metodi algebrici alla serie di Grandi, nonostante essa sia divergente, possiamo ottenerne come valore sia 0 che 1, a seconda di come decidiamo di raggruppare tra parentesi i termini della serie. Allo stesso modo possiamo stabilire un diverso vincitore dalla sfida a seconda di chi decidiamo abbia effettivamente l'ultima parola. Se invece consideriamo il terzo valore che si può decidere di assegnare alla serie, ovvero $\frac{1}{2}$, appare evidente che in realtà non possa esistere un vero e proprio vincitore, ma che entrambe gli argomenti della discussione abbiano la stessa validità, e che la soluzione sia nella insoluta e insolubile tensione tra i due poli.

Ora, se provassimo ad uscire dal carattere ideale e potenziale dell'argomento, e immaginassimo una vera e propria sfida reale nello stile proposto da Lucas, probabilmente una delle due parti prima o poi arriverebbe al limite fisico o della propria capacità di gödelizzazione o della propria capacità di progettare nuove macchine. A questo punto ci sarebbe senza dubbio un vincitore, anche se sarebbe per così dire un vincitore "pratico", dunque la sua vittoria non sarebbe dovuta a ragioni di principio ma solamente a motivazioni e limitazioni fisiche e contingenti. Per questi motivi non sembra essere del tutto corretta la certezza esibita da Lucas

[98] Serie matematica divergente scoperta dal matematico italiano Guido Grandi nel 1703 che consiste nella somma infinita 1-1+1-1+… . Essendo una serie divergente non è possibile assegnarle un valore nel senso usuale del termine, né applicarvi delle normali operazioni che solitamente possono essere applicate ad una serie convergente. Si possono infatti ottenere due valori diversi a seconda di come si decide di raggruppare i termini della serie:
1) (1-1)+(1-1)+(1-1)+…= 0+0+0+…= 0
2) 1+(-1+1) +(-1+1) +(-1+1)+… = 1+0+0+0+…= 1
Utilizzando altri metodi, invece, si può ottenere come risultato di questa somma il valore $\frac{1}{2}$.

di poter avere sempre "l'ultima parola" nel corso del suo confronto con il sostenitore della natura meccanica della mente umana.

Anche per quanto riguarda la coerenza della mente umana, o almeno della coerenza delle sue capacità matematiche (gli argomenti gödeliani prendono in considerazione solamente la componente matematica della nostra mente, escludendo quindi la necessità di provare o supporre la sua totale coerenza), non mancano i problemi. Sono numerosi, infatti, i casi in cui in ambito matematico sono stati provati dei teoremi poi scoperti falsi. Ciò mette fortemente in dubbio la possibilità di una perfetta e assoluta coerenza della nostra capacità matematica. Una soluzione sarebbe quella di imputare tali errori alla natura corporea e finita della nostra capacità di memoria e di calcolo, creando di conseguenza una sorta di dicotomia tra una supposta e ideale abilità matematica completamente corretta e incapace di errori, e la reale applicazione di tale abilità, affetta dalle limitazioni che inevitabilmente affliggono un qualsiasi essere finito, sia esso un uomo o una macchina calcolatrice. Ma possiamo sempre distinguere se un'incoerenza in ambito matematico sia imputabile esclusivamente ad un errore causato dalla nostra natura fisica nell'applicazione di regole e procedure idealmente perfette? Questa distinzione, che Shapiro indica analoga a quella esistente in linguistica tra "competence" e "performance"[99], presupporrebbe di ignorare completamente qualsiasi errore causato dalla componente di performance e di concentrarsi esclusivamente sulla competence, che si deve presupporre come coerente; presupporre perché dimostrare formalmente la coerenza di un sistema matematico del tipo richiesto dagli argomenti gödeliani, ovvero illimitato e illimitabile, è di certo al di là della effettiva capacità matematica di un qualsiasi essere umano reale, o di un qualunque numero finito di esseri umani.

[99] Stewart Shapiro, *Incompleteness, Mechanism and Optimism*, in "The Bulletin of Symbolic Logic", Vol. 4, No. 3 (1998), p. 276

CAPITOLO II - CONOSCENZA TACITA E SENSO COMUNE: IL "FRAME PROBLEM" DELL'IA

Nel corso del capitolo precedente abbiamo illustrato quelli che possono essere considerati come i maggiori argomenti di carattere puramente formale che si oppongono ad una concezione meccanicistica e algoritmica della mente umana, e che di conseguenza pongono l'impossibilità teoretica per una macchina computazionale di essere una esatta rappresentazione di una mente umana.

Tuttavia, come abbiamo indicato, questi argomenti rimangono su un piano estremamente idealizzato sia per quanto riguarda la natura delle macchine, sia per quanto riguarda la natura degli esseri umani e della loro mente. Lo stesso "terreno di scontro" tra le due parti, altro non è che un teorema di logica formale, le cui implicazioni sono connesse alla possibilità di produrre e verificare la veridicità di particolari proposizioni riguardanti sistemi formali molto grandi e complessi. Per questo motivo, è molto difficile sia valutarne l'effettiva validità, sia la loro effettiva rilevanza in contesti pratici e di effettiva interazione tra uomini e sistemi di intelligenza artificiale. In altre parole, non è detto che una teorica possibilità o impossibilità di simulazione del pensiero umano da parte di una macchina si traduca in una effettiva possibilità o impossibilità pratica. Infatti, anche qualora fosse teoricamente possibile per una macchina riprodurre perfettamente i pensieri e le capacità mentali di un essere umano, non è affatto sicuro che si giungerà ad un avanzamento tecnologico tale da poter progettare e realizzare un computer e un software che realizzino concretamente questa possibilità. Allo stesso modo, anche qualora riuscissimo a dimostrare tramite un argomento formale nello stile di quello proposto da Godel o Lucas, che l'essere umano, considerato idealisticamente, possiede una qualche capacità di superare qualsiasi tentativo di descrizione della propria mente in termini puramente meccanicistici, da ciò non seguirebbe necessariamente che un essere umano in carne ed ossa, o l'insieme di tutti gli esseri umani viventi e vissuti, possa fare altrettanto, una volta prese in considerazione anche le inevitabili limitazioni fisiche e temporali a cui sono necessariamente sottoposti gli esseri viventi.

In questo capitolo proveremo ad analizzare il rapporto tra menti umane e intelligenze artificiali da un punto di vista non semplicemente formale e teorico, ma anche da una prospettiva che tenga in conto della dimensione più pratica e concreta del vissuto umano, cercando di stabilire se e fino a che punto si possa attribuire ad una macchina computazionale una capacità di pensiero simile alla nostra.

2.1 Il problema della simulazione

Per prima cosa è necessario specificare con precisione cosa si intende quando parliamo di "simulazione" effettiva del pensiero umano da parte di un computer. Questo compito, tuttavia, non è affatto semplice e porta con sé una serie non indifferente di problematiche. Innanzitutto, il concetto di "simulazione" prevede una relazione tra due oggetti: l'oggetto che viene simulato e che funge da modello e l'oggetto che lo imita. Nel caso della comparazione tra una mente e una macchina computazionale, l'oggetto simulato, ovvero la mente umana, è un concetto estremamente vago e difficilmente definibile con precisione. Per questo motivo, è anche difficile stabilire sia cosa esattamente un computer debba essere in grado di imitare per poter essere definito come una accettabile rappresentazione del pensiero di un essere umano, sia quali siano i criteri attraverso i quali giudicare questa comparazione.

Esistono a questo proposito due diversi tipi di approccio: adottando un punto di vista radicale si può affermare che una simulazione perfetta di una mente umana richiede che si possa stabilire una completa corrispondenza non solo tra l'input e l'output di uomini e macchine, ovvero che date delle informazioni iniziali entrambi giungano alla medesima risposta, ma che debbano essere identici anche i processi cognitivi interni che costituiscono il processo di elaborazione delle informazioni. Questa impostazione si può definire come *strutturale* e comporta una posizione di tipo "forte" nei confronti delle potenzialità dell'intelligenza artificiale: ovvero l'attribuzione ai dispositivi dotati di IA di capacità cognitive pari a quelle umane, nonché della possibilità di comprensione della semantica e anche di una vera e propria coscienza.

Il secondo tipo di approccio alla simulazione della mente umana è invece più moderato e prende il nome di *funzionale*. In questo caso, a differenza del precedente, i meccanismi "interni" della macchina e della mente non devono necessariamente coincidere, quello che importa è che ad un determinato input segua un medesimo output sia da parte della macchina sia della mente. Ora, quando si parla di riprodurre una medesima risposta, non si intende che il requisito per stabilire il successo della simulazione sia che un'intelligenza artificiale debba rispondere nello stesso identico modo dell'essere umano al quale la si sta comparando. Quello che si richiede è che l'intelligenza artificiale produca una risposta che si possa considerare entro i limiti di quello che generalmente si intende come un comportamento "intelligente" ascrivibile ad un essere umano, e che di conseguenza risulti, ad un osservatore esterno, indistinguibile da esso; si tratta, in altre parole, del superamento di un Test di Turing.

Tuttavia, come abbiamo già accennato nel capitolo precedente, il Test di Turing è stato concepito proprio con l'obbiettivo di cercare di sfuggire alle complicazioni derivanti dal dover dare una definizione precisa di concetti come "intelligenza", "pensiero" o "coscienza". Di conseguenza, sebbene il suo superamento possa senza dubbio essere considerato come una condizione necessaria per l'attribuzione dell'intelligenza ad una macchina (qualunque macchina che venga posta come dotata di "intelligenza umana" deve necessariamente essere in grado di superare questo test), esso non è tuttavia anche una condizione sufficiente. Si possono infatti immaginare delle ipotetiche macchine che siano in grado di superare questo test, senza tuttavia essere dotate di una qualsiasi forma di intelligenza. La macchina soggetta a questo test, infatti, deve essere in grado di rispondere a delle domande cercando di fornire delle risposte che sembrino quanto più possibili simili a quelle che fornirebbe un essere umano, tuttavia, l'esaminatore umano non ha conoscenza del modo con cui essa processa internamente le informazioni al fine di elaborare una risposta. Nel caso più estremo si potrebbe anche pensare ad una macchina nella cui memoria siano state immagazzinate un numero elevatissimo di domande e corrispondenti risposte; in questo caso il superamento del test di Turing sarebbe una semplice associazione meccanica di una risposta predeterminata ad una certa domanda, togliendo di conseguenza ogni necessità di una qualsiasi forma di intelligenza.

In aggiunta a questi problemi riguardanti l'approccio con il quale ci si pone nei confronti delle possibilità di simulazione del pensiero umano da parte dell'intelligenza artificiale, ci si potrebbe anche porre una questione ancora più fondamentale, ovvero se abbia davvero senso da un punto di vista pratico simulare artificialmente le capacità mentali di un essere umano. Abbiamo già indicato nel capitolo precedente come nel concreto le ricerche in questo campo non si spingano nella direzione della fedele riproduzione delle facoltà mentali umane da parte di una macchina, e che anzi, nella maggior parte dei casi, una tale fedeltà nella simulazione sarebbe non solo controproducente, ma anche potenzialmente pericolosa. In altre parole, generalmente, nelle applicazioni pratiche e reali dell'intelligenza artificiale, si cerca di progettare macchine dotate di "intelligenza" in senso ampio e non necessariamente di una forma di intelligenza che si basi e cerchi di imitare il modello di quella propria dell'essere umano.

Un'analogia spesso usata a questo proposito mette a confronto lo sviluppo dell'intelligenza artificiale con quello del volo artificiale. Lo studio del volo animale si è rivelato infatti infruttuoso ai fini dello sviluppo dei primi areoplani, mentre fondamentali sono state le scoperte scientifiche nel campo dell'aerodinamica: "direct imitation of natural flight proved a relatively fruitless avenue of resarch (…) working aircrafts were developed by

achieving greater understanding of the principles of aerodynamics"[100]. In questa prospettiva, un test sull'imitazione del volo concepito sul modello di quello proposto da Turing per l'intelligenza appare sostanzialmente inutile. Infatti, anche qualora fossimo in grado di costruire un animale artificiale che voli e sia indistinguibile da un uccello reale, ciò sarebbe di scarsa rilevanza ai fini dell'utilizzo pratico e concreto di macchine artificiali in grado di volare: "some of the purposes for which we use artificial flight, such as the speedy crossing of large distances, are similar to the purposes for which natural flight has evolved, but others, such as controlling the re-entry of spacecrafts, are radically different"[101].

Parallelamente, gli scopi e i settori in cui viene principalmente impiegata l'intelligenza artificiale richiedono un tipo di intelligenza non necessariamente di tipo "umano", ma un tipo di intelligenza intesa come il raggiungimento di uno scopo preciso utilizzando il metodo più efficiente possibile. In questo caso per "efficienza" si intende l'utilizzo del minor tempo e del minor numero di risorse possibili per raggiungere uno scopo, non necessariamente in termini assoluti, ma tenendo anche in considerazione una serie di criteri e di limitazioni atte a non causare dei danni sproporzionati rispetto ai benefici.

L'osservatore umano che funge da arbitro per il test di Turing, tuttavia, non giudica la macchina sotto esame in questi termini, ma utilizza metri di giudizio strettamente legati alla sua considerazione personale di ciò che può essere considerato come una forma di intelligenza umana. Paradossalmente, una macchina dotata di una forma di intelligenza superiore a quella di un essere umano potrebbe non essere in grado di passare questo test, poiché le risposte che fornisce, seppur precise e corrette, non "suonano" come proferite da un essere umano all'esaminatore; mentre una macchina dotata di un'intelligenza inferiore alla prima potrebbe fornire delle risposte scorrette, ma farlo in un modo che rispecchi il modo di parlare di un essere umano, ingannando quindi l'intervistatore. In altre parole, il test di Turing non testa il grado di "intelligenza assoluta" di una macchina, ma la sua capacità di imitare il modo con il quale gli esseri umani pensano, si esprimono e comunicano tra loro.

Di conseguenza, a meno che non si consideri ciò come condizione sufficiente per attribuire ad una macchina una forma di intelligenza umana, e l'intelligenza umana come il paradigma dell'intelligenza in generale, il test di Turing ci dice poco o nulla sulle reali capacità di intelligenza artificiale di un computer:

---

[100] Blay Whitby, *The Turing Test: AI's Biggest Blind Alley?*, in *Machines and Thought*, a cura di Peter Millican, Andy Clark, New York, Oxford University Press, 1996, pag 57

[101] *Ibidem*

> The imitation game proposed by Alan Turing provides a very powerful means of probing humanlike cognition. But when the test is actually used as a real test for intelligence, (…), its very strength becomes a weakness. Turing invented the imitation game only as a novel way of looking at the question "Can machine think?" But it turns out to be so powerful that it is really asking: "Can machines think exactly like human beings?"[102]

L'approccio di tipo simulativo del pensiero umano nel campo dell'intelligenza artificiale, sebbene non sembri produrre dei reali vantaggi in applicazioni pratiche, può invece risultare di grande interesse nello studio della mente umana e delle sue caratteristiche. In questo senso la domanda se le macchine possano pensare in modo simile al nostro o se almeno parzialmente esse siano in grado di simulare alcune delle nostre facoltà mentali, più che contribuire allo sviluppo di sempre più sofisticati sistemi di intelligenza artificiale, può aiutarci a capire meglio il funzionamento della nostra mente. Una macchina che sappia simulare alcuni tratti della nostra intelligenza, infatti, sebbene non necessariamente sia da considerarsi come un perfetto modello della mente umana, sicuramente ne offre un possibile modello interpretativo, ma questo non significa che questo modello interpretativo sia corretto o l'unico possibile:

> The fact that minds and machines both can recognize patterns, for example, no implies that minds are material, than the fact that birds and airplanes both flies implies that aeroplanes are oviparous. By reverse token, even if it were shown that certain forms of mentality cannot be simulated, it would nonetheless remain a possibility that the mind is material[103]

Inoltre, anche qualora un computer mostrasse un comportamento indistinguibile da quello umano, questa stessa identificazione, riscontrata da uno o più esseri umani, avrebbe inevitabilmente alla base un certo modo di vedere, ovvero richiederebbe una certa *interpretazione*. Questa interpretazione non può che essere basata su dei criteri o degli aspetti che coloro che giudicano ritengono essere fondativi del comportamento umano ed essenziali affinché si possa considerare come "umano" un determinato tipo di comportamento. Essa, di

[102] Robert M. French, *Subcognition and the Limits of the Turing Test*, in *Machines and Thought,* a cura di Peter Millican, Andy Clark, New York, Oxford University Press, 1996, pag. 26
[103] Otto Neumaier, *A Wittgensteinian View of Artificial Intelligence*, in *Artificial Intelligence. The Case Against*, a cura di R. Born, Londra-Sydney, Crook Helm, 1986, pag. 160

conseguenza, sarebbe figlia di una certa visione e concezione del mondo e in particolare dell'umano. Non è per nulla scontato, tuttavia, né che questo tipo di visione sia condivisa da tutti gli esseri umani, e di conseguenza che sia solamente propria di una determinata cultura, società e modo di pensare, né che essa rimanga inalterata nel tempo.

In questo senso, come vedremo meglio anche nel corso di questo capitolo, non è scontato che una attribuzione o non attribuzione di "intelligenza" ad una macchina computazionale, su base puramente comportamentale, ad esempio tramite il test di Turing, rimanga necessariamente inalterata nel tempo o sia unanimemente condivisa dall'intero genere umano. Una stessa macchina, considerata oggi come non intelligente, potrebbe invece, in un futuro imprecisato, complice un cambiamento culturale e una ridefinizione dei concetti di "intelligenza", ma anche semplicemente di cosa sia da considerarsi come "umano", essere considerata come tale. Bisogna sempre ricordarsi che un test come quello proposto da Turing non si basa su dei criteri oggettivi scientificamente misurabili e inalterabili, ma sul giudizio dell'essere umano che funge da arbitro al test. E l'insieme dei criteri su cui si basa questo giudizio non può che essere profondamente legato e prodotto dall'insieme dei valori culturali e sociali in cui quell'essere umano è nato, cresciuto e vissuto. In altre parole, per usare un'espressione husserliana, si basa sul suo "mondo della vita", sulla sua particolare visione del mondo, e come tale esso è particolare e contingente, e di conseguenza possibilmente soggetto a cambiamento.

2.2 Mente, Pensiero ed Intelligenza

È giunto ora il momento di dare, o meglio provare a dare, una definizione più precisa di alcuni dei concetti centrali per questa tesi, quali ad esempio "mente", "intelligenza" e "pensiero". Ora, questi termini sembrano a prima vista esprimere dei concetti a noi molto familiari, il cui significato potrebbe risultare quasi scontato e triviale da definire. Tuttavia, come vedremo, essi pongono in realtà delle difficoltà non indifferenti, o forse insormontabili, quando si tratta di definire con precisione cosa si intenda con essi e a quali enti essi si riferiscano.

Ludwig Wittgenstein, tra i massimi filosofi del Novecento, all'interno della sua vasta produzione ha trattato anche di questi termini psicologici, soffermandosi sia su cosa essi indichino, sia sul loro utilizzo all'interno del linguaggio. Queste riflessioni, sebbene non direttamente riferite alla loro applicazione nel campo dell'intelligenza artificiale[104], sono

[104] Il termine stesso "intelligenza artificiale" gli è postumo, anche se le prime riflessioni sulle possibilità di costruire delle macchine computazionali intelligenti, tra cui il celeberrimo *Computing Machinery and*

estremamente significative per l'argomento di discussione di questa tesi, e possono forse costituire degli argomenti a sostegno dell'impossibilità di riprodurre artificialmente l'"intelligenza umana".

Prima di concentrarci nello specifico su cosa significhino di preciso questi termini per Wittgenstein, è opportuno riassumere brevemente la sua posizione filosofica generale in relazione alla psicologia e ai termini psicologici. Semplificando forse troppo, si potrebbe affermare che la filosofia della psicologia di Wittgenstein può essere definita come "explanatory mentalism"[105], una particolare forma di mentalismo[106] che sostiene che la teorie psicologiche sono dei costrutti che possono solamente postulare l'esistenza di processi mentali interiori come potenziale causa dei comportamenti umani esteriori ed empiricamente osservabili, ma mai spiegarli nel modo nel quale, ad esempio, una teoria fisica può descrivere il moto dei corpi. Risulterebbe dunque impossibile provare in modo oggettivo ed assoluto la relazione tra una determinata teoria psicologica e il comportamento umano del quale si suppone essa sia causa.

Da questo si può dedurre come per Wittgenstein non sia possibile attribuire in modo oggettivo e inequivocabile nemmeno ad altri esseri umani termini psicologici che si riferiscono a supposti processi psicologici e mentali interiori, quali ad esempio "dolore", "coscienza" o "intenzionalità". Tutto quello che possiamo affermare è che osserviamo dei fatti e dei comportamenti esteriori negli altri esseri umani e supporre che esistano dei processi interni ad essi che possano fornire una descrizione per questi comportamenti.

Ma secondo quali criteri? Gli esseri umani, fin da quando sono piccoli, sono abituati ad interpretare certi comportamenti, espressioni del volto ed emissioni sonore come delle prove per processi a loro interiori. Ad esempio, siamo soliti attribuire una sensazione di dolore come causa di un grido improvviso o di una certa espressione; ovvero, spieghiamo questo improvviso comportamento supponendo che sia causato da un certo processo interiore. Con l'esperienza e la reiterazione delle medesime situazioni e assunzioni riteniamo di essere giustificati ad assumere che esistano questi processi interiori. In realtà, non potremo mai avere la prova certa che una persona che agisce in un determinato modo stia effettivamente avendo allo stesso tempo una determinata sensazione interiore, ad esempio quella di dolore, o se non stia semplicemente

Intelligence di Alan Turing, furono pubblicate quando Wittgenstein era ancora in vita. Lo stesso Alan Turing fu per un breve periodo un suo studente

[105] Cfr. Otto Neumaier, *A Wittgensteinian View of Artificial Intelligence*, in *Artificial Intelligence. The Case Against*, a cura di R. Born, Londra-Sydney, Crook Helm, 1986, pp. 132-173

[106] Il mentalismo è una corrente di pensiero che afferma che dall'osservazione di fenomeni esteriori, come ad esempio il comportamento umano, possiamo dedurre l'esistenza di processi mentali interni

fingendo di averla. Le prove che possiamo avere per l'esistenza di processi interiori negli altri esseri umani non sono dello stesso tipo di quelle che possiamo avere per l'esistenza di fenomeni fisici empiricamente osservabili; di conseguenza, le condizioni per l'uso dei termini psicologici sono differenti da quelle presenti in altri giochi linguistici.

Il significato dei termini psicologici, dunque, sta nel ruolo che essi hanno nella giustificazione e descrizione di certi fenomeni empiricamente osservabili, nello specifico del comportamento umano. Il linguaggio psicologico si propone di spiegare questi fenomeni assumendo che essi siano causati da certi processi interiori. Tuttavia, non saremo mai in grado di descrivere questi processi interiori o di definirli in alcun modo.

A questo proposito, nelle sue *Ricerche Filosofiche* Wittgenstein propone un celebre esempio: quello del coleottero nella scatola:

> Supponiamo che ciascuno abbia una scatola in cui c'è qualcosa che noi chiamiamo "coleottero". Nessuno può guardare nella scatola dell'altro; e ognuno dice di sapere che cos'è un coleottero soltanto guardando il *suo* coleottero. -Ma potrebbe ben darsi che ciascuno abbia nella sua scatola una cosa diversa. Si potrebbe addirittura immaginare che questa cosa mutasse continuamente.- Ma supponiamo che la parola "coleottero" avesse tuttavia un uso per queste persone!- Allora non sarebbe quello della designazione di una cosa. La cosa contenuta nella scatola non fa parte in nessun caso del giuoco linguistico; nemmeno come un *qualcosa*: infatti la scatola potrebbe anche essere vuota. -No, si può "dividere per" la cosa che è nella scatola; di qualunque cosa si tratti si annulla.[107]

Uscendo dalla metafora, possiamo intendere la scatola come il mondo interiore o la mente di ciascuno, e lo scarafaggio presente in essa come i processi interiori che supponiamo si svolgano al suo interno. A questo punto è evidente come, non potendo noi vedere all'interno delle scatole altrui, ovvero non potendo osservare empiricamente quello che avviene nelle supposte menti delle altre persone, non possiamo nemmeno sapere cosa ci sia al loro interno. Quello che possiamo fare è estendere per analogia anche agli altri quello che per analisi interiore esperiamo della nostra personale mente e dei processi che in essa si svolgono. Ma proprio perché possiamo esperire solo i nostri, e non quelli altrui, non potremo mai usare termini psicologici come "coscienza", "pensiero", "intelligenza", "gioia" o "dolore" in termini

[107] Ludwig Wittgenstein, *Ricerche Filosofiche*, Torino, Einaudi, 1967, pag. 293

descrittivi, nello stesso modo in cui possiamo usare termini che indicano cose e oggetti fisici e naturali e noi esterni, come ad esempio "roccia" o "albero".

Considerare i termini psicologici in maniera descrittiva non può che portare ad un non senso: "Thus, psychological language games are never descriptive, but (at least, some of them) have only their explanatory character in common"[108]. In altre parole, non ha senso usare i giochi linguistici di tipo psicologico in maniera descrittiva, mentre in alcuni casi ha senso usarli in senso esplicativo per il comportamento umano che possiamo osservare.

Questi giochi linguistici psicologici dipendono direttamente dalla *forma di vita* in cui coloro che li usano sono nati e cresciuti. Noi, infatti, impariamo ad associare alcuni particolari comportamenti degli altri esseri umani, alcune loro espressioni o emissioni sonore come segni di un qualche supposto stato interiore a cui associamo un nome, come ad esempio "felicità". Nel fare questo però cogliamo degli aspetti che sono strettamente connessi ed essenziali alla nostra specifica forma di vita. Altre forme di vita potrebbero cogliere aspetti molto diversi dai nostri perché più importanti per la loro forma di vita, e di conseguenza sviluppare altri giochi linguistici psicologici differenti dai nostri, seppur basati sugli stessi comportamenti osservabili.

A questo punto possiamo comprendere cosa intenda realmente Wittgenstein quando nelle sue *Ricerche Filosofiche* afferma: "Ma una macchina non può certo pensare! (…) Solo dell'uomo, e di ciò che è ad esso simile, diciamo che pensa"[109]. Non si tratta in questo caso di una conclusione derivante da un ragionamento di tipo logico-formale come quella a cui è giunto Lucas alla fine del suo argomento che è stato esposto nel primo capitolo. Wittgenstein con queste parole non intende affermare una impossibilità di tipo logico-matematico che impedisca in termini assoluti ad una qualsiasi macchina computazionale di "pensare". Quello che Wittgenstein intende quando afferma che una macchina non può pensare è che: "it makes no sense to apply the notion of "thinking" as well as other psychological terms to computers"[110]. La domanda stessa se una macchina possa pensare è insensata, "L'enunciato: "Una macchina pensa (percepisce, desidera)" sembra in qualche modo assurdo. È come se avessimo domandato: "Ha il numero 3 un colore""[111].

L'insensatezza del chiedersi se le macchine possiedano o meno delle caratteristiche psicologiche simili o uguali a quelle umane deriva proprio dal fatto che tutti i giochi linguistici,

[108] Otto Neumaier, *A Wittgensteinian View of Artificial Intelligence*, in *Artificial Intelligence. The Case Against*, a cura di R. Born, Londra-Sydney, Crook Helm, 1986, pag. 149
[109] Ludwig Wittgenstein, *Ricerche Filosofiche*, Torino, Einaudi, 1967, pag. 360
[110] Otto Neumaier, *A Wittgensteinian View of Artificial Intelligence*, in *Artificial Intelligence. The Case Against*, a cura di R. Born, Londra-Sydney, Crook Helm, 1986, pag. 151
[111] Ludwig Wittgenstein, *Libro Blu e Libro Marrone*, Torino, Einaudi, 1983, pag. 66

non solamente quelli psicologici, sono prodotti e indissolubilmente legati alla forma di vita umana. In questo senso, essi sono "parte della nostra storia naturale come il camminare, il mangiare, il bere, il giocare"[112]. Da questo punto di vista, avrebbe senso chiedersi se un qualcosa di altro da noi abbia degli stati psicologici interiori, quali ad esempio il pensiero, se e solo se, non solamente esso svolgesse nel nostro stesso modo alcune specifiche azioni, ad esempio il rispondere a delle domande o lo scrivere un testo, ma anche se si comportasse *totalmente* come abbiamo imparato gli esseri umani si comportano: "the ascription of psychological states is necessarily connected with the human *form of life*; (…) in other words: we learn to use psychological terms in such a way that their meaning depends upon the total complex of human behaviour"[113].

Ora, anche se il celeberrimo articolo *Computing Machinery and Intelligence* di Alan Turing, nel quale per la prima volta venne proposto quello che verrà chiamato "Test di Turing", venne pubblicato nell'anno della morte di Wittgenstein, e di conseguenza queste riflessioni wittgensteiniane sull'uso dei termini psicologici sono ad esso precedenti, non si può che considerarle come un argomento contro l'attribuzione di intelligenza ad una macchina sulla base di questo test. Come abbiamo già illustrato, infatti, il Test di Turing prevede una interazione uomo macchina solamente testuale: la macchina che viene testata non possiede un corpo fisico con fattezze umane, non si muove, non parla e soprattutto non è nata e cresciuta in un ambiente umano. In altre parole, non fa parte della forma di vita umana. Per questo motivo, anche qualora le risposte fornite da questa macchina fossero in tutto e per tutto indistinguibili da quelle date da un essere umano, essa non potrà mai essere considerata come intelligente nello stesso modo in cui possiamo considerare intelligente un altro essere umano. Anzi, lo stesso chiedersi se una tale macchina sia intelligente non avrebbe nemmeno senso. Turing, cercando con il suo test di evitare le inevitabili e intrinseche difficoltà nel definire cosa sia l'intelligenza, e spostando il centro dell'attenzione solo su una minima parte dell'insieme del comportamento umano, ovvero nella capacità di rispondere testualmente a delle domande, avrebbe in realtà reso priva di senso la stessa domanda alla quale il suo test cerca di fornire una risposta.

Ritornando a quanto detto qualche paragrafo sopra, il nostro utilizzo dei termini psicologici e la loro attribuzione ad altri esseri umani si basa su criteri esteriori, su delle prove empiricamente osservabili che però non sono equivalenti ai criteri che possiamo avere per attribuire determinate caratteristiche a degli oggetti naturali. In altre parole, non sono evidenze

[112] Ludwig Wittgenstein, *Ricerche Filosofiche*, Torino, Einaudi, 1967, pag. 23
[113] Otto Neumaier, *A Wittgensteinian View of Artificial Intelligence*, in *Artificial Intelligence. The Case Against*, a cura di R. Born, Londra-Sydney, Crook Helm, 1986, pag. 152

di tipo scientifico, ma prove che si basano sulla nostra esperienza e comprensione del complesso della forma di vita umana.

Anche se esperienzialmente potremmo pensare che le nostre ipotesi circa l'attribuzione di stati interiori negli altri esseri umani sono confermate, questo tipo di conferma non è la stessa che ricaviamo quando ad esempio osserviamo un certo fenomeno naturale e formuliamo delle ipotesi per stabilirne la causa. Tuttavia, nonostante le evidenze che possediamo per l'esistenza degli stati psicologici non siano le stesse che abbiamo per l'esistenza dei fenomeni naturali, questa loro somiglianza, soprattutto quando confermata dalla nostra esperienza nella vita di tutti i giorni, può facilmente portarci a considerarle alla stessa stregua e a ritenere, di conseguenza, che tali processi psicologici interiori abbiano una propria esistenza e siano in tutto e per tutto degli enti indipendenti dagli esseri umani.

Se, infatti, termini quali "intelligenza", "pensiero", "coscienza" avessero una propria esistenza indipendente dalla forma di vita umana, e fossero scientificamente provabili come tali, allora il chiedersi se una macchina possa pensare avrebbe perfettamente senso, e un test come quello di Turing sarebbe senz'altro una valida possibilità per stabilire una risposta. Tuttavia, almeno secondo Wittgenstein, non è assolutamente possibile, anzi è totalmente privo di senso, sia astrarre questi concetti dalla forma di vita umana, sia considerarli come reali cause del comportamento umano.

Come abbiamo già indicato, inoltre, questi termini psicologici non possiedono alcun carattere descrittivo, ma solamente esplicativo; in altre parole, non è possibile definirli nello stesso modo in cui definiamo e descriviamo gli oggetti naturali. Non possiamo, per esempio, attribuire loro delle caratteristiche specifiche, come peso, forma, colore o dimensione, al pari di quanto possiamo fare con questi ultimi.

La caratteristica fondamentale dei termini psicologici è perciò proprio il loro essere indefiniti. Il complesso della forma di vita umana e di conseguenza anche lo stesso comportamento umano, che di essa fa parte, è troppo complesso e variabile per essere definito in modo univoco "variability itself is a characteristic of behaviour without which behaviour would be to us something completetely different"[114]. La nostra possibilità di ascrivere processi interiori agli altri esseri umani dipende, paradossalmente, proprio da questa stessa impossibilità di descrivere la complessità e mutevolezza del comportamento umano.

[114] Otto Neumaier, *A Wittgensteinian View of Artificial Intelligence*, in *Artificial Intelligence. The Case Against*, a cura di R. Born, Londra-Sydney, Crook Helm, 1986, pag. 151

Ora, a differenza degli esseri umani, i programmi di intelligenza artificiale[115] sono per noi completamente aperti e trasparenti. Ovvero, conosciamo benissimo e possiamo addirittura leggere, scrivere e modificare il loro codice; in altre parole, sappiamo alla perfezione tutto ciò che sta alla base del comportamento della macchina che li esegue, e in questo senso possiamo giustamente affermare che il programma è effettivamente la causa descrittiva del suo comportamento. I processi interiori della macchina sono perciò radicalmente diversi da quelli che supponiamo essere presenti negli esseri umani. I termini psicologici che noi umani usiamo nei giochi linguistici psicologici derivano da una particolare forma di vita che ha la caratterista fondamentale di essere indefinita, continuamente mutevole e soprattutto opaca e non conoscibile. Non abbiamo, di conseguenza, alcun bisogno di applicare tali termini a delle macchine controllate da programmi ben definiti, comprensibili e perfettamente trasparenti.

Possiamo quindi affermare categoricamente che, in una concezione dell'umano e del linguaggio wittgensteiniana, non ci sia veramente alcuno spazio per l'intelligenza artificiale o per computer che pensano? La risposta in apparenza sembrerebbe affermativa, in quanto essi "do not share our human context, what Wittgenstein termed our *forms of life*. For example, computers are not interested in food, humour, or companionship; nor are they hurt when hit, or sympathetically prompted to go to the aid of someone who is"[116].

Ma è proprio vero che i computer non fanno parte della nostra forma di vita? O, più generalmente, possiamo affermare che essi non facciano parte di alcuna forma di vita?

Proviamo a rispondere alla prima di queste due domande. Noi, in quanto esseri umani, possiamo comprendere solamente ciò che fa parte e condivide la nostra stessa forma di vita, sia esso una macchina, o anche più semplicemente un altro essere umano. Ora, è ovvio che i computer, non condividendo con noi interessi, sentimenti, relazioni, non fanno parte della forma di vita umana nello stesso modo in cui ne fanno parte gli esseri umani in carne ed ossa. Tuttavia, allo stesso tempo è anche vero che è possibile programmare un computer in modo che sia in grado di simulare almeno alcuni tratti della nostra intelligenza. In questo senso, essi non sono completamente incomprensibili per noi.

In secondo luogo, i computer ormai sono parte integrante della nostra forma di vita per il fatto che sono sempre più presenti e ricoprono un ruolo sempre maggiore nelle nostre vite. Questo progressivo e sempre maggiore ingresso dei computer nella nostra forma di vita rende

[115] Da intendersi qui come i programmi di intelligenza artificiale classici, ovvero composti da linguaggio di programmazione scritto da esseri umani. Il nuovo approccio all'intelligenza artificiale basato sulle reti neurali verrà trattato nel terzo capitolo della tesi

[116] Otto Neumaier, *A Wittgensteinian View of Artificial Intelligence*, in *Artificial Intelligence. The Case Against*, a cura di R. Born, Londra-Sydney, Crook Helm, 1986, pag. 152

sempre più facile riuscire a comprenderli. La comprensione tra forme di vita diverse, infatti, è più semplice a seconda delle similitudini e della quantità di rapporti tra esse, anche se non è definibile con precisione *quanto* una determinata forma di vita debba essere simile ad un'altra affinché sia possibile una comprensione tra le due: "it isn't clear *how much* has to be similar for us to have a right to apply (…) the concept "thinking", which has its home in *our* life"[117]. In questo senso, quanto più i computer faranno parte della nostra forma di vita, tanto più saremo in grado di comprenderli. Un cambiamento della nostra forma di vita, causato dal mutare delle circostanze e dal sempre maggiore utilizzo dei mezzi tecnologici, potrebbe cambiare anche il nostro stesso modo di intendere i computer e, di conseguenza, in un futuro imprecisato, potrebbe anche essere sensato attribuire anche ad essi dei termini di tipo psicologico:

> We might conclude that Wittgenstein's views on the (im-) possibility of "thinking" machines depend on the background of *his* life, whereas the application of psychological terms to computers depends upon the altered conditions of *our* modern life; it is, then, only a matter of empirical, technical progress that we are successively enabled to apply our common psychological terms to computers.[118]

Se un cambiamento nella forma di vita può portare anche ad un cambiamento nei criteri di utilizzo delle parole all'interno dei giochi linguistici, allora, di conseguenza, potrebbe divenire sensato affermare che anche le macchine possano "pensare" o essere "intelligenti": "in relation to Wittgenstein's "form of life", it would be abnormal to ascribe the ability of "thinking" to computers, but this cannot prevent some to imagine another "form of life", the world of AI, where this case becomes *normal*"[119]. All'interno della forma di vita di cui faceva parte Wittgenstein, dunque, il fatto che una macchina possa "pensare" non può che essere un qualcosa di insensato, ma non è escluso in nessun modo che in altre forme di vita ciò sia perfettamente sensato.

Detto questo, rimane valido il fatto che, almeno nel senso tradizionale del termine "intelligenza", non è possibile definire intelligenti i computer. Essi, infatti, non esibiscono alcun comportamento che possiamo definire come intelligente nel senso tradizionale e comune del termine, né solitamente consideriamo come intelligente un essere umano solamente sulla base della sua capacità di eseguire compiti esattamente nello stesso modo in cui lo farebbe un

[117] *Ivi*, pag. 157
[118] *Ibidem*
[119] *Ivi*, pag. 158

computer. Se un giorno cambieranno i criteri sulla base dei quali noi definiamo termini psicologici come "intelligenza", "pensiero" e "coscienza" allora sarà possibile applicarli, in una prospettiva linguistica wittgensteiniana, anche a delle macchine computazionali.

Tornando al punto di partenza di questa discussione, il motivo per cui, almeno secondo Wittgenstein, facciamo uso dei termini psicologici è per spiegarci il motivo di un certo comportamento, empiricamente osservabile da noi, esibito da altri esseri umani. Siccome non possiamo in nessun modo "guardare" dentro la testa delle altre persone per vedere cosa accade in essa, siamo costretti ad assumere che esistano dei determinati stati mentali che ne sono la causa. Possiamo solamente, al massimo, misurare le reazioni fisiologiche e neurologiche concomitanti a determinati comportamenti, ma esse non corrispondono a quello che intendiamo quando usiamo termini psicologici quali ad esempio "dolore" o "gioia".

Ora, a differenza di quanto accade con le persone, noi siamo perfettamente in grado di "guardare" dentro un computer, sia a livello fisico (possiamo, ad esempio, misurare fisicamente i passaggi di corrente elettrica nei circuiti dei suoi componenti), ma anche e soprattutto a livello di programma, ovvero di ciò che determina il comportamento stesso della macchina computazionale. Per questa ragione, e in luce di quanto detto sopra riguardo la genesi e l'utilizzo dei termini psicologici, in una prospettiva wittgensteiniana non abbiamo alcun bisogno di attribuire ad essi degli stati psicologici interiori. Noi, infatti, non solo sappiamo con esattezza quali sono le cause del funzionamento di un computer, ma siamo anche in grado di prevederlo con ragionevole certezza (al di là della possibilità di errori non previsti nel codice o nella componente fisica della macchina che producano risultati diversi da quelli aspettati).

Anche qualora fosse possibile costruire un automa in tutto e per tutto identico ad un essere umano, ed esso esibisse comportamenti indistinguibili da quelli umani (ad esempio una reazione di dolore in seguito ad un colpo subito), non avremmo bisogno di attribuirgli alcuno stato interiore di dolore; potremmo semplicemente andare a leggere il suo programma e scoprire che quella reazione è semplicemente un certo comportamento codificato in un certo linguaggio di programmazione da un programmatore umano. La causa di una reazione di dolore esibita da un automa di questo tipo è per noi perfettamente conoscibile.

Un discorso analogo possiamo farlo anche per l'attribuzione dell'"intelligenza" ad un automa. Ovvero, anche qualora esso rispondesse perfettamente alle domande del test di Turing e riuscisse ad ingannare l'esaminatore umano, non avremmo bisogno di assumere che la causa di ciò sia una qualche forma di "intelligenza" analoga a quella umana. Infatti, andando a leggere il suo codice di programmazione, scopriremmo il modo in cui esso è stato programmato al fine di poter rispondere "umanamente" a determinate domande. In questo senso, non possiamo usare

né la componente fisica del computer, né il suo programma come prove esteriori di possibili stati psicologici interiori simili o analoghi a quelli che invece attribuiamo agli altri esseri umani: "This strongly indicates that programs too represent no appropriate outward evidence for mentalistic assumptions within AI"[120].

Questo argomento fa riferimento ad una concezione classica sia della programmazione che dell'intelligenza artificiale, in cui il programma è concepito come una vera e propria trasposizione esplicita in termini logici di quello che si vuole rappresentare, o far eseguire ad una macchina, e codificata in un certo linguaggio di programmazione. Questi linguaggi non sono ovviamente dei linguaggi naturali, ma una volta appresi sono perfettamente comprensibili per gli esseri umani.

Lo stesso non si può però affermare dei programmi che stanno alla base dei più recenti approcci nel campo dell'intelligenza artificiale, come ad esempio le reti neurali di cui tratteremo nel prossimo capitolo, che non risultano trasparenti e comprensibili nello stesso modo nel quale lo sono i programmi "classici". In altre parole, in alcuni casi questi nuovi programmi producono dei risultati che nemmeno coloro che li hanno scritti e progettati sono in grado di giustificare con esattezza, e si potrebbe quindi ipotizzare un utilizzo giustificato di termini analoghi a quelli che siamo soliti utilizzare per giustificare il comportamento umano, anche se ovviamente non sarebbero gli stessi. Lasciando questa discussione ad una più approfondita disamina nel corso del terzo capitolo, proviamo a tirare le somme di questa interpretazione in chiave wittgensteiniana delle possibilità dell'intelligenza artificiale.

Innanzitutto, possiamo affermare che quando parliamo di simulazione da parte di un computer dell'intelligenza umana non stiamo parlando dell'intera intelligenza umana, composta da innumerevoli e difficilmente definibili sfaccettature, ma di alcune e ben specifiche sue caratteristiche (ad esempio analisi e classificazione di dati, uno degli ambiti in cui effettivamente oggi è più utilizzata l'intelligenza artificiale), generalmente quelle che sono più consone e utili a particolari scopi pratici.

In secondo luogo, quando associamo la parola "intelligenza" ad un computer, in realtà non stiamo parlando dello stesso tipo di intelligenza che siamo soliti attribuire ad altri esseri umani. Abbiamo infatti visto come l'utilizzo di questo genere di termini psicologici sia indissolubilmente e inequivocabilmente legato ad una particolare *forma di vita*, quella umana, di cui le macchine (ancora) non fanno parte e di conseguenza non possiamo attribuire loro questo genere di attributi.

[120] *Ivi*, pag. 164

In terza battuta, abbiamo mostrato come, almeno per i programmi di intelligenza artificiale "classici", non ci sia in realtà alcun bisogno di alcuna assunzione di tipo mentalistico per spiegarne il comportamento: "even if AI programs succeed (…) in "simulating" human abilities which would underly the performance of similar tasks, this does not prove that mentalistic assumptions are necessary for AI; we should therefore be (ontologically) more cautious and renounce the assumption of entities we do not need for explaining"[121].

Tutte queste argomentazioni dimostrano dunque inequivocabilmente che in una prospettiva wittgensteiniana l'intelligenza artificiale sia completamente impossibile? Non necessariamente, infatti "according to Wittgenstein, we may plausibly assume that our language use (and, hence, our total form of life) changes in such a way that computers, in our opinion, *can* "think""[122]. Quello che possiamo invece affermare con certezza è che se anche ciò accadesse, non si tratterebbe dello stesso tipo di "intelligenza" che intendiamo quando usiamo questo (e altri termini psicologici) in riferimento agli esseri umani e alla loro *forma di vita*.

2.3 Conoscenza Esplicita e Conoscenza Tacita

Una delle questioni fondamentali all'interno del dibattito sulle possibilità dell'intelligenza artificiale è sicuramente quella riguardante la completa formalizzazione del pensiero e della mente umani; ovvero della possibilità di tradurre in un linguaggio di tipo logico-formale l'insieme completo dei nostri pensieri, ragionamenti e comportamenti. Si tratta di una questione fondamentale perché, qualora ciò non fosse possibile, sarebbe anche necessariamente impossibile che una macchina computazionale possa simulare la totalità delle nostre facoltà cognitive, in quanto, come abbiamo già affermato, tutto il funzionamento di una macchina di questo tipo deve essere espresso esplicitamente in un linguaggio formale. Lucas, con il suo controverso argomento, aveva cercato di dimostrare logicamente che gli esseri umani possiedono almeno una capacità non codificabile, ovvero quella di "vedere" la verità delle proposizioni godeliane all'interno di determinati sistemi formali. Il suo argomento, tuttavia, è andato incontro ad una serie di problematicità a livello sia puramente logico che teoretico che rendono, a detta di molti, invalida la sua argomentazione. Proveremo ora a vedere se possiamo trovare nel contesto del pensiero umano, e al di fuori dell'ambito puramente logico-matematico, un qualcosa che sia totalmente o anche solo parzialmente non codificabile e dunque non riproducibile da un computer.

[121] *Ivi*, pag 166
[122] *Ibidem*

Quello della conoscenza è senza dubbio uno dei grandi temi e problemi della filosofia fin dalle sue origini, e parallelamente essa rappresenta anche una delle più grandi sfide a cui gli scienziati che si occupano di intelligenza artificiale si sono trovati a far fronte. Cosa sia la conoscenza, in che modo essa si ottiene e come la si può esprimere sono delle domande con cui chi si propone di realizzare un modello artificiale della mente umana è inevitabilmente costretto a confrontarsi. Se il cosa sia la conoscenza è forse una problematica più prettamente filosofica, il come la si ottenga e il come la si possa esprimere invece sono dei veri e propri ostacoli che in passato, e anche al giorno d'oggi, hanno posto dei seri problemi a coloro che si occupano di sistemi di intelligenza artificiale. Cominciamo a discutere dell'ultimo di questi problemi, ovvero di come possiamo esprimere la nostra conoscenza.

È indubbio che una gran parte della nostra conoscenza sia esprimibile in un qualche tipo di linguaggio, sia esso naturale o formale, e di conseguenza sia anche comunicabile. Possiamo chiamare questo tipo di conoscenza "esplicita". Fa parte di questo tipo di conoscenza tutto ciò che troviamo scritto in manuali, enciclopedie, media di vario genere, documenti. La conoscenza esplicita può essere espressa in molti modi, non solo verbalmente, ma anche in formato visivo, tramite illustrazioni o video. Questo tipo di conoscenza, essendo facilmente esprimibile e codificabile non pone alcun tipo di problema nel campo dell'intelligenza artificiale o della computer science in generale, almeno per quanto riguarda la sua formalizzazione.

L'altro tipo di conoscenza, che risulta invece molto più problematica da esprimere, possiamo definirla come "tacita" o "implicita". Questo tipo di conoscenza, a differenza della prima, non è solitamente espressa in forme esplicite e la sua esplicitazione è senza dubbio difficoltosa, se non addirittura impossibile. Possiamo definire come forme di conoscenza implicita la saggezza o sapienza, l'esperienza, il saper fare e l'intuizione. Non essendo comunemente espressa esplicitamente, questo tipo di conoscenza può essere appresa solo nella pratica, oppure essere innata: saper suonare uno strumento musicale, saper guidare una macchina, usare un linguaggio naturale sono solo alcuni di innumerevoli esempi di conoscenza tacita. È un tipo di conoscenza che implica necessariamente una dimensione sociale, interpersonale e soprattutto esperienziale per essere appresa.

L'espressione "conoscenza tacita" fu usata per la prima volta da Michael Polanyi nel suo libro del 1958 *Personal Knowledge*[123], e al giorno d'oggi essa viene suddivisa in tre differenti categorie a seconda delle sue caratteristiche e possibilità di essere codificata. Il primo tipo viene definito come "conoscenza tacita relazionale", è un tipo di conoscenza che potenzialmente

[123] Michael Polanyi, *Personal Knowledge*, Chicago University Press, 1974

potrebbe essere resa esplicita, ma che per ragioni sociali, relazionali e contingenti viene mantenuta implicita. È tacita non tanto per fattori intrinsechi alla sua particolare natura ma piuttosto per una più o meno deliberata e conscia scelta di coloro che la posseggono. È costituita, dunque, da segreti od omissioni di informazioni causate da deliberate scelte di ragione politica, sociale o relazionale, ma anche più semplicemente da ragioni contingenti di spazio e tempo. Tolti questi ostacoli, non si presenta alcun problema sostanziale che ne impedisca la codificazione, dunque, questo tipo di conoscenza tacita non presenta particolari complicazioni nel campo dell'intelligenza artificiale.

Il secondo tipo di conoscenza tacita viene definita "conoscenza tacita somatica" e comprende tutte quelle azioni e movimenti che svolgiamo quotidianamente con il nostro corpo in modo inconscio o automatico, e che non necessitano di un razionale controllo. Suonare uno strumento musicale, camminare, guidare la macchina e innumerevoli altri compiti che svolgiamo quotidianamente sono esempi di conoscenza tacita somatica. Questo tipo di conoscenza, sebbene solitamente non venga codificata o espressa esplicitamente dagli esseri umani, sia per ragioni di complessità, che per il fatto che nella maggior parte dei casi una sua codificazione sarebbe inutile o superflua, è potenzialmente esprimibile e codificabile. La difficoltà della sua esplicitazione risiede nel fatto che risulta difficile per noi esprimere in un linguaggio naturale o formale come il nostro cervello diriga e controlli i movimenti corporei. Nonostante ciò, non solo è in linea di principio potenzialmente possibile farlo, ma in molti casi una sua esplicitazione è già stata realizzata. Sono infatti molti gli esempi di automi di forma umana o animale che si muovono e compiono azioni in modo molto simile ai viventi sui quali essi vengono modellati.

Il terzo e ultimo tipo di conoscenza tacita viene definito "conoscenza tacita collettiva" ed è la tipologia di conoscenza tacita che pone i più grandi ostacoli nella rappresentazione e codificazione della conoscenza umana. Si tratta infatti di un tipo di conoscenza che è situata nella società umana e che può essere appresa solamente essendo immersi in un determinato sostrato sociale. Si può definirla come: "the ability to absorb ways of going on from the surrounding society without being able to articulate rules in detail"[124]. Lo stesso corpo umano ricopre un ruolo fondamentale nell'acquisizione di questo tipo di conoscenza tacita. Esempi di questo tipo di conoscenza sono: il parlare un linguaggio naturale, gestire relazioni di potere tra individui, saper come comportarsi in determinate situazioni sociali. È impossibile da codificare

[124] Louis Sanzogni, Gustavo Guzman, Peter Busch, *Artificial Intelligence and Knowledge Management: Questioning the Tacit Dimension*, "Prometheus", 35 (2017), pag. 43

perché l'ambiente sociale e umano nel quale essa si sviluppa non può essere semplicemente ridotto ad un insieme di regole e principi che lo regolano a causa della sua complessità, multiformità, vaghezza, astrattezza e mutevolezza. Si tratta di un tipo di conoscenza che nasce ed è esclusiva di un particolare "mondo della vita" e può essere appresa solamente da coloro che ne fanno parte. Anche qualora fosse possibile esprimerla, sarebbe impossibile da comprendere per chiunque che sia esterno ad esso.

Ricapitolando, possiamo affermare che la conoscenza esplicita è compatibile con un approccio di tipo oggettivistico alla natura della conoscenza stessa. Un approccio di questo tipo implica che tutta la conoscenza sia una realtà oggettiva che può essere definita, misurata, codificata e trasferita, e di conseguenza non pone limitazioni a ciò che può essere rappresentato ed elaborato all'interno di una macchina computazionale. Se la conoscenza fosse tutta di questo tipo, essa non rappresenterebbe alcun ostacolo nello sviluppo dell'intelligenza artificiale. Tuttavia, questo tipo di concezione della conoscenza non può che escludere gli aspetti contingenti, mutevoli e soggettivi della conoscenza stessa.

Un altro tipo di concezione della conoscenza può essere considerato come "interpretativo". In questo caso vengono sottolineati e messi in luce gli aspetti più taciti e intangibili della conoscenza. Questo tipo di approccio tiene in conto della componente tacita della conoscenza e del suo essere personale, relazionale, socialmente costruita e comporta inevitabilmente un certo grado di incertezza, diversità interpretativa e ambiguità. Non viene messa in dubbio la possibilità di codificare e comunicare la conoscenza esplicita e parte di quella tacita, anche se in diversi gradi, ma una chiara e oggettiva trasposizione della conoscenza tacita in toto è impossibile.

Infine, possiamo definire come approccio "pratico" una concezione della conoscenza che ne mette in luce il suo aspetto più radicato nella pratica e nell'azione umana. Secondo questa prospettiva una delle caratteristiche fondamentali e fondanti della conoscenza è il suo essere radicata nella prassi e nel comportamento umano. Gli esseri umani spesso sono incapaci di spiegare formalmente un gran numero di attività che svolgono, anche coscientemente, sia a causa del fatto che il tradurre queste attività pratiche in un qualche linguaggio formale, o anche semplicemente naturale, è un compito molto arduo, sia perché sono attività che essi stessi hanno imparato nella pratica e non ne hanno mai ricevuto una spiegazione formale. Possiamo, inoltre, anche essere inconsciamente ignari di qualcosa che sappiamo.

Secondo la prospettiva "pratica" della conoscenza, esiste una parte di conoscenza tacita che non può essere formalizzata, e di conseguenza che non può essere nemmeno codificata e replicata da una macchina computazionale. Questo perché parte della nostra conoscenza è

contingente, ovvero viene prodotta e appresa in particolari situazioni e contesti che possono mutare nel tempo; possiede inoltre un carattere di tipo relazionale e sociale le cui regole non possono essere codificate, anzi spesso gli individui in certi casi "need to break rules in order to adapt performing actions to local conditions of operation"[125].

Un altro aspetto della conoscenza che viene tenuto in considerazione dall'approccio pratico, e che non è né codificabile né replicabile da una macchina, è quello più strettamente personale e individuale, ovvero il fatto che la conoscenza personale di un individuo è influenzata da fattori come pregiudizi e cultura personale, ma anche da emozioni, sentimenti e dalla sua personalità. Questa dimensione privata e personale della conoscenza non è ovviamente replicabile o codificabile, né da una macchina né da un altro essere umano e non si possono trovare regole o principi che la regolino.

Appare di conseguenza evidente come gli approcci interpretativo e pratico allo studio della conoscenza "help in appreciating the limitations of state-of-the-art AI (…) since there is recognition of the complex and multifaceted nature of tacit knowledge"[126].

Questi diversi tipi di conoscenza, il loro essere più o meno codificabili ed esplicitabili, uniti alla natura multiforme, mutevole e contingente degli esseri umani e del mondo in cui viviamo, contribuiscono a porre dei limiti a ciò che è possibile realizzare nel campo dell'intelligenza artificiale, almeno nei suoi aspetti che riguardano lo sviluppo di modelli cognitivi computazionali che si propongono di imitare o simulare il funzionamento della mente umana.

Quello che i sistemi di intelligenza artificiale possono fare è "store articulated rules and apply these to increasingly complicated situations"[127], dunque finché si tratta di articolare e codificare conoscenze di tipo esplicito e alcuni tipi di conoscenze tacite che abbiamo sopraindicato, non sussistono particolari problemi di tipo epistemologico che ostacolino il progresso dell'intelligenza artificiale. Si possono immaginare anche situazioni di carattere sociale o pratico, particolarmente delineate e controllate e in cui sia possibile definire delle ben precise regole di comportamento, che rendano possibile delle parziali e limitate applicazioni di sistemi di intelligenza artificiale anche nel campo della conoscenza tacita di tipo relazionale e pratico. Tuttavia, appena si esce da questi ambienti controllati, i sistemi di intelligenza artificiale devono fare i conti con la complessità e la imprevedibilità della dimensione

[125] *Ivi*, pag. 42
[126] *Ibidem*
[127] *Ivi*, pag. 43

relazionale e sociale umana: “machines cannot socialise or be meaningfully embedded in a social milieu since machines are different in kind and materially”[128].

Il comportarsi e interagire in ambienti sociali e umani richiede non solamente la comprensione di una serie di regole di condotta e di pratiche che non possono essere formalizzate e comprese se non dagli stessi esseri umani che ne fanno parte, ma anche e soprattutto una certa flessibilità e adattabilità nell’applicazione di queste regole medesime. Anche qualora fosse possibile codificare in maniera esaustiva un ipotetico insieme di norme comportamentali che regolino il comportamento umano e le interazioni sociali nella loro interezza, non sarebbe comunque possibile ricavare in modo deduttivo da esse tutte le possibili risposte, comportamenti e reazioni che un essere umano può esibire.

Il mondo nel quale ci troviamo a vivere, infatti, è in costante cambiamento, evoluzione e trasformazione, e di conseguenza le stesse regole sociali e comportamentali di coloro che lo abitano sono in costante aggiornamento. Ogni situazione è particolare e contingente e richiede un certo grado di adattamento e mutamento delle stesse regole non espresse che gli esseri umani seguono, molto spesso inconsciamente, nella loro vita di tutti i giorni. Un programma di intelligenza artificiale, tuttavia, altro non è che una (lunghissima) lista di proposizioni ben definite e codificate in un certo linguaggio di programmazione che ne regolano il comportamento. Di conseguenza, un tale programma, qualora fosse progettato in modo da essere un modello di tipo computazionale della coscienza e delle abilità intellettive umane, dovrebbe non solamente racchiudere tutta le conoscenze e le regole di comportamento che si presuppone scandiscano la vita umana, ma anche essere in grado di aggiornarle e modificarle costantemente. In altre parole, anche prescindendo dalle effettive possibilità di esplicitazione della conoscenza umana, non è possibile formalizzare in un sistema formale l’insieme del “mondo della vita” degli esseri umani a causa del suo essere in continuo divenire. Ogni formalizzazione non sarebbe altro che un’istantanea di un mondo in continua trasformazione, una immagine formale di un mondo che però ormai è già passato, ed è già divenuto altro.

In questo senso, dunque, si potrebbe interpretare quella capacità che Lucas attribuiva agli esseri umani di “andare oltre” qualsiasi sistema formale ben definito, come la capacità tipicamente umana sia di adattarsi alle situazioni sempre diverse che di volta in volta ci si presentano in un mondo in continua evoluzione, sia di essere parte integrante e motore di questo cambiamento. Il ridefinire costantemente regole e metodi comportamentali, ma anche la capacità di ignorarli e produrne di nuovi che si adattino meglio alla situazione contingente in

[128] *Ibidem*

cui ci si trova, è una capacità che i sistemi di intelligenza artificiale ancora non sono in grado di attuare nello stesso modo in cui lo fanno gli esseri umani. Anche qualora queste regole potessero essere considerate come oggettive e immutabili, bisognerebbe comunque tenere in conto il fatto che ogni essere umano è particolare, possiede una sua propria personalità, modo di intendere, comprendere e vedere il mondo, che sono influenzati non solamente dal suo sottofondo culturale e sociale di provenienza, ma anche dalle sue emozioni, sentimenti e modi di sentire, talvolta contingenti e transitori, e sarebbe di conseguenza impossibile per una macchina replicare in maniera fedele la totalità dei comportamenti umani.

2.4 Senso comune e "Frame Problem"

Nella sezione precedente abbiamo indicato quali siano i problemi che sopraggiungono quando si cerca di codificare ed esplicitare in maniera formale il complesso della conoscenza umana al fine di renderlo utilizzabile nell'ambito dell'intelligenza artificiale e dei modelli computazionali della cognizione umana. In particolare, abbiamo indicato come alcuni tipi di conoscenza, ovvero quelli taciti, offrono delle notevoli sfide a coloro che lavorano in questi ambiti a causa del loro essere sostanzialmente non separabili dai contesti pratici e sociali in cui sorgono e si sviluppano. Proviamo ora a definire meglio dal punto di vista filosofico cosa si intende per questo complesso di conoscenze inespresse, e talvolta inconsce, e per quale motivo esse hanno dato origine ad uno dei grandi problemi dell'intelligenza artificiale, ovvero il cosiddetto "Frame Problem".

La dicotomia tra la conoscenza esplicita, formale, oggettiva e stabile che si fonda su conclusioni dedotte tramite metodi di ragionamento formali considerati come logicamente validi a partire da premesse verificabili, e la conoscenza "comune", posseduta dalla maggior parte degli uomini senza che alla sua base ci sia una particolare evidenza di tipo logico o scientifico, è presente nella storia della filosofia fin dai suoi primordi. Già in Parmenide, infatti, troviamo la distinzione tra la via della Verità stabile, oggettiva e incontrovertibile e quella invece dell'opinione comune, infondata e fallace.

In ambito filosofico, ma anche scientifico, il senso comune ha generalmente sempre avuto una connotazione tendenzialmente negativa o è stato comunque considerato come subordinato alla conoscenza fondata su criteri razionali e scientifici. Dare una definizione specifica di "senso comune" è un compito difficile data la vaghezza del suo concetto. Si tratta di uno di quei concetti del cui significato tutti possediamo una qualche conoscenza intuitiva, ma che fatichiamo a delineare precisamente. Si potrebbe definire come "un complesso di atteggiamenti

conoscitivi e di relativi contenuti che un certo gruppo sociale, o una determinata epoca storica, condividono in modo più o meno immediato e irriflesso, prescindendo da competenze "specialistiche" e che, pertanto, si tende a ritenere che costituisca una sorta di patrimonio conoscitivo comune a tutta la specie umana"[129].

È un insieme di credenze, conoscenze più o meno consce ed inespresse, che generalmente vengono accettate dalla maggior parte degli individui come ovvie, senza una vera e propria analisi critica circa la loro fondatezza. Queste convinzioni si caratterizzano per il fatto di "non "esibire" le credenziali della loro validità, in quanto sono spontanee e irriflesse e l'unico sostegno indiretto della loro validità sembra rappresentato dal fatto di essere larghissimamente "condivise""[130]. Questo è uno dei principali motivi per cui il senso comune viene spesso considerato negativamente in ambito filosofico e scientifico, anche se in realtà il suo non essere propriamente "giustificato" in molti casi deriva dal fatto di essere composto da conoscenze che "si impongono *spontaneamente*, ossia che sono *ovvie*"[131].

È indubbio che molte delle credenze facenti parte del senso comune, dal punto di vista logico non siano altro che fallacie prodotte da ragionamenti scorretti o false premesse. Al contempo, tuttavia, esso, proprio perché è così condiviso, immediato e "primitivo", risulta essere "il presupposto di ogni conoscenza "ulteriore", ivi comprese quelle che, in una certa misura, possono indurre a modificarne o correggerne alcuni contenuti"[132]. È chiaro, infatti, come la maggior parte delle cose, prima di essere oggetto di una qualche scienza o disciplina particolari, sia inizialmente considerata secondo le categorie del cosiddetto senso comune, e non già secondo quelle scientifiche o filosofiche. Per questo motivo "ogni discorso specializzato è concettualmente "successivo" al senso comune e determina in seno ad esso i propri *referenti*, anche se, svolgendosi, modifica più o meno profondamente quella realtà, quell'unità dell'esperienza, entro cui si svolge"[133].

Considerato in questo modo, il senso comune sembra acquisire una maggiore importanza e dignità ontologica rispetto alla considerazione negativa che generalmente gli è stata affidata da scienza e filosofia. Fossilizzandosi troppo su una descrizione del reale secondo parametri puramente razionali, formalmente validi e scientificamente provabili, si rischia di perdere una dimensione importante della stessa realtà che si sta cercando di descrivere. L'assolutizzazione

[129] Evandro Agazzi (a cura di), *Valore e Limiti del Senso Comune*, Milano, Franco Angeli, 2004, pag. 9
[130] *Ivi*, pag 13
[131] *Ibidem*
[132] *Ivi*, pag. 17
[133] Evando Agazzi, *Il Senso Comune e l'Unità dell'Esperienza*, in *Valore e Limiti del Senso Comune*, a cura di Evandro Agazzi, Milano, Franco Angeli, 2004, pag. 33

della scienza comporta inevitabilmente l'attribuzione ad essa di una "esclusiva per quanto concerne la conoscenza della verità"[134] e il "non riconoscere che la sua portata è solo *parziale* rispetto alla complessità di quel *mondo della vita* (…) che viene accolto nell'"immagine manifesta" del senso comune"[135].

Ovviamente, con ciò non si intende sminuire l'importanza e la validità epistemica della scienza, né tantomeno attribuire al senso comune un ruolo di maggior rilievo rispetto ad essa. Si tratta semplicemente di riconoscere che quella scientifica è "una visione *ridotta* della realtà e che, affinché essa non diventi una visione *riduttiva*, è necessario tenere aperta l'ottica sulla *pienezza* del reale"[136], la quale può essere mantenuta solamente a patto di non trascurare la componente del vissuto umano costituita dal senso comune.

Fatta questa breve precisazione circa il significato della nozione di "senso comune", vediamo ora come esso può essere considerato nell'ambito dell'intelligenza artificiale e che genere di problematiche esso produce.

Se consideriamo il campo dell'intelligenza artificiale come il campo in cui si cerca sia di costruire delle macchine che si comportino e svolgano dei compiti in maniera simile o uguale (e possibilmente in modo più preciso, rapido ed efficiente) agli esseri umani, è inevitabile che prima o poi ci si debba confrontare anche con quella parte della conoscenza umana che abbiamo definito come "senso comune".

Se nel fare ciò si adotta un approccio che nella sezione precedente abbiamo definito come "oggettivistico", dovrebbe essere possibile esplicare e tradurre in linguaggio formale, almeno in linea di principio, anche il senso comune. Tuttavia, non appena proviamo a razionalizzarlo scopriamo quanto in realtà esso sia lontano dalla logica e quanti paradossi sorgano da una sua razionalizzazione: "quando si cerca di spiegare il senso comune come insiemi di credenze o come struttura complessa di insiemi di credenze parzialmente sovrapposti, ogni piccola frazione di essa -come si manifesta, ad esempio, in un ragionamento discorsivo oppure attraverso un determinato comportamento- appare in stretta connessione con una rete inestricabile di conoscenza o pseudo-conoscenza implicita"[137].

In generale, possiamo affermare che quando ci troviamo a dover rappresentare in modo esplicito i contenuti del senso comune ci scontriamo inevitabilmente con il suo essere dipendente dai contesti nel quale esso è sorto. In altre parole, non è possibile produrre una

---

[134] *Ivi*, pag. 36
[135] *Ibidem*
[136] *Ivi*, pp. 36-37
[137] Luisa Montecucco, *Il Senso Comune come "Teoria" e come "Limite"*, In: Evandro Agazzi (a cura di), *Valore e Limiti del Senso Comune*, Franco Angeli, Milano, 2004, pag. 61

"teoria generale" del senso comune: il suo essere un prodotto di particolari contesti socio-culturali ne impedisce ogni tentativo di generalizzazione. Il senso comune, sebbene sia costituito da molte delle nostre intuizioni riguardanti il mondo e il nostro modo di vivere, è tuttavia "vacillante di fronte ai criteri di correttezza formale, di rigore metodologico, di confronto empirico che convalidano le teorie scientifiche"[138]. Per concludere possiamo affermare che il senso comune, a causa del suo essere così vago, indefinito, per certi versi infondato e contraddittorio non può essere oggetto di una teoria o di insiemi di teorie.

Ora, anche se il senso comune, come abbiamo appena affermato, sembra essere impossibile da formalizzare nei termini di una teoria o un insieme di teorie generali, può essere sicuramente esplicitato e definito almeno in parte. Qualunque sistema di intelligenza artificiale che sia progettato con lo scopo di svolgere un qualche compito in maniera simile a quella di un essere umano o di agire nel mondo, ha, infatti, necessariamente bisogno di una certa base interpretativa attraverso la quale categorizzare e analizzare le informazioni che riceve dall'esterno. Un programma di intelligenza artificiale, solitamente, o raccoglie in maniera autonoma input provenienti dall'esterno e raccolti tramite sensori, oppure da insiemi di dati forniti dai programmatori stessi in un formato già utilizzabile dalla macchina. Il problema della rappresentazione e codificazione della conoscenza nell'ambito dell'informatica è duplice: bisogna, infatti, stabilire sia il modello descrittivo sia il formato con cui codificare la rappresentazione del mondo all'interno della macchina.

Il formato non è altro che un insieme di proposizioni codificate in un linguaggio di programmazione basato su una logica di un certo ordine. I linguaggi di programmazione, linguaggi artificiali creati appositamente dai programmatori, sono innumerevoli e differiscono tra loro per la capacità di codificare in modo più o meno efficiente ed esaustivo le informazioni e le istruzioni. Alcuni di essi sono molto più adatti per realizzare certi scopi di altri, ma generalmente è possibile tradurre un programma in un qualsiasi linguaggio di programmazione.

A differenza dei linguaggi di programmazione, che semplificando un po' possiamo definire come interscambiabili, il modello interpretativo ricopre un ruolo fondamentale e costituisce le stesse condizioni di possibilità del corretto funzionamento del programma. Ad esempio, una macchina che abbia come fine quello di calcolare il moto di un corpo può essere dotata di un modello interpretativo che classifica gli input provenienti dall'esterno come particelle soggette a leggi fisiche. Tuttavia, è evidente che un tale modello non può essere

[138] *Ivi*, pag. 62

applicato anche ad una macchina che invece abbia come scopo quello di tradurre dei testi da un linguaggio naturale ad un altro.

Tornando all'argomento principale della nostra tesi, quando si tratta di programmare macchine che simulino le capacità cognitive o il comportamento umani, è evidente che ci si trovi davanti il problema della formalizzazione del senso comune. Abbiamo infatti indicato come molti dei modi in cui noi rappresentiamo il mondo e ci rapportiamo con esso abbiano alla base molte delle assunzioni implicite e inconsce che generalmente vengono indicate come "senso comune". Il pensiero, l'agire e il comportamento umani sono tutti profondamente influenzati da queste assunzioni:

> We can understand how the problem of commonsense understanding arises when we reflect that the computer comes into our world even more alien than a Martian. It does not have a body, needs, or emotions, and it is not formed by a shared language and other social practices. If the machine is to interact intelligently with us, it has to be endowed with and understanding of the human form of life.[139]

A questo proposito, prendiamo in considerazione un esempio proposto da Dennet[140] per mostrare meglio la quantità e la complessità delle assunzioni implicite che sono implicitamente sottointese anche ad azioni molto comuni che compiamo tutti i giorni. Supponiamo di dover prepararci un panino imburrato; il piano d'azione in questo caso sembra molto semplice: basta andare in cucina, prendere gli ingredienti dai luoghi in cui sappiamo che sono conservati e metterli insieme. Tuttavia, se analizziamo meglio la situazione ci accorgiamo che dietro questo procedimento all'apparenza molto banale, si nasconde un grandissimo e complesso insieme di conoscenze tacite di cui nemmeno ci rendiamo conto: "of course I couldn't do this without knowing a good deal – about bread, spreading mayonnaise, opening the fridge, the friction and inertia that will keep the turkey between the bread slices and the bread on the plate as I carry the plate over to the table"[141].

Facendo un'introspezione ancora più approfondita possiamo arrivare anche ad assunzioni implicite ancora più fondamentali di queste, ad esempio il fatto che due cose non possono essere contemporaneamente nello stesso posto, il fatto che se una cosa si trova in un posto non può

[139] Hubert L. Dreyfus, Stuart E. Dreyfus, *Mind over* Machine, New York, Free Press, 1986, pag. 79
[140] Daniel C. Dennet, *Cognitive Wheels: The Frame Problem of AI*, in *The Philosophy of Artificial* Intelligence, a cura di M. Boden, New York, Oxford University Press, 1990, pag. 153
[141] *Ivi*, pag. 152

essere contemporaneamente anche in un altro, che le situazioni cambiano in seguito alle nostre azioni.

Questi sono solo alcuni esempi delle innumerevoli implicazioni, assunzioni implicite, conoscenze inespresse che si trovano dietro le nostre azioni più semplici e comuni. Molte di esse sono state apprese tramite l'esperienza e la ripetizione delle stesse azioni, altre potrebbero addirittura essere innate, in ogni caso è evidente che "no agent that did not *in some ways* have the benefit of the information could perform such a simple task"[142]. Al fine di programmare una macchina che svolga questo tipo di azioni è di conseguenza necessario stabilire quali siano le conoscenze implicite ed esplicite che colui che agisce deve possedere al fine di portare a termine il compito.

Ora, il problema risiede nel fatto che il computer di per sé non è altro che una *tabula rasa*, tutte le informazioni che possiede gli devono essere fornite in formato da lui utilizzabile dai programmatori, o devono essere apprese in qualche modo: "the tasks set by AI start at zero: the computer to be programmed to simulate the agent (or the brain of the robot, if we are actually going to operate in the real, non-simulated world), initially knows nothing at all "about the world""[143]. Di conseguenza, o tutto questo bagaglio di informazioni deve essere installato all'interno del computer da un programmatore umano o deve essere appreso da esso tramite l'esperienza (ovviamente in questo secondo caso bisogna comunque fornire al computer dei metodi di apprendimento, cosa che verrà trattata nel prossimo capitolo).

Nel caso si decida di procedere con il primo metodo, ci si trova davanti sostanzialmente a tre problemi: il primo, che abbiamo già ampiamente illustrato, è quello della possibilità stessa di codificazione e sistematizzazione del senso comune, il secondo è costituito dal come installare effettivamente tutte queste informazioni in un computer e il terzo dal come poi utilizzare questa immensa mole di conoscenze al fine di risolvere problemi e creare schemi di azione.

Supponiamo ora che sia in linea di principio possibile codificare in un qualche modo tutte le conoscenze esplicite e implicite che un essere umano possiede e che gli permettono di agire e comprendere il mondo. In che formato possiamo inserirle in un computer? Si potrebbe pensare di creare una sorta di lista di proposizioni che contengano ogni singola conoscenza e regola che possediamo, come in una sorta di sterminata enciclopedia: "ideally an entire encyclopedia would somehow be stored in computer-accessible form, not as text but as a collection of

[142] *Ivi*, pag. 153
[143] *Ibidem*

thousands of structured, multiply indexed units"[144]. Tuttavia, questo richiederebbe che fosse possibile esprimere la conoscenza in un formato totalmente oggettivo ed esplicito, cosa che abbiamo più volte indicato essere impossibile. Anche la comprensione di un semplice articolo di una ipotetica enciclopedia di questo tipo, infatti, richiederebbe "a large body of common-sense knowledge not yet shared by computer software"[145].

Infine, un elenco enciclopedico di questo genere avrebbe ovviamente dimensioni difficilmente immaginabili e comporterebbe di conseguenza un nuovo problema: trovare in un tempo ragionevole le informazioni e le conoscenze pertinenti al compito che la macchina deve svolgere: "The demand for an efficient system of information storage (…) it is a time limitation, for stored information that is not reliably accessible for use in the short real-time spans typically available to agents in the world is of no use at all"[146]. Alcune delle caratteristiche fondamentali di un comportamento intelligente sono, infatti, la velocità e la prontezza con le quali affrontiamo le situazioni nel mondo reale: "a creature that can solve any problem given enough time -say a million years- is not in fact intelligent at all[147]". Ora, è vero che il progresso tecnologico ci ha fornito di computer sempre più veloci ed efficienti nel processare le informazioni a loro disposizione, ma un ipotetico insieme esplicito di tutta la conoscenza umana sarebbe comunque troppo grande per essere utilizzabile efficacemente anche dai più moderni computers, e probabilmente anche da quelli che verranno progettati e costruiti nel prossimo futuro.

Un secondo modo di approcciare il problema potrebbe essere quello di limitarsi ad indicare una sorta di relativamente contenuto e limitato insieme di "assiomi" dai quali poi dedurre a seconda della situazione conoscenze e schemi di azione. Anche questo secondo metodo sembra essere troppo utopistico sia dal punto di vista più strettamente pratico della sua realizzazione effettiva, sia dal punto di vista più teoretico e filosofico. Basti semplicemente pensare a quanto spesso fatti esperienziali imprevisti ci costringono a riconsiderare e modificare le nostre conoscenze pregresse e i nostri modelli interpretativi.

Sia che si decida di adottare un approccio di tipo esplicito ed enciclopedico, sia uno di tipo deduttivo, ci si trova davanti al problema della rilevanza, ovvero dello stabilire quali delle conoscenze e regole a disposizione di un sistema di intelligenza artificiale siano rilevanti in un determinato contesto. È un problema duplice, in quanto si tratta di dover decidere non solo quali informazioni siano pertinenti al contesto e al compito che deve svolgere, ma anche quali siano

[144] Hubert L. Dreyfus, Stuart E. Dreyfus, *Mind over* Machine, New York, Free Press, 1986, pag. 79
[145] *Ibidem*
[146] Daniel C. Dennet, *Cognitive Wheels: The Frame Problem of AI*, in *The Philosophy of Artificial* Intelligence, a cura di M. Boden, New York, Oxford University Press, 1990, pag. 155
[147] *Ibidem*

ininfluenti e di conseguenza ignorabili. Se gli esseri umani decidono per lo più automaticamente e inconsciamente cosa sia rilevante e appropriato in un particolare contesto, le macchine computazionali devono essere in grado di fare altrettanto però sotto forma di regole: "determination of relevance will have to be based on further facts and rules, but the question will again arise as to which fact and rules are relevant for making each particular determination[148]. In altre parole, "the sort of rules human beings are able to ariculate always contain *ceteris paribus* conditions, that is, the rules are applicable "everything else being equal""[149].

Si tratta di isolare, di volta in volta, e a seconda della particolare situazione contingente tutte le regole e le conoscenze pertinenti, e considerare come ininfluenti e secondarie tutte quelle che invece non hanno direttamente a che fare con essa. Tuttavia, essendo questo un tipo di ragionamento necessariamente dipendente dal contesto particolare della sua applicazione, non è possibile generalizzarlo e formalizzarlo in un modo che possa essere indistintamente applicato da una macchina in qualunque situazione.

Molto raramente, infatti, si possono applicare in modo esatto e preciso conoscenze e regole in un determinato contesto. Più spesso sono necessari un qualche adattamento, modifica e integrazione delle stesse per meglio adattarle ad esso. Inoltre, volendo esplicare in modo formale i procedimenti che utilizziamo inconsciamente per stabilire la rilevanza di determinate informazioni in determinati ambiti, ci potremmo trovare in una sorta di regresso all'infinito in cui di volta in volta dovremmo giustificare le nostre scelte di rilevanza con altri criteri di rilevanza: "to explain our actions and our rules, we must eventually fall back on our everyday practises and eventually say "this is what one does" or "that's what it is to be a human being"[150].

Il fatto è che noi esseri umani in molte occasioni non ci rendiamo nemmeno conto di quali meccanismi utilizziamo per attingere dal nostro bagaglio di conoscenze pregresse quando ci troviamo di fronte ad una certa situazione. Certamente, soprattutto quando dobbiamo svolgere compiti al di fuori dei nostri ambiti abituali, utilizziamo degli schemi di pianificazione aperti all'introspezione. Ad esempio, ci immagiamo svolgere particolari azioni in una determinata situazione e cerchiamo di prevederne gli esiti prima di agire. Tuttavia, "what happens backstage, as it were, to permit this "seeing" (…) is utterly inaccessible to instrospection"[151]. Questa nostra incapacità di introspezione nei confronti dei processi che

[148] Hubert L. Dreyfus, Stuart E. Dreyfus, *Mind over* Machine, Free Press, New York, 1986, pag. 80
[149] *Ibidem*
[150] *Ivi*, pag. 81
[151] *Ibidem*

impieghiamo nell'utilizzo delle conoscenze a nostra disposizione, è ancora più evidente nel caso di azioni consuete e di routine nelle quali non è nemmeno presente questa dimensione di pianificazione cosciente.

Poiché non abbiamo una vera e propria comprensione del nostro modo di ragionamento e pianificazione non riusciamo nemmeno a formalizzarlo in modo tale da farlo replicare ad una macchina: "Do we have any model for how such unconscious information-appreciation might be accomplished? The only model we have *so far is conscious*, deliberate information-appreciation. Perhaps, AI suggests, that is good model. If it isn't, we are all utterly in the dark for the time being"[152].

Tutte queste problematicità riguardanti l'implementazione del senso comune nei sistemi di intelligenza artificiale vengono generalmente indicate con il nome di "frame problem". Il cosiddetto *frame problem* venne descritto per la prima volta da John McCarthy e Patrick J. Hayes nel loro articolo del 1969 intitolato *Some Philosophical Problems from the Standpoint of Artficial Intelligence*[153]. In origine, il termine fu usato per indicare un particolare problema di rappresentazione che sorge quando si cerca di descrivere dei fatti riguardanti il mondo utilizzando la logica del primo ordine. In particolare, si tratta del fatto che specificare che alcune condizioni sono cambiate in seguito ad una determinata azione non implica che tutte le altre, anche se non direttamente coinvolte, siano rimaste immutate. In altre parole, per ogni azione è necessario specificare tramite assiomi che tutto ciò che non è direttamente condizionato da una qualche azione rimane immutato e nello stato che precedeva quell'azione. In seguito, il termine "frame problem" ha assunto un significato più generale fino ad indicare l'insieme dei problemi che sorgono dalla rappresentazione, in formato logico o di linguaggio di programmazione, di contesti mutevoli e in continua trasformazione. Se lo specifico problema di carattere puramente logico e matematico, per il quale il termine è stato coniato, è stato negli anni risolto in vari modi, la questione filosofica di fondo che gli sta dietro è rimasta: "the attempt to capture human, temporal, situated, continously changing know-how in a computer as a static, de-situated, discrete, knowing-that has become known as the frame problem"[154].

Il frame problem, sia nella sua versione "originale" sia nel suo significato più ampio e generale, è da sempre uno dei maggiori ostacoli per quella che possiamo chiamare Intelligenza

[152] Daniel C. Dennet, *Cognitive Wheels: The Frame Problem of AI*, in *The Philosophy of Artificial* Intelligence, a cura di M. Boden, New York, Oxford University Press, 1990, pag. 156

[153] John McCarthy, Patrick J. Hayes, *Some Philosophical Problems from the Standpoint of Artificial Intelligence*, in *Machine Intelligence 4*, a cura di B. Meltzer, D. Michie, Edinburgh, Edinburgh University Press, 1969, pp. 463-502

[154] Hubert L. Dreyfus, Stuart E. Dreyfus, *Mind over* Machine, New York, Free Press, 1986, pag. 82

Artificiale convenzionale. Con “Intelligenza Artificiale convenzionale” indichiamo tutti quei programmi di IA nei quali tutte le informazioni e le istruzioni devono essere esplicitamente scritte e codificate in un qualche linguaggio di programmazione da un programmatore umano.

Riassumendo quello che abbiamo già esposto in questa sezione, un approccio di questo tipo alla creazione di sistemi di intelligenza artificiale si scontra necessariamente con una serie di problematicità sia puramente tecniche e tecnologiche, che più profonde, fondamentali e filosofiche.

La possibilità di esprimere esplicitamente tutta la conoscenza umana; quella di tradurre in termini di logica formale i processi cognitivi, anche inconsci, che sono alla base del nostro comportamento e del nostro modo di rapportarci e comprendere il mondo e infine quella di poter raffigurare in un insieme finito di proposizioni la mutevolezza e la vaghezza di un mondo in continuo divenire, sono tutte incognite che sembrano limitare le potenzialità dell’intelligenza artificiale.

Facendo un parallelo con quanto espresso nel primo capitolo, in merito alla presunta capacità degli esseri umani di essere in grado di andare oltre l’incompletezza di un qualsiasi sistema formale, possiamo affermare che anche in questo caso, a distinguerci dalle macchine computazionali, sia la nostra capacità di “completare” e “andare oltre” una qualsiasi formalizzazione statica del mondo in cui viviamo. Ogni presunto sistema di regole che normi quello che generalmente intendiamo come “mondo” deve, infatti, essere inevitabilmente incompleto. Non è possibile fornire una descrizione esaustiva del mondo che tenga in conto ogni possibile evenienza o situazione contingente. In molti casi due regole riferite ad un medesimo ambito o ambiente possono entrare in conflitto tra loro, e la decisione in merito a quale delle due applicare, non viene presa dal singolo essere umano, che si trova in quella situazione, necessariamente tramite un ragionamento di tipo logico-matematico esplicito, ma anche sulla base delle sue esperienze, del suo vissuto, del senso comune e di conoscenze tacite. Queste capacità umane di applicare flessibilmente delle regole, di adattarsi a situazioni nuove ed impreviste, di modificare i propri schemi comportamentali e i propri insiemi di credenze, di crearne di nuovi a seconda della situazione, sono forse i principali ostacoli con cui coloro che si occupano di intelligenza artificiale si sono trovati, e si trovano ancora, a dover fare i conti.

Quel superamento dell’incompletezza in ambito logico-matematico che Lucas aveva indicato come tratto distintivo tra le potenzialità di una mente umana e quelle di una macchina computazionale, può essere, più in generale, e forse più appropriatamente, individuato anche in tutti quegli ambiti che hanno a che fare con il “mondo della vita” umano. Non si sta affermando qui l’impossibilità assoluta e categorica che un giorno possa esistere un’intelligenza artificiale

che pensi e agisca indistintamente da un essere umano. Esistono già, infatti, intelligenze artificiali che in ambiti specifici eseguono dei compiti anche più velocemente ed efficacemente degli esseri umani. Di conseguenza non può essere escluso a priori che l'avanzamento tecnologico possa un giorno permetterci di programmare intelligenze artificiali che siano in grado di pensare, agire e rapportarsi al mondo nella sua totalità nello stesso modo in cui lo facciamo noi umani. Tuttavia, allo stato attuale delle cose, un'intelligenza artificiale generale di questo tipo sembra ancora molto lontana dal poter essere realizzata. La ragione di ciò la possiamo trovare anche, e soprattutto, nella difficoltà di replicare in modo formale ed esplicito tutta quella dimensione implicita, tacita e più prettamente "umana" della nostra vita.

CAPITOLO III - APPRENDIMENTO MECCANICO E RETI NEURALI

Dopo aver trattato della comparazione tra intelligenza umana e artificiale in termini logico-formali nel primo capitolo e da un punto di vista epistemologico nel secondo, in questo terzo e ultimo capitolo ci concentreremo invece proprio sui sistemi di intelligenza artificiale veri e propri e sulle più recenti innovazioni in questo campo.

Ci concentreremo in particolare sulle cosiddette "reti neurali", ovvero dei sistemi di intelligenza artificiale basati su particolari algoritmi che consentono processi di apprendimento molto più efficaci e avanzati rispetto ai programmi classici. I programmi di intelligenza artificiale sviluppati attraverso metodi di apprendimento basati sulle reti neurali vengono sempre più utilizzati in molti e disparati campi, e stanno rivoluzionando sempre più settori non solamente di ambito tecnologico. Vengono usati, ad esempio, nell'analisi di dati, nel settore della guida autonoma, nel riconoscimento e nella trascrizione del parlato e recentemente anche nella generazione di testi, immagini e video. In aggiunta a queste applicazioni, algoritmi di intelligenza artificiale addestrati tramite reti neurali sono presenti in molte delle applicazioni e social media che usiamo tutti i giorni, anche se non ce ne rendiamo conto. Sono responsabili, ad esempio, della raccolta di informazioni riguardanti i gusti e delle preferenze di ciascun utente. In base a questi dati vengono poi scelti e proposti i contenuti, ma anche le inserzioni pubblicitarie, che meglio si adattano a ciascuno. Per questo motivo, data la grandissima e sempre maggiore rilevanza e ruolo dei social media nella nostra società, essi possono influenzare e plasmare i gusti, i costumi e la moda in modo molto marcato. Tuttavia, come vedremo, il funzionamento di questi algoritmi non è del tutto chiaro nemmeno per coloro che li hanno progettati e programmati. Di conseguenza possiamo forse affermare che per certi versi, queste tecnologie stanno diventando in un certo senso "indipendenti" rispetto a noi, e potrebbero modificare in modi imprevisti e imprevedibili il nostro stesso modo di vivere. Diventa di conseguenza essenziale, anche a livello più teoretico e "filosofico", cercare di capirli sempre meglio.

Usando espressioni già impiegate nel secondo capitolo possiamo affermare che, poiché queste intelligenze artificiali stanno sempre di più entrando a far parte della nostra *forma di vita*, e addirittura contribuendo a modificarla, è quanto meno auspicabile che da parte nostra ci sia non solo una maggiore comprensione del loro funzionamento, ma anche una maggiore attenzione nel loro impiego al fine di evitare effetti imprevisti e indesiderati.

Ora, il dibattito in merito alle questioni etiche connesse all'utilizzo delle intelligenze artificiali è un argomento vastissimo e che esula dal tema principale della nostra tesi. Ci limiteremo, di conseguenza, a provare a interpretare secondo i paradigmi espressi nei precedenti capitoli queste nuove tecnologie e a cercare di capire se e in che modo questi nuovi e più moderni approcci nel campo dell'intelligenza artificiale possono superare alcune di quelle problematiche e ostacoli che abbiamo espresso in precedenza.

3.1 Macchine che apprendono

Una delle principali e più importanti abilità della mente umana è sicuramente quella dell'apprendimento. Con apprendimento non si intende la semplice acquisizione e memorizzazione di nuove nozioni e informazioni, ma anche la capacità di imparare nuovi modelli di interpretazione e di modificare quelli già esistenti.

Fin dalle prime ricerche nel campo dell'intelligenza artificiale, gli studiosi si sono chiesti se fosse possibile progettare un computer che non si limitasse semplicemente ad elaborare le informazioni che gli vengono fornite secondo le regole logiche contenute nella programmazione che gli viene imposta, ma se fosse possibile realizzare anche una macchina che sia in grado di modificare la propria programmazione imparando dalla propria esperienza e migliorando il modo in cui esegue i compiti che le vengono assegnati.

Uno dei problemi più grandi che si riscontrano nel campo della programmazione di sistemi di intelligenza artificiale deriva dalla complessità del loro programma, ovvero dell'insieme di istruzioni che il programmatore deve fornire loro. I moderni computer digitali sono infatti costruiti sul modello dell'architettura ideata dal logico e matematico John von Neumann attorno alla metà del ventesimo secolo. Questa architettura prevede fondamentalmente una unità centrale di calcolo (CPU) che svolge le computazioni, una memoria che contiene sia il programma che controlla la CPU sia le informazioni che esso ha a disposizione e su cui opera, sistemi di input e output che consentono l'interazione dell'operatore con il computer e infine il sistema di collegamento di queste diverse unità.

Ora, in un sistema di questo tipo tutte le operazioni che la CPU svolge sono necessariamente dettate dal programma, il quale consiste in un serie di istruzioni codificate e contenute nella memoria del computer e che vengono scritte dal programmatore[155]. Per questo

[155] Nei moderni computer digitali il processore riceve istruzioni ed esegue operazioni nel cosiddetto "linguaggio macchina", ovvero un linguaggio basato su un alfabeto binario che comprende due simboli: lo 0 e l'1. Questo linguaggio è tuttavia di difficile comprensione per l'essere umano, per questo motivo, per scrivere i programmi, vengono utilizzati dei "linguaggi di programmazione", ovvero dei linguaggi più vicini sia alla logica umana che

motivo ogni singola operazione che il processore svolge deve essere indicata, e quindi scritta, da un essere umano. Di conseguenza, lo sviluppo di programmi di intelligenza artificiale che riescano a svolgere dei compiti molto complessi, come ad esempio la conversazione con un essere umano, è necessariamente un processo estremamente dispendioso sia in termini strettamente pratici, in quanto sono costituiti da un numero enorme di linee di codice, sia da un punto di vista teorico, in quanto richiederebbero la traduzione in termini di istruzioni logiche eseguibili da un computer di elementi molto astratti e difficilmente quantificabili in termini logici. Basti pensare alla già precedentemente indicata difficoltà insita nel tradurre in un insieme di istruzioni logiche delle azioni fisiche relativamente semplici, come ad esempio camminare, per rendersi conto del compito monumentale in cui consisterebbe fare la medesima cosa con alcune delle funzioni più elevate della nostra mente, tra le quali appunto la comprensione del linguaggio.

Per ovviare a queste difficoltà già Alan Turing, nel suo fondamentale articolo *Computing machinery and intelligence*, propose un modello teorico di una "learning machine", ovvero di una macchina che, al pari di un bambino, sia dotata di una versione basilare e semplificata delle caratteristiche che le si vuole far simulare, e che riesca, tramite l'esperienza e l'apprendimento, ad affinarle fino a farle giungere ad un livello paragonabile a quelle di un essere umano adulto: "Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain"[156]. Turing divise il compito di progettare una macchina di questo tipo in due parti: da un lato la programmazione vera e propria della "child machine" e dall'altra il suo processo educativo.

Per quanto riguarda la parte della programmazione vera e propria, propose un approccio di tipo evoluzionistico, ovvero partendo da un programma iniziale molto basilare vi si introducono via via delle modificazioni, e giudicandone i risultati, si valuta quali altre modifiche apportare o se scartare completamente il programma, applicando una sorta di selezione naturale. Per quanto invece concerne il processo di apprendimento, Turing propose un modello educativo basato su un sistema di ricompense e punizioni:

ai linguaggi naturali. Le istruzioni descritte in linguaggio di programmazione vengono poi tradotte in linguaggio macchina eseguibile dal processore dal computer stesso.

[156] Alan Turing, *Computing Machinery and Intelligence*, in *The Philosophy of Artificial Intelligence*, a cura di M. Boden, New York, Oxford University Press, 1990, pp. 62

> We normally associate punishments and rewards with the teaching process. Some simple child machines can be constructed or programmed on this sort of principle. The machine has to be so constructed that events which shortly preceded the occurrence of a punishment signal are unlikely to be repeated, whereas a reward signal increased the probability of repetition of the events which lead up to it. These definitions do not presuppose any feelings on the part of the machine.[157]

Non si tratta dunque di dotare una macchina di "sentimenti", Turing non ha in questo senso la pretesa di realizzare una macchina cosciente che sappia provare emozioni in modo simile a quello di un essere umano, si tratta semplicemente di dare una valutazione quantitativa al risultato di una determinata computazione e, sulla base di essa far decidere alla macchina se e in che modo modificare parte del suo programma e dei suoi parametri per cercare di ottenere un risultato, e di conseguenza anche una "ricompensa", quantitativamente migliore. Tale modifica del programma non avviene ovviamente sul piano strutturale e fondamentale dello stesso, "the rules which get changed in the learning process are of a rather less pretentious kind, claiming only an ephemeral validity"[158]. In questo senso la macchina non cade nel paradosso derivante dal fatto di cambiare le proprie stesse regole di funzionamento, il nucleo fondamentale del suo programma rimane infatti invariato: "How can the rules of operation of the machine change? They should describe completely how the machine will react whatever its history might be, whatever changes it might undergo. The rules are thus quite time-invariant"[159].

Alcune di queste idee proposte da Turing si sono rivelate seminali e fondamentali per le più recenti applicazioni di "machine learning" di cui ora discuteremo.

3.2 Reti Neurali

Vediamo ora più in dettaglio in che cosa consistono e come funzionano le già citate reti neurali che negli ultimi anni hanno rivoluzionato l'intero campo dell'intelligenza artificiale.

Innanzitutto, ricordiamo che in matematica e informatica per "funzione" intendiamo, informalmente, un procedimento che permette di assegnare ad un determinato valore x un altro valore y. Data una certa funzione possiamo sempre stabilire per qualsiasi input un determinato output. Supponiamo ora di non conoscere una certa funzione f ma di possedere solamente alcuni

[157] *Ivi*, pp. 62-63
[158] *Ibidem*
[159] *Ibidem*

dei suoi risultati *y* a partire da certi input *x*. Possiamo noi, a partire da questi dati, risalire alla funzione *f* che li ha generati? Questo procedimento, che in gergo tecnico si chiama "ingegneria inversa", è esattamente quello che fa una rete neurale. A partire da un insieme di dati conosciuti, costituito da input e corrispettivi output, una rete neurale cerca di ricavare la funzione responsabile della generazione di quell'insieme. Non è necessario, tuttavia, che la funzione ricavata dalla rete neurale sia matematicamente quella esatta, nella maggior parte dei casi pratici è sufficiente trovare una funzione che approssimi quanto meglio possibile quella che stiamo cercando. In altre parole, il compito di una rete neurale è di trovare, a partire da un insieme di dati, una funzione approssimata che li generi. Tramite un procedimento che viene definito come "apprendimento" o "addestramento", la rete neurale genera via via funzioni sempre più precise e simili a quella ideale che si sta cercando.

Il termine "rete neurale" deriva dal fatto che la loro struttura è ispirata e ricorda quella del cervello umano. Infatti, come quest'ultimo, sono costituite da dei "blocchi", chiamati anche "neuroni", interconnessi tra loro fino a formare una vera e propria rete. Inoltre, proprio come nelle sinapsi del nostro cervello, ogni blocco riceve e a sua volta invia delle informazioni a tutti gli altri neuroni a cui è collegato. Precisiamo che si tratta solamente di analogie, le reti neurali non sono sistemi che si propongono di imitare a livello strutturale il funzionamento di un cervello. Sia i loro neuroni che le connessioni sono, infatti, simulate da un normale computer; non hanno bisogno a livello di hardware di una macchina computazionale speciale o che sia in qualche modo fisicamente simile a un cervello.

Dopo aver descritto quale sia lo scopo di una rete neurale, vediamo ora meglio come esse funzionano. Una rete neurale, come abbiamo appena indicato, è formata da strati progressivi di blocchi o neuroni che ricevono un'informazione in entrata e a partire da questa ne generano una in uscita. Ogni neurone riceve tutti gli input dei neuroni dello strato precedente e a partire da questi genera un output che passa a quello successivo. In altre parole, ogni neurone è una funzione matematica.

Ciascun neurone moltiplica tutti gli input ricevuti per un certo numero chiamato "weight" (peso) e li somma tra loro aggiungendo un'altra cifra denominata "bias". Pesi e bias sono dei parametri specifici di ciascun neurone e sono le variabili che vengono modificate nel processo di apprendimento della rete neurale. Possiamo pensare ai neuroni come a dei mattoni da costruzione che possono essere modificati e riassemblati al fine di realizzare lo scopo che vogliamo dare alla rete neurale. Se consideriamo ciascun neurone come una funzione matematica semplice, il "weight", moltiplicando la x, ha il compito di "ruotare" il grafico della funzione, mentre il "bias" che si somma ad essa lo "muove" su e giù e a destra e sinistra.

Se combiniamo tutte le innumerevoli funzioni lineari rappresentate da ognuno dei neuroni che costituiscono la rete neurale, otteniamo una funzione molto complessa che altro non è che la somma di esse. Questa funzione risultante è quella che idealmente rappresenta l'insieme di dati di partenza e che stiamo cercando di approssimare tramite la rete neurale. Ora, matematicamente parlando, una qualsiasi somma di funzioni lineari è a sua volta una funzione lineare. Tuttavia, raramente l'insieme di dati di partenza è rappresentabile tramite una semplice funzione lineare, nella maggior parte dei casi è il risultato di una funzione molto complessa. Di conseguenza, a ciascun neurone viene applicata una cosiddetta "activating function", la quale, in parole semplici, "trasforma" la funzione lineare di partenza in una non lineare, e in questo modo anche la funzione complessiva risultante dalla somma di tutte le singole funzioni semplici potrà essere non lineare.

Abbiamo detto che pesi e bias sono le variabili su cui possiamo agire per modificare la funzione risultante della rete neurale. In altre parole, un cambiamento in uno qualsiasi di questi valori in uno degli innumerevoli neuroni che costituiscono la rete modificherà (minimamente) anche la funzione finale. Il progressivo cambiamento di questi parametri al fine di modellare e approssimare la funzione risultante a quella che idealmente rappresenta l'insieme di dati di partenza, costituisce l'"apprendimento" della rete neurale. Ora, modificare "manualmente" pesi e bias in ciascuno delle migliaia di neuroni della rete è un compito troppo dispendioso, di conseguenza vengono impiegati degli algoritmi che lo fanno automaticamente. Questi algoritmi, semplificando, calcolano un valore che rappresenta l'errore tra la funzione generata e quella ipotetica che si sta cercando e modificano pesi e bias in modo da ridurre quanto più possibile questo valore. Successive reiterazioni di questo procedimento approssimano sempre di più la funzione a quella che si sta cercando.

È inoltre possibile provare rigorosamente che le reti neurali altro non sono che delle cosiddette "Universal Function Approximators"[160], ovvero delle funzioni che possono generare o approssimare a qualunque grado di precisione qualsiasi altra funzione computabile. Da ciò possiamo dedurre che, almeno a livello teorico, una rete neurale può approssimare e imitare qualunque cosa sia esprimibile nei termini di una funzione computabile. Si può anche dimostrare matematicamente che le reti neurali sono "Turing Complete", ovvero possono svolgere qualsiasi problema che un computer tradizionale o una qualsiasi macchina computazionale può svolgere. Questo comporta che, sempre in linea teorica, è possibile

[160] Si definisce *universal function* una funzione computabile che è capace di calcolare qualsiasi altra funzione computabile. In altre parole, è il corrispettivo astratto della macchina di Turing universale. L'esistenza di questa funzione è una conseguenza diretta della numerazione godeliana che abbiamo citato nel primo capitolo.

generare tramite una rete neurale un qualsiasi programma informatico “tradizionale”. In altre parole, sarebbe possibile generare tramite le reti neurali tutti i programmi che attualmente sono scritti da programmatori umani e che sono composti da lunghissime catene di istruzioni e informazioni. Questi programmi “tradizionali” sono gli stessi di cui abbiamo discusso nel corso del secondo capitolo.

3.2.1 Reti neurali: vantaggi e limitazioni

Dopo aver brevemente e informalmente descritto cosa sono e come funzionano le reti neurali, discuteremo ora dei principali vantaggi che esse offrono rispetto ai programmi di intelligenza artificiale tradizionali e delle loro inevitabili limitazioni e problematicità derivanti dal loro utilizzo.

Iniziando dai vantaggi, possiamo affermare che indubbiamente le reti neurali risolvono molti dei problemi derivanti dalla codifica e trascrizione in linguaggio di programmazione di molte delle nostre conoscenze tacite, che nel precedente capitolo abbiamo indicato essere storicamente uno dei principali problemi nel campo dell’intelligenza artificiale. In particolare, esse sono superiori ai normali programmi soprattutto in quei campi che richiedono una certa dose di “intuizione” o di logica “fuzzy”, problemi tipicamente e storicamente difficili da risolvere per i computers tradizionali sia a causa della loro stessa architettura che della nostra difficoltà di formalizzare tali capacità.

Facciamo ora un esempio per esplicare meglio questo punto. Supponiamo di voler programmare un computer affinché riconosca una determinata cosa, ad esempio un cane, in delle immagini o fotografie. In primo luogo, è opportuno Il settore del riconoscimento visivo di oggetti (denominato anche “computer vision”), uno dei campi in cui maggiormente vengono impiegati sistemi di intelligenza artificiale, è relativamente esente da molte delle problematiche di carattere epistemologico che abbiamo indicato nel capitolo precedente. Questo perché la traduzione di immagini in un formato che può essere utilizzato da un computer non comporta particolari complicazioni, di conseguenza supponiamo che il computer possa “vedere” e analizzare informazioni visive senza problemi. Fatta questa breve precisazione, torniamo al nostro esempio.

Adottando un approccio di programmazione tradizionale, la prima cosa che dobbiamo fare è trovare ed esprimere formalmente tutte le caratteristiche estetiche che ci fanno riconoscere una certa cosa come un “cane”. Si tratta di un compito sicuramente arduo, sebbene non in linea di principio impossibile. Noi infatti, generalmente, quando riconosciamo un

determinato oggetto, animale o persona, non pensiamo consciamente ed esplicitamente a tutte le caratteristiche che esso deve possedere per essere tale, ma lo facciamo istintivamente e immediatamente sulla base delle nostre esperienze pregresse. Un computer, tuttavia, come abbiamo indicato nel capitolo precedente, è essenzialmente una *tabula rasa*, di conseguenza dobbiamo fornirgli esplicitamente tutte le proprietà che ci permettono di riconoscere un cane in un'immagine visiva.

Tuttavia, non appena proviamo a definire quelle che potrebbero essere le caratteristiche distintive che fanno di un certo animale un "cane", ci rendiamo subito conto della complessità di questo compito. Forma, colore e dimensioni, ad esempio, sono estremamente variabili a seconda della razza, di conseguenza non possono essere definite come tratti distintivi dell'"essere cane". Potremmo provare a indicare alcune caratteristiche comuni a tutte le razze canine, ad esempio l'avere quattro zampe o due orecchie. Tuttavia, sia tali proprietà sono comuni a moltissime altre specie di animali, sia si potrebbe dare il caso che nell'immagine considerata non si vedano tutte e quattro le zampe del cane. Sorge qui, inoltre, un altro e ben più fondamentale problema, ovvero quello del regresso (potenzialmente infinito) della definizione dei concetti. Per poter riconoscere che qualcosa possiede quattro zampe, infatti, è necessario sapere cosa sia una "zampa", il concetto della quale per essere definito avrebbe però bisogno di altre definizioni e così via.

Il punto è che il nostro concetto di "canità" (ma anche della maggior parte dei concetti e delle idee che possediamo), ovvero dell'insieme di tutte le qualità e caratteristiche che fanno di un certo animale un cane, non solo non è formalmente ed esplicitamente definito, ma anche qualora lo fosse, ci sarebbe comunque bisogno di una certa flessibilità e creatività nell'applicazione di tali proprietà al fine di riconoscere un cane in ogni occasione particolare. Noi, ad esempio, siamo in grado di identificare un cane anche osservandolo da dietro, quindi senza aver bisogno di vederne l'aspetto del muso, oppure sappiamo classificare come cane anche un esemplare di una razza che non abbiamo mai visto prima, e le cui qualità peculiari non facevano ancora parte del nostro concetto di "canità".

In altre parole, possiamo affermare che è molto difficile per noi esplicitare formalmente, e di conseguenza tradurre in un linguaggio di programmazione utilizzabile da un computer, tutti quei ragionamenti e concetti che nel capitolo precedente abbiamo indicato come facenti parte della nostra conoscenza implicita e del nostro "senso comune".

Tramite le reti neurali, tuttavia, possiamo in un certo senso ignorare tutte queste problematiche legate alla formalizzazione del nostro modo di pensare e ottenere dei risultati

che molto difficilmente sarebbero raggiungibili tramite programmi di intelligenza artificiale tradizionali. Vediamo come.

Nella sezione precedente abbiamo affermato che le reti neurali sono dei sistemi che non fanno altro che approssimare delle funzioni che idealmente sono alla base di un certo insieme di dati. Riprendendo in considerazione il nostro esempio del riconoscimento di un cane in un'immagine, possiamo immaginare che la nostra capacità di identificazione di esso sia rappresentabile o approssimabile da una funzione matematica. Noi, ovviamente, non siamo in grado di tradurre in termini di funzioni matematiche le nostre capacità cognitive, tuttavia possiamo servirci di una rete neurale per approssimare tale supposta funzione quanto più possibile.

Nello specifico viene raccolta una quantità molto elevata di immagini e per ognuna di esse il programmatore indica semplicemente se contiene o meno un cane. Possiamo considerare l'insieme delle immagini come l'ipotetica *x* di una funzione e il fatto che contenga o meno un cane come la *y*. La rete neurale, a partire da queste immagini, cerca di approssimare la funzione matematica che meglio rappresenta questo insieme di dati. Il processo di apprendimento, come abbiamo precedentemente indicato, si basa sulla modifica di pesi e bias in relazione all'errore che viene calcolato tra la funzione trovata dalla rete e quella idealmente perfetta che si sta cercando. Dopo un certo numero di diverse generazioni di funzioni si arriva ad una che approssima in modo soddisfacente quella voluta. A questo punto si può inserire come input della rete neurale un'immagine qualsiasi e la rete neurale, applicando la funzione che ha "appreso", riesce a riconoscere la presenza o meno di un cane in essa, senza che sia più necessaria alcuna specifica in tal senso da parte umana.

Come possiamo notare, utilizzando le reti neurali non è più necessaria alcuna formalizzazione delle caratteristiche fisiche che fanno di un certo animale un cane. Tutto quello che il programmatore deve fare è indicare, usando le sue implicite conoscenze e abilità di riconoscimento, se in ciascuna delle immagini tramite le quali viene addestrata la rete neurale sia presente o meno un cane. Le reti neurali, in altre parole, permettono di "saltare" interamente tutta la formalizzazione ed esplicitazione della componente tacita della conoscenza e delle abilità umane, riuscendo a produrre dei risultati impossibili, o comunque molto difficili da raggiungere per i programmi di intelligenza artificiale tradizionali.

Ovviamente, quello del riconoscimento degli oggetti nelle immagini è solamente uno, e tra i più semplici, dei compiti che una rete neurale può imparare a svolgere. Altre applicazioni possono essere il riconoscimento e la trascrizione del parlato, la traduzione di testi in altri linguaggi naturali, la guida autonoma di veicoli, l'analisi di immense moli di dati, ma anche

diagnosi mediche, predizioni economiche e un'infinità di altre. In breve, possiamo affermare che qualunque insieme di dati che può essere rappresentato come una funzione è potenzialmente apprendibile e approssimabile da una rete neurale.

Ma le reti neurali possono veramente apprendere potenzialmente qualunque cosa? In realtà esistono alcune limitazioni alle loro capacità e potenzialità. Innanzitutto, esse sono limitate dall'effettiva potenza di calcolo della macchina computazionale sulla quale sono simulate e addestrate. A livello teoretico, infatti, le funzioni ricercate dalla rete neurale possono essere approssimate alla perfezione solamente utilizzando un numero infinito di neuroni. Ovviamente, nessuna rete neurale esistente potrà mai avere un numero infinito di neuroni, di conseguenza non si potrà mai raggiungere un grado di approssimazione della funzione perfetto. Certamente, nella maggior parte delle applicazioni pratiche, non è mai richiesto un tale grado di approssimazione; tuttavia, si potrebbe dare il caso che alcune cose siano intrinsecamente troppo complesse per essere efficacemente simulate e apprese da una rete neurale. L'insieme di tutte le capacità cognitive che costituiscono la mente umana potrebbe essere, ad esempio, una di queste cose. Sebbene infatti, grazie alle reti neurali, un numero sempre maggiore di capacità umane (dalla generazione di testi o immagini alla composizione di musica) stia venendo simulato e riprodotto da intelligenze artificiali, un sistema di tipo "olistico" che le racchiuda tutte sembra essere al giorno d'oggi al di fuori della portata dell'attuale tecnologia.

Una seconda limitazione è legata all'insieme dei dati attraverso il quale la rete viene addestrata. Questi dati, infatti, devono essere quanto più numerosi e precisi possibile al fine di approssimare al meglio la funzione. Nell'esempio del riconoscimento visivo di oggetti che abbiamo proposto i dati di partenza, ovvero le immagini, non presentavano particolari problemi di questo tipo, in quanto come abbiamo affermato, non ci sono particolari problemi nella traduzione di immagini in un formato utilizzabile da un computer e nemmeno limitazioni nella quantità di esse che abbiamo a disposizione. Tuttavia, in molti casi i dati potrebbero essere errati, il loro formato essere corrotto o molto difficile da tradurre, oppure insufficienti.

Un'altra problematica legata ai dati è quello dei cosiddetti "bias" ovvero delle distorsioni sistematiche degli stessi causate da una parziale o condizionata selezione di essi. Gli insiemi di dati considerati, infatti, possono rispecchiare quelli che abbiamo definito come i giudizi personali e culturali di coloro che li selezionano. Ad esempio, sempre facendo riferimento al nostro esempio del cane, una persona europea potrebbe avere la tendenza a inserire nell'insieme di dati immagini contenti specie canine tipiche del continente in cui vive, tralasciando parzialmente o completamente specie autoctone di altre parti del mondo. In questo caso la rete

avrebbe difficoltà a riconoscere come cani esemplari di tali specie, poiché scarsamente presenti nelle immagini tramite le quali è stata addestrata.

Al di là di questo esempio piuttosto banale, quella dei *bias* dei dati è una problematica sistemica e molto profonda all'interno del campo delle reti neurali e che non genera solamente problemi di natura tecnica ma potenzialmente anche etica. Programmi di intelligenza artificiale sono, infatti, sempre più impiegati anche nella creazione e schedatura di profili di persone che poi possono venire utilizzati da grandi aziende per la selezione del personale oppure da istituti di credito per stabilire il grado di rischio connesso alla concessione di un mutuo. Ora, prescindendo da un discorso più generale a riguardo della legittimità di queste pratiche, è evidente quanto alto sia il rischi che dei pregiudizi, ad esempio di natura etnica o sociale, anche involontari e inconsci, insiti nei dati con i quali la rete neurale viene addestrata, possano poi avere delle serie ripercussioni anche sulla vita delle persone. Questo e molti altri temi sono al giorno d'oggi al centro del dibattito sull'etica dell'intelligenza artificiale, ma sono argomenti che tuttavia esulano dal focus di questa tesi.

Altre limitazioni di carattere tecnico sono legate ad eventuali errori e imprecisioni negli algoritmi che gestiscono l'apprendimento della rete neurale oppure da malfunzionamenti della macchina fisica su cui vengono eseguiti tali programmi, tuttavia, questo genere di limitazione è per lo più di natura contingente ed è comune a qualsiasi programma informatico.

Per concludere, indichiamo ora una potenziale limitazione dell'impiego e dell'utilizzo delle reti neurali che esula da questioni di natura tecnica o pratica, ma che invece tocca delle questioni più fondamentali e filosofiche. Abbiamo infatti affermato che le reti neurali altro non sono che sistemi universali di approssimazione di funzioni, ovvero possono, almeno a livello teorico e prescindendo dalle problematiche che abbiamo appena indicato, generare una qualsiasi funzione a partire da un insieme di dati. Ora, se consideriamo il mondo nella sua interezza, sia nei suoi aspetti fisici che in quelli legati all'umano, come potenzialmente spiegabile e descrivibile in termini matematici, allora le reti neurali possono imitare, modellare e riprodurre ogni cosa.

Torniamo, dunque, qui alla domanda di partenza della nostra tesi: è effettivamente tutto riconducibile a una descrizione in termini matematici o di logica formale, o esiste qualcosa, nello specifico una qualche facoltà cognitiva umana, che non può essere ridotta in questi termini? Infatti, sebbene le reti neurali sembrino essere in grado di sorpassare molte delle limitazioni legate alla formalizzazione della conoscenza tacita e del senso comune di cui abbiamo discusso nel secondo capitolo, esse rimangono pur sempre delle macchine di Turing.

Ed essendo delle macchine di Turing non possono che essere soggette all'argomento di Lucas concernente i teoremi di incompletezza di Godel di cui abbiamo trattato nel primo capitolo.

Come abbiamo già indicato, però, quell'argomento non sembra essere in grado di dare una risposta inequivocabile a questo quesito. Inoltre, è possibile che una rete neurale possa, grazie alle sue capacità di apprendimento, imparare a replicare la capacità umana di riconoscere la veridicità delle proposizioni godeliane di un certo numero di sistemi formali. Infatti, tornando a quanto detto nel primo capitolo, se si prende in considerazione un insieme finito di insiemi formali e dei loro corrispettivi enunciati godeliani, allora senz'altro una rete neurale saprà approssimare una funzione che descrive questo insieme e di conseguenza imitare la supposta capacità degli esseri umani di essere "superiori" alle macchine computazionali. Quello che non possiamo sapere è se le reti neurali possano fare lo stesso per qualunque sistema formale, non semplicemente per un insieme finito di essi. Tuttavia, date le inevitabili limitazioni spazio-temporali a cui siamo soggetti, non possiamo stabilire a priori nemmeno se gli esseri umani siano in grado di farlo e, di conseguenza, affermare formalmente che le reti neurali non possono in linea di principio essere in grado di replicare totalmente una mente umana.

Quello che possiamo affermare tuttavia, è che anche da un punto di vista puramente materialistico, le reti neurali potranno al massimo essere una simulazione ed approssimazione, per quanto fedele e accurata, delle nostre capacità cognitive. In altre parole, esse rientrano in una concezione "debole" dell'intelligenza artificiale, e non in una di tipo "forte". Vediamo perché.

Innanzitutto, è opportuno ricordare che "To date, all natural laws discovered through scientific endeavor are stated in terms of mathematical equations that relate *physical* properties of matter"[161]. Questo cosa comporta nel nostro caso? Ebbene, anche qualora le nostre facoltà cognitive fossero effettivamente rappresentabili in termini scientifici come equazioni e funzioni matematiche, esse sarebbero in ogni caso univocamente legate alla particolare composizione chimica e fisica della materia che si suppone le generi, ovvero il nostro cervello. Tuttavia, le reti neurali altro non sono che programmi computer, ovvero delle simulazioni digitali, ovvero "instantiations of an algorithm, and such is by definition multiply realizable, that is, it depends not on the physical composition of the system that implements it but rather on its formal organization"[162].

[161] Tomer Fekete, Shimon Edelman, *The (Lack of) Mental Life of Some Machines*, in *Being in Time*, a cura di Shimon Edelman, John Benjamins Publishing Company, 2012, pag. 97
[162] *Ibidem*

Supponiamo per esempio che una delle nostre facoltà, ad esempio la preferenza nella scelta tra due diverse alternative sia definibile tramite una qualche funzione matematica. A questo punto possiamo immaginare che una certa macchina computazionale T venga dotata di questa funzione. Se quanto abbiamo indicato poco sopra è vero, ovvero questa funzione di preferenza descrive univocamente una certa proprietà della materia, allora dovrebbe essere possibile risalire, a partire da tale funzione, alla composizione chimico-fisica della macchina che la esegue: "If this is right, than the predicate "T prefers A to B" should be definable in terms of the physical-chemical composition of our Turing Machines"[163].

Tuttavia, come abbiamo già indicato in precedenza, le macchine di Turing possono essere fisicamente realizzabili in innumerevoli modi, e una qualsiasi realizzazione fisica di una macchina di Turing è potenzialmente in grado di svolgere i medesimi algoritmi di tutte le altre. In altre parole, i programmi di una macchina di Turing possono essere eseguiti da macchine che possiedono caratteristiche fisiche completamente diverse tra loro. Di conseguenza: "there is no logically valid inference from the premiss that one of our Turing machines has a certain physical-chemical composition to the conclusion that it prefers A to B, (…), nor from the premiss that it prefers A to B to the conclusion that it has a certain physical-chemical composition"[164].

Da ciò possiamo dedurre che, anche qualora le facoltà cognitive e la coscienza umane fossero descrivibili in termini di funzioni matematiche, non potremmo considerare come effettivamente cosciente una rete neurale che apprenda tali funzioni. La rete neurale potrà al massimo simulare tali funzioni. Nello stesso modo in cui una simulazione computerizzata di, ad esempio, un evento atmosferico al fine di prevederne il suo sviluppo *non è* quell'evento atmosferico, e non possiede le sue caratteristiche fisico-chimiche, una ipotetica simulazione della coscienza umana non può essere *la* coscienza umana, e di conseguenza non possiamo considerare come cosciente la macchina che la esegue nello stesso modo in cui lo siamo noi esseri umani.

Secondo questo ragionamento sembrerebbe dunque impossibile attribuire predicati mentali, facoltà cognitive e coscienza ad una macchina computazionale solamente sulla base della sua organizzazione formale, tralasciando la sua composizione psico-fisica. Una simulazione digitale tramite rete neurale dell'attività cerebrale e neuronale umana di conseguenza sarà "at best partial, and, furthermore, (…) fundamentally incapable of realizing

[163] Hilary Putnam, *The Mental Life of Some Machines*, 1967, in *Mind Language and Reality*, H. Putnam, New York, Cambridge University Press, 1975, pag. 414
[164] *Ibidem*

both some of the essential properties of actual neuronal systems and some of the fundamental properties of experience"[165]. Di conseguenza possiamo affermare che: "if machine consciousness is at all possibile, conscious experience can only be instantiated in a class of machines that are entirely different from digital computers, namely, time-continuous, open, analog, dynamical systems"[166].

3.2.2 Reti neurali: il problema del "Black box"

Nel corso del secondo capitolo di questa tesi abbiamo indicato come, in una prospettiva wittgensteiniana, l'attribuzione di termini psicologici alle macchine computazionali sia insensata per due motivi; in primo luogo perché esse non fanno parte della nostra *forma di vita*, e in secondo luogo poiché in realtà il loro funzionamento è totalmente trasparente e comprensibile per noi e di conseguenza non abbiamo alcun bisogno di servirci di termini come "intelligenza" o "pensiero" per spiegare il loro comportamento. Questo secondo punto, certamente valido per i programmi di intelligenza artificiale tradizionali, può forse venire messo in discussione dai sistemi di intelligenza artificiale basati su reti neurali. Vediamo ora perché.

In precedenza abbiamo definito una rete neurale come un insieme di innumerevoli strati di *neuroni*, che altro non sono che delle funzioni lineari semplici che si differenziano tra loro per il fatto di possedere due parametri, *pesi* e *bias*, che vengono modificati nel processo di apprendimento. Una volta addestrata la rete tramite un insieme di dati fornito dai programmatori, è possibile fornirle come input altri dati dello stesso tipo di quelli su cui è stata addestrata, e ottenere in risposta degli output che corrispondono alle ipotetiche "y" della funzione che la rete sta approssimando. Dal punto di vista generale, di conseguenza, le reti neurali sono, al pari di qualsiasi altro programma informatico, nient'altro che sistemi di elaborazione che generano degli output a partire da determinati input. Tuttavia, a differenza dei normali programmi, il funzionamento interno della rete neurale non è per noi trasparente e comprensibile. Le reti neurali sono infatti dei sistemi che in gergo tecnico vengono definiti come "black box", ovvero sistemi di cui possiamo conoscere solamente gli input e gli output, ma non il processo interno che li ha generati. Spieghiamo meglio questo punto.

Un tradizionale programma informatico è composto da un serie di linee di codice che altro non sono che la traduzione in uno specifico linguaggio di programmazione di una serie di

[165] Tomer Fekete, Shimon Edelman, *The (Lack of) Mental Life of Some Machines*, in *Being in Time*, a cura di Shimon Edelman, John Benjamins Publishing Company, 2012, pag. 95
[166] *Ibidem*

istruzioni logiche. Conoscendo il linguaggio di programmazione è possibile, pertanto, comprendere perfettamente il funzionamento del programma, e i procedimenti che hanno portato alla generazione di determinati output. Se invece proviamo a "guardare dentro" una rete neurale, quello che possiamo osservare è un vastissimo numero di neuroni, ovvero funzioni lineari semplici, ma non possiamo capire né come l'insieme di tutti i neuroni "cooperi" per giungere al risultato finale, né quale sia l'importanza e il ruolo specifico di ciascun neurone all'interno del sistema: "currently, there is no mechanistic or quantitative explanation for how the distributed activities of individual neurons give rise to network outputs or cognitive experiences, especially in the deeper layers responsible for mapping features to inputs"[167]. In aggiunta a ciò, anche la memoria stessa delle reti neurali non è contenuta in un luogo ben definito ma è distribuita tra tutti i neuroni della rete. In un computer tradizionale, infatti, tutti i dati e i programmi non solo sono esplicitamente codificati e immagazzinati, ma possiedono un vero e proprio luogo fisico all'interno della macchina in cui sono conservati. Qualora volessimo recuperare una certa informazione in un computer tradizionale, dovremmo semplicemente andare all'"indirizzo" in cui essa è presente per trovarla. Tutto questo non è applicabile invece alle reti neurali, sia perché non esiste un luogo fisico particolare in cui le conoscenze della rete vengono immagazzinate, sia perché, anche qualora esistesse, non sarebbe codificato in un formato a noi comprensibile.

Si tratta, facendo un parallelo, della stessa situazione cui ci troviamo di fronte studiando il nostro cervello. Tramite scansioni elettromagnetiche noi infatti possiamo individuare quali parti del nostro cervello sono responsabili delle nostre facoltà mentali, e vedere i meccanismi di attivazione e il funzionamento dei singoli neuroni che lo compongono. Tuttavia, non riusciamo ancora a comprendere come dalle reazioni fisiche e chimiche che avvengono nei singoli neuroni possano scaturire dei fenomeni come la coscienza. Anche il nostro stesso cervello può essere definito come un black box: possiamo certamente "guardare" al suo interno e misurare la sua attività in termini chimici e fisici, ma non possediamo una teoria che spieghi come questa attività sia poi responsabile dei nostri comportamenti o della nostra coscienza e intelligenza.

Stando così le cose, riprendendo alcuni dei ragionamenti fatti nel secondo capitolo possiamo forse concludere che, almeno da un certo punto di vista, sono cambiate alcune delle condizioni che rendevano insensata e non necessaria l'attribuzione dei cosiddetti "termini

[167] Paul Blazek, Milo Lin, *Explainable Neural Networks that Simulate Reasoning*, Nature Computational Science, 2021

psicologici" in riferimento alle macchine computazionali. Lo stesso motivo per cui le reti neurali siano così efficaci nello svolgere determinati compiti sfugge alla nostra comprensione; o meglio appare paradossale come una rete neurale che, in termini matematici, altro non è che "an overparameterized, opaque model to minimize a loss function on some training data"[168], possa poi concretamente svolgere compiti che generalmente riteniamo come necessitanti di intelligenza, come ad esempio scrivere dei testi, comporre della musica o dipingere un quadro.

Possiamo dunque considerare come effettivamente "intelligenti" le reti neurali? Ancora una volta il problema risiede nel significato che diamo a questo termine. Quello che possiamo affermare con una certa sicurezza è che le reti neurali, anche qualora fossero "intelligenti", non lo sarebbero di certo nello stesso modo in cui lo siamo noi. Sebbene, come abbiamo appena indicato, il funzionamento delle reti neurali sfugga in un certo qual senso alla nostra comprensione, possiamo tuttavia affermare che sicuramente le loro operazioni sono di natura esclusivamente *sintattica* e non *semantica*. In altre parole, l'operato di una rete neurale altro non è che una manipolazione logica di simboli e numeri, senza che ci sia da parte loro una qualsiasi capacità di comprensione, astrazione e creazione di concetti: "traditional deep learning does not learn generalizations, it is simply a sophisticated pattern-matching technology, not a true learning method"[169].

Una rete neurale, infatti, non fa altro che tradurre in termini di funzioni degli schemi presenti in determinati insiemi di dati, cercando di collegare in maniera migliore possibile un certo input ad un output. La funzione che essa "apprende" può essere applicata solamente ad insiemi di dati che abbiano la medesima forma e struttura dei dati sui quali la rete è stata addestrata. Una rete neurale non apprende generalizzazioni e concetti che le permettano di astrarre dal contesto specifico del suo addestramento e operare in situazioni diverse da esso:

> Consider a monkey that is painstakingly taught how to impressively play some pieces on the piano. Does it have the same understanding of music as a classically trained pianist? Of course not. The former has learned how to approximate a copy of its training, while the latter has used prior experiences to develop the deep and general understanding necessary to play new music, to improvise, and to appreciate the music player by others. [170]

[168] Paul Blazek, *How Aristotle is Fixing Deep Learning's Flaws*, The Gradient, 2022
[169] *Ibidem*
[170] *Ibidem*

Le reti neurali se paragonate ai classici programmi informatici riescono certamente a svolgere dei compiti ritenuti in passato impensabili, con una precisione e un'accuratezza che in molti casi rendono impossibile distinguere il risultato del loro operato da quello umano. In questo senso esse posseggono sicuramente più criteri per essere definite come "intelligenti". Tuttavia, anche prescindendo dalla delicata questione dell'attribuzione di facoltà cognitive e coscienza a delle macchine, appare evidente come al giorno d'oggi siamo ancora molto distanti dal poter attribuire l'*intelligenza artificiale generale*[171] ad un qualunque sistema di intelligenza artificiale. Una rete neurale, infatti, può solamente svolgere il compito ben specifico per il quale è stata addestrata, e può lavorare solamente su insiemi di dati che corrispondano esattamente nella forma a quelli dai quali ha appreso. Finché questa eccessiva specializzazione e queste mancanze di generalizzazione e astrazione delle reti neurali non verranno superate sarà di conseguenza impossibile parlare quanto meno di intelligenza artificiale generale in relazione a questi sistemi.

Alcuni studiosi, al fine di risolvere questi problemi strutturali, hanno proposto una potenziale commistione tra le reti neurali e i vecchi programmi di intelligenza artificiale. In altre parole, si tratterebbe di unire l'approccio cosiddetto "simbolico" dei programmi tradizionali a quello "connettivista" delle reti neurali: "to understand the neural basis of cognition and pursue artificial general intelligence, it is necessary to bridge the gap between symbolism and connectivism"[172]. I classici programmi di intelligenza artificiale di tipo simbolico, ispirati alla psicologia, cercano, infatti, di codificare esplicitamente i processi cognitivi umani in termini di linguaggio logico e di programmazione, offrendo il vantaggio di essere trasparenti e comprensibili per noi, ma lo svantaggio di trascurare la componente fisiologica che sta alla base del funzionamento cerebrale. Le reti neurali, d'altro canto, sono ispirate al funzionamento del nostro cervello e simulano il comportamento dei nostri neuroni, generando però il problema del black-box di cui abbiamo trattato poc'anzi, fallendo nel replicare "symbolic manipulation and other fundamental cognitive processes such as high-level reasoning and deliberation"[173]. Si tratta, in ogni caso, di campi di studio molto recenti e la ricerca in questo settore è ancora nelle sue fasi iniziali, e i progressi sono di conseguenza ancora limitati.

---

[171] Con "Intelligenza Artificiale Generale" (Artificial general Intelligence) si intende la capacità di un agente di comprendere ed eseguire qualsiasi compito intellettuale al pari di un essere umano. A differenza dell'Intelligenza artificiale forte (Strong AI) non è necessario che la macchina debba poter essere considerata anche come cosciente.
[172] Paul Blazek, Milo Lin, *Explainable Neural Networks that Simulate Reasoning*, Nature Computational Science, 2021
[173] *Ibidem*

Un ultimo punto su cui ci soffermeremo è la possibilità che le facoltà cognitive umane, la coscienza, l'intelligenza e il pensiero siano necessariamente ed esclusivamente causate dalla particolare natura chimica e neurofisiologica del cervello umano. Questa posizione, nota con il nome di "naturalismo biologico" è la stessa sostenuta da John Searle nel suo celebre argomento della Stanza Cinese. I sostenitori di questa posizione sono concordi sul fatto che una macchina possa pensare, considerano infatti il cervello come una macchina, ma solo qualora possegga le stesse esatte caratteristiche fisico-chimiche del cervello. Coscienza, intenzionalità, pensiero, intelligenza sarebbero infatti dei fenomeni biologici naturali che non possono essere replicati se non da macchine "naturali": "Whatever else intentionality is, it is a biological phenomenon, and it is as likely to be as causally dependent on the specific biochemistry of its origins as lactation, photosynthesis, or any other biological phenomena"[174]. In altre parole, si tratta di una posizione filosofica opposta al dualismo di matrice cartesiana che vede la mente come un'entità separata dal corpo e caratterizzata da processi di natura logico-formale indipendenti dalla materia corporale. Secondo i sostenitori del naturalismo biologico, di conseguenza, nessun programma, sia esso "tradizionale" o basato su reti neurali, che operi su una qualsiasi macchina che non sia il cervello, potrà mai essere considerato come "intelligente" o "cosciente".

Ora, al di là di queste considerazioni di natura filosofica sulla natura del pensiero e della coscienza, è indubbio che le intelligenze artificiali basate su reti neurali stiano sempre di più entrando a fare parte delle nostre vite, svolgendo sempre più compiti e in maniera sempre più efficace. Sembra dunque sempre più necessario comprendere meglio, anche a livello filosofico, il funzionamento di questi sistemi, non solamente per limitarne utilizzi potenzialmente dannosi, ma anche per poter trarre nuove conoscenze sul modo in cui noi stessi pensiamo e ragioniamo.

[174] John Searle, *Minds, Brains, and Programs*, 1980, *The Philosophy of Artificial Intelligence*, a cura di M. Boden, New York, Oxford University Press, 1990, pp. 86-87

Conclusioni

Giunti alla conclusione di questa tesi, ritorniamo alla domanda da cui siamo partiti. Siamo noi esseri umani delle macchine? Gli argomenti di natura logico-matematica, esposti nel primo capitolo, che si propongono di dimostrare logicamente l'esistenza di almeno una facoltà intellettiva umana non riducibile alla computazione, non sembrano poter dare una risposta definitiva e incontrovertibile a questo dilemma. Nonostante il loro innegabile fascino e il merito di aver aperto un lunghissimo e acceso dibattito sul tema dell'intelligenza artificiale, essi sono infatti forse troppo astratti e ideali per poter essere applicati appropriatamente agli esseri umani e alla macchine reali.

Come abbiamo indicato nel secondo capitolo, la risposta alla domanda che dà il titolo a questa tesi, risiede forse in grandissima parte nel significato che attribuiamo a termini come "intelligenza" e "pensiero". Sicuramente possiamo affermare che, almeno per il momento, a nessuna macchina computazionale può essere attribuita la nostra stessa intelligenza, né alcun programma di intelligenza artificiale ha ancora raggiunto quella che in gergo tecnico viene definita "artificial general intelligence". Questo, tuttavia, non esclude necessariamente che il nostro cervello sia anch'esso a tutti gli effetti una macchina, e nello specifico una macchina computazionale. I progressi tecnologici nel campo dell'intelligenza artificiale sembrano infatti ridurre sempre di più la differenza tra le abilità e le prestazioni di una macchina e quelle di un essere umano.

Quello che però possiamo affermare con relativa sicurezza è che esiste una dimensione propria dell'essere umano che difficilmente verrà mai riprodotta artificialmente da una macchina. Si tratta di quella dimensione sociale, pratica e relazionale che abbiamo trattato nel secondo capitolo di questa tesi e di cui possono far parte necessariamente solamente coloro che condividono la stessa *forma di vita*. I nuovi sistemi di intelligenza artificiale basati sulle reti neurali e sull'apprendimento, infatti, nonostante stiano rivoluzionando il mondo e producendo risultati impensabili fino a qualche anno fa, sono e rimangono fondamentalmente diversi da noi. Per quanto possa apparire stupefacente il fatto che una rete neurale produca dei testi letterari, componga della musica o dipinga un quadro (per citare alcune delle più applicazioni più "umane" dell'intelligenza artificiale), è pur sempre vero che tali opere sono il risultato di operazioni di natura esclusivamente *sintattica*. Le reti neurali non possiedono concetti, valori, o esperienze, non possono comprendere né il significato di quello che stanno facendo né il motivo per cui lo fanno.

Noi esseri umani, d'altro canto, pensiamo a noi stessi come dotati di *semantica*. I testi che scriviamo o i discorsi che pronunciamo non sono riconducibili al solo rapporto sintattico tra le parole, o alla manipolazione di simboli secondo regole logiche, ma possiedono significato e intenzionalità. Anche qualora un testo prodotto da un'intelligenza artificiale risultasse come indistinguibile da uno scritto da un essere umano, il processo attraverso il quale esso è stato creato rimarrebbe fondamentalmente differente.

In ultima istanza, inoltre, qualsiasi sistema di intelligenza artificiale presente e passato ha sempre ignorato la natura biologica e fisica del cervello umano da cui supponiamo sorgano le nostre facoltà mentali e la nostra coscienza. Si è infatti sempre cercato di realizzare dei modelli che riproducessero il nostro modo di pensare prescindendo dalla stessa struttura materiale che lo produce. In questo senso qualunque sistema di intelligenza artificiale basato su computers digitali potrà al massimo essere considerato come un'imitazione e una simulazione del nostro pensiero, ma mai come avente le nostre medesime facoltà mentali e coscienza.

Quello che possiamo affermare dunque è che, anche qualora fossimo delle "macchine", è molto improbabile che la nostra natura sia esclusivamente riducibile alla computazione e al calcolo:

> More fundamentally, it is not even clear yet that computation of *any* kind, whether involving serial or parallel processing, provides the key to human cognitive capacity. Computation clearly is *one* of our cognitive capacities - one that we have found so useful that we have developed machines to do it for us. But to regard the operations of those machines as providing a model for *all* of our cognitive activities looks suspiciously like the overworking of a metaphor. It appears to ignore too many facets of our mental life which are inseparable from human cognition, such as sensation and emotion, and to disregard the biological aspects of our nature.[175]

[175] E. Lowe, *An Introduction to the Philosophy of Mind*, Cambridge, Cambridge University Press, 2004, p. 228

Bibliografia

AA. VV., *Intelligenza Naturale e Intelligenza Artificiale*, Genova Marietti, 1991

AA. VV., *La Filosofia degli Automi*, Torino, Boringhieri, 1965

AA. VV., *Simulation and Knowledge of Action*, Amsterdam-Philadelphia, John Benjamins Publishing Company, 2002

AA. VV., *Valori e Limiti del Senso Comune*, Milano, Franco Angeli, 2004

Agazzi E., *Il Senso Comune e l'Unità dell'Esperienza*, in *Valore e Limiti del Senso Comune*, a cura di E. Agazzi, Milano, Franco Angeli, 2004, pp. 25-38

Agazzi E., *Valore e Limiti del Senso Comune*, Milano, Franco Angeli, 2004

Benacerraf P., *God, The Devil and Gödel*, "The Monist", 51 (1967), pp. 9-32

Berto F., *Logica da Zero a Gödel*, Bari, Laterza, 2007

Berto F., *Tutti pazzi per Gödel!*, Bari, Laterza, 2008

Blazek P., *How Aristotle is Fixing Deep Learning's Flaws*, The Gradient, 2022

Blazek P., Lin M., *Explainable Neural Networks that Simulate Reasoning*, Nature Computational Science, 2021

Boden M., *Escaping from the Chinese Room*, in *The Philosophy of Artificial Intelligence*, a cura di M. Boden, New York, New York University Press, 1990, pp. 89-104

Bringsjord S., *What Robots Can and Can't Be*, Dordrecht, Springer Science+Business Media, 1992

Burgin M., *Theory of Knowledge*, Singapore, World Scientific, 2017

Busch P., Sanzogni L., Guzman G., *Artificial Intelligence and Knowledge Management: Questioning the Tacit Dimension*, "Prometheus", 35 (2017), pp. 37-56

Cappelen H., J. Dever, *Making AI Intelligible*, Oxford, Oxford University Press, 2021

Cartesio, *Discorso sul Metodo*, tr. it. di Armando Carlini, Bari, Laterza, 1965

Dennet D. C., *Cognitive Wheels: The Frame Problem of AI*, in *The Philosophy of Artificial Intelligence*, a cura di M. Boden, New York, New York University Press, 1990, pp. 147-170

Dreyfus H., Dreyfus S., *Making a Mind Versus Modelling the Brain: Artificial Intelligence Back at a Branch Point*, in *The Philosophy of Artificial Intelligence*, a cura di M. Boden, New York, New York University Press, 1990, pp. 309-333

Dreyfus H., S. Dreyfus, *Mind over Machine*, New York, Free Press, 1986

Dreyfus H., *Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian*, "Artificial Intelligence", 171 (2007), pp. 1137-1160

Edelman S., Fekete T., *The (Lack of) Mental Life of Some Machines*, in *Being in Time*, a cura di S. Edelman, John Benjamins Publishing Company, 2012, pp.

Feferman S., *Are There Absolutely Unsolvable Problems? Gödel's Dichotomy*, "Philosophia Mathematica Advance Access", III (2006), pp. 1-19

Feferman S., *Gödel, Nagel, Minds and Machines*, "The Journal of Philosophy", 106 (2009), pp. 201-219

Franzen T., *Gödel's Theorem: an Incomplete Guide to its Use and Abuse*, Wellesley, A K Peters, 2005

French R. M., *Subcognition and the Limits of the Turing Test*, in *Machines and Thought*, a cura di P. Millican, A. Clark, New York, Oxford University Press, 1996, pp. 11-26

Galton A., *The Church-Turing Thesis: Its Nature and Status*, in *Machines and Thought*, a cura di P. Millican, A. Clark, New York, Oxford University Press, 1996, pp. 137-164

Gloess P., *Capire l'Artificiale*, Padova, Franco Muzzio Editore, 1985

Gödel K., *Collected Works, Volume III: Unpublished Essays and Lectures*, a cura di S. Feferman et al., New York-Oxford, Oxford University Press, 1995

Gödel K., *Collected Works, Volume V: Correspondence H-Z*, a cura di S. Feferman et al., New York-Oxford, Oxford University Press, 1995

Goldkind S., *Machines and Intelligence*, New York, Greenwood Press, 1987

Hofstadter D., *Gödel, Escher, Bach: An Eternal Golden Braid*, New York, Basic Books, 1979

Lindstrom P., *Penrose's New Argument*, "Journal of Philosophical Logic", XXX (2001), pp. 241-250

Lindstrom P., *Remarks on Penrose's New Argument*, "Journal of Philosophical Logic", XXXV (2006), pp. 231-237

Lowe E., *An Introduction to the Philosophy of Mind*, Cambridge, Cambridge University Press, 2004

Lucas J. R., *Minds, Machines and Gödel*, "Philosophy", XXXVI (1961), pp. 112-127

Lucas J. R., *Minds, Machines and Gödel: a Retrospect*, in *Machines and Thought*, a cura di P. Millican, A. Clark, New York, Oxford University Press, 1996, pp. 103-124

Magnani L., *Ingegnerie della Conoscenza*, Milano, Marcos y Marcos, 1997

Marbach E., *Mental Representation and Consciousness*, Dordrecht, Kluwer Academic Publishers, 1993

McCarthy J., Hayes P., *Some Philosophical Problems from the Standpoint of Artificial Intelligence*, in *Machine Intelligence 4*, a cura di B. Meltzer, D. Michie, Edinburgh, Edinburgh University Press, 1969, pp. 463-502

Micheli G., *Il Concetto di Automa nella Cultura Greca dalle Origini al Sec. IV A.C.*, "Rivista della Storia della Filosofia", 53, no.3 (1998), pp. 421-462

Montecucco E., *Il Senso Comune come "Teoria" e come "Limite"*, in *Valore e Limiti del Senso Comune*, a cura di E. Agazzi, Milano, Franco Angeli, 2004, pp. 57-72

Moravia S., *L'Enigma della Mente*, Roma-Bari, Laterza, 1986

Moser P., *The Oxford Handbook of Epistemology*, New York, Oxford University Press, 2002

Nagel E., Newman J., *Gödel's Proof*, New York, NYU Press, 2001

Nagel T., *What Is Like to Be a Bat*, "The Philosophical Review, 83 (1974), pp. 435-450

Neumaier O., *A Wittgensteinian View of Artificial Intelligence*, in *Artificial Intelligence. The Case Against*, a cura di R. Born, Londra-Sydney, Crook Helm, 1986, pp. 132-173

Newell A., *The Knowledge Level*, "Artificial Intelligence", 18 (1982), pp. 87-127

Norvig P., Russel S., *Artificial Intelligence a Modern Approach*, Pearson Education Limited, 2022

Parisi D., *Intervista sulle Reti Neurali*, Bologna, Il Mulino, 1989

Penrose R., *Beyond the Doubting of a Shadow*, "Psyche", 2 (1996)

Penrose R., *La Mente Nuova dell'Imperatore*, Milano, Rizzoli, 1992

Penrose R., *Shadows of the Mind*, Oxford, Oxford University Press, 1994

Polanyi M., *Personal Knowledge*, Chicago University Press, 1974

Pratt V., *Macchine Pensanti*, Bologna, Il Mulino, 1990

Putnam H., *Brains and Behaviour*, in *Mind Language and Reality*, H. Putnam, Cambridge, Cambridge University Press, 1975, pp. 325-341

Putnam H., *Minds and Machines*, in *Mind Language and Reality*, H. Putnam, Cambridge, Cambridge University Press, 1975, pp. 362-385

Putnam H., *Robots: Machines or Artificially Created Life?*, "The Journal of Philosophy, 21 (1964), pp. 668-691

Putnam H., *The Mental Life of Some Machines*, in *Mind Language and Reality*, H. Putnam, Cambridge, Cambridge University Press, 1975, pp. 408-428

Raatikainen P., *On the Philosophical Relevance of Gödel's Incompleteness Theorems*, "Revenue Internationale de Philosophie", 59 (2005), pp. 513-534

Searle J., *Minds, Brains, Programs*, "Behavioral and Brain Sciences", 3 (1980), pp. 417-457

Shapiro S., *Incompleteness, Mechanism and Optimism*, "The Bulletin of Symbolic Logic", 4 (1998), pp. 273-302

Shapiro S., *Mechanism, Truth and Penrose's New Argument*, "Journal of Philosophical Logic", XXXII (2003), pp. 19-42

Simon H., *Machine as Mind*, in *Machines and Thought*, a cura di P. Millican, A. Clark, New York, Oxford University Press, 1996, pp. 81-102

Tabossi P., *Intelligenza Naturale e Intelligenza Artificiale*, Bologna, Il Mulino, 1988

Thagard P., *Cervelli a Confronto*, trad. it. di Piero Corsini, Milano, Franco Angeli, 2021

Turing A,. *On Computable Numbers, with an Application to the Entscheidungsproblem*, "Monatsheftc Math. Phys.", 38 (1931), pp. 173-198

Turing A., *Computing Machinery and Intelligence*, "Mind", LIX, no. 2236 (1950), pp. 433-460

Wang H., *Dalla Matematica alla Filosofia*, tr. it. di Alberto Giacomelli, Torino, Boringhieri, 1984

Whitby B., *The Turing Test: AI's Biggest Blind Alley?*, in *Machines and Thought*, a cura di P. Millican, A. Clark, New York, Oxford University Press, 1996, pp. 53-62

Wittgenstein L., *Libro Blu e Libro Marrone*, tr. it. di Amedeo Conte, Torino, Einaudi, 1983

Wittgenstein L., *Ricerche Filosofiche*, tr. it. di Mario Trinchero, Torino, Einaudi, 1967